Chapter 1: Introduction

Chapter 2: Travel Demand

Chapter 3: Costs

Chapter 4: Pricing

Chapter 5: Investment

1. INTRODUCTION

At the heart of all modern economic activity is trade. People trade labor and ideas for cash, and cash for goods and services; firms trade technology, expertise, financial capacity, intermediate goods, administrative functions, and many other things with each other, with individuals, with governments. All these transactions require communications and most require transportation of goods or of people — to work, shopping, tourist sites, meeting locations. Thus it is fair to say that transportation is central to economic activity.

Cities exist when there are special advantages to carrying out economic activities in proximity, advantages often called “economies of agglomeration” because costs are lower when certain groups of activities locate close to each other. The primary reason for agglomeration economies, especially in a world with low communication costs, is that transportation costs are still significant and proximity reduces them. A corollary is that anything that reduces transportation costs within an urban area increases the extent to which its activities are easily linked to each other, and thus takes further advantage of agglomeration economies. In a world of many competing urban centers, those with efficient transportation systems have an advantage.

The study of transportation involves researchers trained in many disciplines, including engineering, economics, geography, planning, business, and regional science. Regardless of its disciplinary origin, transportation research has become increasingly sophisticated in its use of economics. This trend has brought a solid practical footing to policy analysis, by showing how the ideas and goals generated within various analytical frameworks can be reconciled with the actual behavior of users and with the resource constraints of a real economy.

This book reviews the contributions that economics can make to the analysis of urban transportation. It concentrates mainly on industrialized nations, but many of the principles are equally applicable to developing nations. This is especially true in light of the ubiquitous rise in amount of travel by private automobile—which, in developed nations at least, is closely related to the dispersion of urban population and employment. Because of the predominance of this trend, we place heavy emphasis on highway transportation and use it as a topic in which to present a comprehensive set of formal models.

1.1 The Scope of Urban Transportation Economics

The boundaries of transportation economics are neither well defined nor static. Nevertheless, the reader is entitled to know what principles we use to limit the scope of this review. Aside from the inevitable one that we try to write about what we know best, the following observations are guidelines.

Transportation economics is, of course, a branch of economics. Hence it focuses on resource allocation and how the interactions among independent agents bring about a self-consistent outcome. It draws from and interacts closely with transportation engineering, urban planning, and other disciplines, but has a somewhat different emphasis. Engineering emphasizes facility design and implementation, while economics emphasizes behavioral principles and resource allocation. The disciplines of management, public administration, and urban planning are concerned with the formulation of workable transportation policies, for example by studying decision processes and organizational structures. An important role of economics is to inform these disciplines about the complex ways in which transportation policies exert their influence. Economics is well suited to predict the ultimate results of behavioral shifts among interacting economic actors in response to policy implementations or exogenous events. It also can identify tradeoffs between efficiency and other goals.

These orientations of the subject give it certain characteristics that are evident in this review. Transportation economics tends to focus on models that illustrate concepts, as opposed to those whose output is the actual design of facilities or regulations. Hence its models are usually at a coarse rather than a fine geographical scale, lending themselves to "sketch planning" of the broad features of a transportation system.

Analysis often proceeds by defining a demand structure and a supply structure for a set of goods or services, then searching for an outcome that is consistent with both structures. This is a normal microeconomic approach, although the nature of transportation creates ambiguities in the boundary between demand and supply: for example, is the time required for a trip an attribute affecting demand, or is it part of the cost? Either viewpoint is valid so long as demand and cost are consistently defined; doing so is the task of Chapters 2 and 3, respectively. Their interconnection is made explicit in the discussion of value of time in Section 2.6.

Demand and supply structures are complex, involving many types of people, modes, locations, and times. For this reason, finding a consistent solution — an equilibrium, in economics parlance — requires considerable analytical sophistication. A pioneering study by Beckmann, McGuire, and Winsten (1955) showed how to do this for arbitrary transportation networks given certain assumptions. In the presence of increasing returns to scale and lack of efficient pricing (conditions which, as we shall see, characterize urban transportation), this equilibrium may differ in drastic and surprising ways from a configuration satisfying optimality conditions. In other words, what happens when people make decisions looking out for themselves is not necessarily what they would choose collectively. The common thread throughout this review is that transportation economics should provide the tools to understand and quantify such differences, and to design policies that address them.

Three important types of policies to consider are pricing, investment, and industry organization (including regulations contingent on the form of organization). Pricing and investment have long been the hallmark of urban transportation economics, and are accorded full treatment here (chapters 4 and 5). Industrial organization has likewise been of great concern, and recent changes throughout the world have brought to the fore questions of oligopolistic markets, regulation, and government versus private provision of services. Chapter 6 considers a number of these questions.

Transportation potentially affects the nature of the urban area itself. If transportation were costless, participants in an economy would have no economic reason to locate close to one another. The study of this influence is clearly germane to transportation policy. To analyze it fully requires the full power of disciplines such as urban geography, urban economics, and regional science, which seek to explain the shape of urban development. Important as this is, it is too ambitious for this book and we limit ourselves to a brief summary in Section 2.1.6.

1.2 The Scope of This Book

This book aims at providing a self-contained introduction to the major subject areas of research. We have chosen to explore certain topics in great detail in order to illustrate how the analysis can be put to practical use, and others much more cursorily. In both cases, extensive

citations provide the reader opportunities to expand on any particular theme. Thus we aim to make the book suitable both as an initial textbook for students with a good technical background and as a reference work for practicing researchers and professionals.

Above all, we attempt to show how to construct a set of workable models that can be adapted, refined, and combined in subsequent research. In the sections on highway congestion, for example, the models presented incorporate critical features such as queueing, trip scheduling, and peak shifting that are often omitted for simplicity. There is evidence that these phenomena greatly affect the outcomes from policies to relieve congestion, and their analysis has been made tractable by recent technical advances. It is our hope that such models will provide a common language for researchers in urban transportation, facilitating comparisons among theoretical innovations and among empirical applications.

Models presented here are suitable as building blocks for comprehensive models of an entire urban transportation system. Such comprehensive models would permit investigation of questions such as: What would the overall transportation system look like under drastically different policies designed to improve economic efficiency? Would the small and dwindling share of trips by public transit become much larger? Would rapid-transit systems that seem uneconomical under present conditions become desirable? More pragmatically, such models would enable researchers to better predict how policies in one area, such as public transit, are developed. These questions are further discussed in the concluding chapter.

2. TRAVEL DEMAND

In order to plan transportation facilities, it is necessary to forecast how much they will be used. In order to price them rationally and determine the best operating policies, it is necessary to know how users respond to prices and service characteristics. In order to evaluate whether a project is worthwhile at all, it is necessary to have a measure of the benefits it produces. All these requirements are in the province of travel demand analysis.

The demand for travel takes place in a multi-dimensional setting. The traditional sequential framework used by many metropolitan transportation planning agencies considers four choice dimensions: trip generation (the total number of trips originating from an area); trip distribution (the locations of the trips' destinations); modal choice (the means of travel, such as car, bus, train, bicycle, or walking); and trip assignment (the exact route used). More recently, researchers have paid greater attention to other dimensions of choice, such as residential and job location, household automobile ownership, and the time of day at which trips are taken.

Travel is a derived demand: it is usually undertaken not for its own consumption value, but rather to facilitate a complex and spatially varied set of activities such as work, recreation, shopping, and home life. This observation links the study of travel demand to studies of labor supply, firms' choices of technologies, and urban development. It furthermore calls attention to an increasingly common form of travel: the linking together of several trip purposes into one integrated travel itinerary or *tour*, a process known as *trip chaining*.

The chapter begins (Section 2.1) by describing what happens when a conventional aggregate approach to economic demand is applied to transportation. It then moves on to disaggregate models (Section 2.2), also known as "behavioral" because they depict the individual decision-making process. Section 2.3 presents examples of models explaining some key travel choices: mode, trip time, and express-lane option. More specialized topics are then discussed (Sections 2.4-2.5). Finally, Section 2.6 analyzes two behavioral results from travel-demand studies that have special importance for policy: travelers' willingness to pay for travel-time savings and improved reliability.

Other reviews of travel demand include Bates (1997) and Small and Winston (1999). Collections of more specialized papers include Garling, Laitila and Weston (1998), Hensher and Button (2000), and Mahmassani (2002). Our discussion, like most of the literature, is mainly

mainly about passenger transportation; studies of the demand for urban freight transportation tends to use similar methods, although with many details specific to characteristics of the industry (D'Este, 2000).

2.1 Aggregate Models

The approach most similar to standard economic analysis of consumer demand is an aggregate one. The demand for some portion of the travel market is explained as a function of variables that describe the product or its consumers. For example, total transit demand in a city might be related to the amounts of residential and industrial development, the average transit fare, the costs of alternative modes, some simple measures of service quality, and average income.

Much can be learned simply from cross-tabulations of survey data. For example, Pucher and Renne (2003) tabulate results from the 2001 National Household Travel Survey (NHTS) in the United States.[1] Among the many interesting findings are that public transit accounts for just 1.6 percent of all daily trips in 2001, considerably less than walking and bicycling (9.5 percent). For work trips transit has a higher share (3.7 percent). However, there is a 15-fold difference in transit share across the nine Census regions of the US, the highest being the region that includes New York City. Nearly 92 percent of US households own a car, varying across five income groups from 73.5 percent (lowest income) to 98.5 percent (highest income).

Hu and Young (1999, Fig. 13), using 1995 U.S. data from the National Personal Transportation Survey, present variations by trip purpose and time of day. Only during early morning hours, roughly 4-7 a.m., do work trips constitute a majority. Nevertheless, it would be a mistake to conclude that work trips are no longer important for urban congestion: they account for more than half of all trips in selected Belgian cities (Van Dender, 2001, p. 103), and a very high fraction of peak-period trips on some of the most notoriously congested Los Angeles freeways (Giuliano, 1994, p. 261).

We now turn to more formal statistical techniques for measuring travel demand.

[1] This survey continues the well-known National Personal Transportation Survey (NPTS), conducted previously at approximately six-year intervals and covering daily travel, as well as the American Travel Survey covering long-distance travel.

2.1.1 Cross-Sections of Metropolitan Areas

Several studies have used aggregate data, such as are commonly reported by transit authorities or local governments, to study influences on travel behavior across cities. Gordon and Willson (1984) compile an international data set of 91 cities with light rail systems, and estimate simple regression equations explaining the ridership (per kilometer of line) on those systems. Using the semilogarithmic form with just four variables, they show that ridership is positively related to city population density and gross national product per capita; and that it is 52 percent lower in the U.S. and 55 percent higher in eastern Europe than elsewhere, other variables being held constant.[2] Applying this model to three U.S. and two Canadian cities with light-rail systems then under construction, they predict far lower ridership than the official forecasts -- about half in most cases, but less than one-seventh in the case of Detroit. It has been shown subsequently by Pickrell (1992) that nearly every modern rail system built in the U.S. has in fact attained less than half the originally forecast patronage.

Pucher (1988) compares auto and transit use in twelve western European and North American nations. He then presents data on those nations' levels of transit and highway subsidies, transit service, gasoline and motor vehicle taxes, licensing and parking policies, and land-use policies; these data suggest that the main explanations for differences in travel modes are automobile taxation and land-use policies. But he also finds that auto ownership in western Europe is growing much faster than in the U.S. and that ownership rates per capita appear to be converging to a common value.

Black (1990) uses regression analysis on data from 120 U.S. metropolitan areas to see what factors influence the fraction of metropolitan workers who walk to work. This fraction varies from 1.9 to 15.7 percent, averaging 5.4 percent. It is higher in cities with military bases or universities, in small cities, and where incomes are low. Black's paper is a useful reminder that walking can be an important journey-to-work mode, its prevalence being sensitive to land uses and demographics. More recent data suggest that the importance of walking to work has

[2] Their model 3, p. 137.

declined, but still its share was 3.4 percent for work trips in 2001.[3] Plaut (2004) reports that in Tel Aviv, 9.4 percent of workers walk to work, a fraction that varies by gender, age, and household status.[4]

2.1.2 Cross-Sections within a Metropolitan Area

Statistical analysis can also be used to analyze trip-making in different parts of one metropolitan area. This approach, known as *direct demand modeling*, was introduced by Domencich and Kraft (1970) to explain the number of round trips between zone pairs, by purpose and mode, in the Boston area. An example is the analysis by Kain and Liu (2002, Table 5-10) of mode share to work in Santiago, Chile. The share is measured for each of 34 districts ("communas") and its logarithm is regressed on variables such as travel time, transit availability, and household income. The most powerful predictors for automobile share are vehicle ownership and household income, both of which increase the share. The share for the Metro rail system is strongly increased, quite naturally, by presence of a Metro station in the district, and is strongly decreased by high household income.

Dagenais and Gaudry (1986), analyzing Montreal data, find it important to include observations of zone pairs with zero reported trips, using the standard "tobit" model for limited dependent variables to account for the fact that the number of trips between two zones cannot be negative. This illustrates a pervasive feature of travel-demand analysis: many of the variables to be explained are limited in range so that ordinary regression analysis, which typically assumes the dependent variable to have a normal (bell-shaped) distribution, is inappropriate. For this reason, travel-demand researchers have contributed importantly to the development of techniques, discussed later in this chapter, that are appropriate for such data (McFadden, 2001). We note here one such technique that is applicable to aggregate data.

Suppose the dependent variable of a model can logically take values only within a certain range. For example, if the dependent variable x is the modal share of transit, it must lie between

[3] Pucher (2003, Table 3). The decline in walking is seen in Pucher's Table 4 for 1977-1995, which tallies all daily travel; the subsequent trend cannot be discerned from the NPTS and NHTS data due to changes in survey methodology in 2001.

[4] Computed as the weighted average of "percent walking to work" for the columns pertaining to Tel Aviv in Plaut's Table 7, p. 246.

zero and one. Instead of explaining x directly, we can explain the logistic transformation of x as follows:

$$\log\left(\frac{x}{1-x}\right) = \beta' z + \varepsilon \tag{2.1}$$

where β is a vector of parameters, z is a vector of independent variables, and ε is an error term with infinite range. Equivalently,

$$x = \frac{\exp(\beta' z + \varepsilon)}{1 + \exp(\beta' z + \varepsilon)}. \tag{2.2}$$

This is an *aggregate logit* model for a single dependent variable.

In many applications, several dependent variables x_i are related to each other, each associated with particular values z_i of some independent variables. For example, x_i might be the share of trips made by mode i, and z_i a vector of service characteristics of mode i. If the characteristics in the z variables encompass all the systematic influences on mode shares, then a simple extension of equation (2.2) ensures that they sum to one:

$$x_i = \frac{\exp(\beta' z_i + \varepsilon_i)}{\sum_{j=1}^{J} \exp(\beta' z_j + \varepsilon_j)} \tag{2.3}$$

where J is the number of modes.[5] Anas (1981) and Mackett (1985b) use counts of interzonal flows to estimate models similar to this, in order to explain location and mode choice in Chicago and in Hertfordshire (England), respectively.

2.1.3 Studies Using Time Series or Panel Data

One can estimate demand equations from aggregate time-series data from a single area. For example, Greene (1992) uses US nationwide data on vehicle-miles traveled (VMT) to examine the effects of fuel prices. Several studies have examined transit ridership using data over time from a single metropolitan area or even a single transit corridor—for example Gaudry (1975) and Gómez-Ibáñez (1996) use this method to study Montreal and Boston, respectively. Time-series studies are quite sensitive to the handling of auto-correlation among the error terms, which refers to the tendency for unobserved influences on the measured dependent variable to persist over

[5] Eqn (2.2) is a special case of (2.3) in which J=2 and we define x=x_1, z=z_1-z_2 and ε=ε_1-ε_2.

time. They may also postulate "inertia" by including among the explanatory variables one or more lagged values of the variable being explained. For example, Greene considers the possibility that once people have established the travel patterns that produce a particular level of VMT, they change them only gradually if conditions such as fuel prices suddenly change. The coefficients on the lagged dependent variables then enable one to measure the difference between short- and long-run responses, but this measured difference is especially sensitive to the treatment of autocorrelation.

It is common to combine cross-sectional and time-series variation using *panel data*, also called *longitudinal data*. Such data often combine observations from many separate locations and two or more time periods. Kitamura (2000) provides a general review. For example, Voith (1997) analyzes ridership data from 118 commuter-rail stations in metropolitan Philadelphia over the years 1978–91 to ascertain the effects of level of service and of demographics on rail ridership. Surprisingly, he find that demographic characteristics have little independent effect; rather, he suggests that much of the observed correlation between demographic factors and rail ridership arises from reverse causation. For example, a neighborhood with good rail connections to the central business district (CBD) will attract residents who work in the CBD. A corollary of this finding is that the long-run effects of changes in fares or service levels, allowing for induced changes in residential location, are considerably greater than the short-run effects. Another study using panel data is that by Petitte (2001), who estimates fare elasticities from station-level data on Metrorail ridership in Washington, D.C.

Studies using panel data need to account for the fact that, even aside from autocorrelation, the error terms for observations from the same location at different points in time cannot plausibly be assumed to be independent. Neglecting this fact will result in an unnecessary loss of efficiency and an over-statement of the precision of the estimates; for nonlinear models, it may also bias the estimates. To account for the panel structure, at least three approaches are available. One is to "first difference" all variables, so that the variable explained describes *changes* in some quantity rather than the quantity itself; this reduces the size of the data set by N if N is the number of locations.[6] A second is to estimate a "fixed effects" model, in which a separate constant is estimated for every location; this retains all observations but adds

[6] A good example is the study of transit use in US cities by Baum-Snow and Kahn (2000).

N-1 new coefficients to be estimated (assuming one constant would be estimated in any case). A third is a "random effects" model, in which a separate random-error term is specified that varies only by location (not time), usually with an assumed normal distribution; this specification adds only one parameter to be estimated (the standard deviation of the new error term) and so is especially useful where only a few time periods are available. Statistical tests are available to determine whether the more restrictive random-effects model is justified. Voith (1997) uses both the first-difference and fixed-effects approaches.

2.1.4 Summary of Key Results

Several literature surveys have compiled estimates of summary measures such as own- and cross-elasticities of demand for auto or public transit with respect to cost and service quality. Service quality is typically proxied by annual vehicle-miles or vehicle-hours of service.

As a rough rule of thumb, a 10 percent increase in transit fare reduces transit demand by four percent: that is, transit's own-price elasticity is approximately -0.4 on average (Pratt *et al.* 2000: 12-9). This elasticity is higher for trips to a central business district, trips on buses (compared to urban rail), and off-peak trips (Lago et al. 1981). Goodwin (1992) and Pratt et al. (2000, ch. 12) provide thorough reviews, which also include some studies using the disaggregate techniques described later in this chapter. Transit elasticities with respect to service quality tend to be higher, especially where service quality is poor (Chan and Ou 1978).

Demand for automobile work trips is similarly more sensitive to service quality than to cost. For example, time- and cost-elasticities are measured at -0.8 and -0.5, respectively, for Boston; and at -0.4 and -0.1 for Louisville, Kentucky (Chan and Ou 1978: 43). Overall demand for travel in personal vehicles has more often been measured as a function of fuel price and/or fuel cost per mile; typical results show elasticities between 0.1 and 0.3, with short-run elasticities typically smaller (in absolute values) than long-run elasticities elasticities typically smaller (in absolute values) than long-run elasticities.[7] The demand for fuel itself is apparently two to three

[7] Luk and Hepburn (1993); Greening, Greene, and Difiglio (2000, section 3.1.4); Graham and Glaister (2002, Table 2). Johansson and Schipper (1997) estimate a useful breakdown of changes of per capita fuel consumption into changes in vehicle stock, average fuel economy, and average usage per vehicle.

times as large, indicating that changes in fuel price affect the composition of the motor-vehicle fleet more than its usage.

2.1.5 Travel Budgets

Some researchers have suggested that better predictions of travel behavior can be obtained by making use of certain regularities in the per-capita or per-household expenditures of time and money on travel. For example, Tanner (1961) notes that distances traveled by households varied only slightly across large and small urban areas and rural areas in Britain. Schafer (2000) finds similar stability across thirty data sets in more than ten nations. In a series of unpublished papers and reports, Yacov Zehavi has made these relationships the centerpiece of a model of travel demand known as Unified Mechanism of Travel (UMOT). No satisfactory theoretical foundation has ever been provided for the model, despite a heroic attempt by Golob, Beckmann, and Zehavi (1981).

The travel-budget approach is thoroughly reviewed by Gunn (1981) and Schafer (2000), who verify many intriguing regularities. Nevertheless, the evidence suggests that more conventional models can explain them and, furthermore, that the regularities are approximations and that violations of them occur as predicted by economic theory. For example, Kockelman (2001) statistically rejects the hypothesis of fixed travel-time budgets using data from the San Francisco Bay Area. She finds that total time spent traveling declines as nearby activities are made more accessible and as distant activities are made less accessible, the latter suggesting that substitution (of nearby for distant destinations) more than offsets the direct effects of increased travel time to distant locations.

2.1.6 Transportation and Land Use

Land-use patterns are among the most important factors influencing travel decisions. This fact has led to contentious policy debates as people seek to use land-use controls to solve transportation-related problems. Giuliano (2004) provides a thorough review.

Understanding the effects of land use on travel demand requires careful disentangling of related factors. First, one must control for variables, like income and fuel price, that independently affect travel and are also correlated with land-use characteristics. Some influential studies claiming far-reaching effects of urban density on travel, such Newman and Kenworthy

(1989), have been severely criticized for neglecting this fundamental requirement of statistical analysis.

Second, should one also control for variables measuring the provision of transportation infrastructure and services? Such provision is partly a response to travel choices and to that extent, including such variables as controls could lead to an overstatement of their effect on travel and an understatement of the effect of land-use patterns (since some of their effects would be attributed instead to the transportation service variables). One way to handle this is to control for specific policies that are deemed to be exogenous, such as federal cost-sharing formulas, rather than the actual amount of infrastructure or transit service. Another is to use such policies, or other variables correlated with infrastructure and services but not directly affecting travel itself, as instrumental variables.

Third, land-use patterns themselves are not entirely exogenous, but rather respond strongly to transportation systems, especially fixed infrastructure. As a result, transportation policies may cause unexpected feedback. For example, expanding highways to relieve congestion may attract development that undermines the intended effect; this is just one of many causes of "induced demand," discussed in Chapter 5. Expanding mass transit can even exacerbate highway congestion because the induced development, even if relatively transit-oriented, still generates many automobile trips. For example, the Bay Area Rapid Transit system in San Francisco is credited with causing Walnut Creek, an outlying station, to develop into a major center of office employment – but despite its good transit access, 95 percent of commuting trips to this center are by automobile (Cervero and Wu 1996, Table 5).

Fourth, land-use policy acts only indirectly on urban structure and thus it is misleading to treat land-use patterns as though they could be modified by fiat. Even in countries with very strong land-use authority, such as the Netherlands, land-use policies do not always bring about the changes that were intended (Schwanen, Dijst, and Dieleman, 2004). In others, such as the United States, the changes in land use that can feasibly be accomplished through policy are quite limited. As Downs (2004, ch. 12) points out, the urbanized area of Portland, Oregon, remains relatively low density despite three decades of stringent policies aimed at increasing urban density. One reason is documented by Jun (2004): Portland's policies have apparently diverted urban development to more outlying jurisdictions outside its control.

We turn now to empirical findings on how land use affects travel demand. At the level of an entire metropolitan area, modest effects have been documented. Within a cross-section of 49 US metropolitan areas, Keyes (1982) shows that per capita gasoline consumption rises with total urban population and with the fraction of jobs located in the central business district, and falls with the fraction of people living in high-density census tracts (defined as more than 10,000 people per square mile). Gordon, Kumar, and Richardson (1989) examine average commuting time among individual respondents to the 1980 National Personal Transportation Study. Commuting time, unlike gasoline consumption, goes *up* with residential density; presumably this reflects longer or more congested commutes. But commuting time goes *down* with the proportion of the metropolitan population that lives outside the central city. This last finding suggests that polycentric or dispersed land-use patterns enable people to bypass central congestion, which in turn may explain a paradox in commuting trends: even while congestion on particular roads has gotten worse over time, average commuting times have mostly not become longer (Gordon and Richardson, 1994).

Bento *et al.* (2004) define a number of land-use and transportation variables related to "urban sprawl," and ask how they influence travel choices across 114 US metropolitan areas. While no one factor explains very much of the variation, all of them taken together make a significant difference. To illustrate, the authors predict annual vehicle use for a national sample of households if they all lived in metropolitan areas with specified characteristics related to land use and transportation supply. If the characteristics are those of Atlanta, their model predicts 16,900 vehicle-miles per household; if the characteristics are changed to those of Boston, the same households would travel 25 percent less. While many characteristics contribute to this difference, the most important is population centrality, which indicates that a higher proportion of Boston's population than Atlanta's lives within the central portions of the total urbanized land area. The next two most important characteristics are Boston's higher supply of rail transit and its less circular shape, both of which favor greater use of public transit. Boston also has a higher urban population density and a more even balance between jobs and households at the zip-code level, both of which contribute more modestly to its lower vehicle use.

The results just cited control for transportation infrastructure and services. If one were to view such variables instead as consequences of travel decisions, and so omit them from the model, the effects of land use are even larger – by about two-thirds in Keyes' study.

European contexts may yield different results. Schwanen *et al*. (2004) examine the effects of urban size and form on travel for both commuting and shopping trips in The Netherlands. For any given mode (auto or transit), they find that the largest cities have the shortest and fastest commutes, medium-sized cities and outlying "growth centers" have the longest, while suburbs are in between. They also report exceptionally high use of walking and bicycling, amounting to 40 percent of work trips and 67 percent of shopping trips in the three largest cities, with lower rates in other areas. Specific policies in The Netherlands may explain some of these results, but also the local culture and habits, and possibly external economies of scale in cycling, will matter (and may in fact partly justify these policies).

Turning to the neighborhood level, the evidence on how land use affects travel is quite mixed. Crane (2000) provides a useful review. It is clear that high-density neighborhoods near transit stops support higher use of public transit. It is less clear how much of this is due simply to sorting of households: those who want to use transit choose transit-oriented neighborhoods. This is not a concern if one simply wants to predict transit patronage in a proposed transit-oriented development, but if one wants to know the aggregate effects of building many such developments one needs to control statistically for sorting. When this is done, the effects of land use on the travel behavior are found to be much more modest, and to depend on a number of collateral factors such as how centralized the entire urban area is. For example, Cervero and Gorham (1994) find that transit-oriented development encourages more walking and bicycling in local areas within the San Francisco region, but not in those within the Los Angeles region.

One of the factors limiting the influence of land use is that people travel far greater distances than are required by land-use patterns alone (Giuliano and Small, 1993). Even when jobs-housing balance is achieved within a community, people do not predominately choose nearby jobs (Giuliano, 1991; Cervero, 1996). Thus, for example, new exurban communities intended to be relatively self-contained have generally discovered that most residents work elsewhere, even in the Netherlands (Schwanen *et al.*, 2004).

Recent work has suggested that most of the effect of land use on travel can be understood by examining how trip costs are affected and how people respond to changes in trip costs (Boarnet and Crane, 2000). This observations opens the door to research explaining more specifically why land use has impacts in some settings and not others.

2.2 Disaggregate Models

An alternative approach, known as *disaggregate* or *behavioral* travel-demand modeling, is now far more common for travel demand research. Made possible by micro data (data on individual decision-making units), this approach explains behavior directly at the level of a person, household, or firm. Disaggregate models are more efficient in their use of survey data when it is available, and are based on a more satisfactory microeconomic theory of demand, a feature that is particularly useful when applying welfare economics. Most such models analyze choices among discrete rather than continuous alternatives and so are called *discrete-choice models*.[8]

Discrete-choice modeling of travel demand has mostly taken advantage of data from large and expensive transportation surveys. Deaton (1985) shows that it can also be used with household-expenditure surveys, which are often conducted for other purposes and are frequently available in developing nations.

2.2.1 Basic Discrete-Choice Models

The most widely used theoretical foundation for these models is the additive random-utility model of McFadden (1973). Suppose a decision maker n facing discrete alternatives j=1,...,J chooses the one that maximizes utility as given by

$$U_{jn} = V\left(z_{jn}, s_n, \beta\right) + \varepsilon_{jn} \tag{2.4}$$

where $V(\cdot)$ is a function known as the *systematic utility*, z_{jn} is a vector of attributes of the alternatives as they apply to this decision maker, s_n is a vector of characteristics of the decision maker (effectively allowing different utility structures for different identifiable groups of decision makers), β is a vector of unknown parameters, and ε_{jn} is an unobservable component of utility which captures idiosyncratic preferences. U_{jn} and V are known as *conditional indirect utility* functions, since they are conditional on choice j and, just like the indirect utility function of standard consumer theory, they implicitly incorporate a budget constraint.

The choice is probabilistic because the measured variables do not include everything relevant to the individual's decision. This fact is represented by the random terms ε_{jn}. Once a

[8] Reviews with a transportation focus include Ben-Akiva and Bierlaire (1999), Koppelman and Sethi (2000), and the books by Ben-Akiva and Lerman (1985) and Train (2003).

functional form for V is specified, the model becomes complete by specifying a joint cumulative distribution function (cdf) for the random terms, $F(\varepsilon_{1n},...,\varepsilon_{Jn})$. Denoting $V(z_{jn},s_n,\beta)$ by V_{jn}, the choice probability for alternative i is then

$$\begin{aligned} P_{in} &= \Pr[U_{in} > U_{jn} \qquad \text{for all } j \neq i] \\ &= \Pr[\varepsilon_{jn} - \varepsilon_{in} < V_{in} - V_{jn} \qquad \text{for all } j \neq i] \\ &= \int_{-\infty}^{\infty} F_i(V_{in} - V_{1n} + \varepsilon_{in}, \ldots, V_{in} - V_{Jn} + \varepsilon_{in}) d\varepsilon_{in} \end{aligned} \tag{2.5}$$

where F_i is the partial derivative of F with respect to its i-th argument. (F_i is thus the probability density function of ε_{in} conditional on the inequalities in (2.5).)

Suppose the cdf $F(\cdot)$ is multivariate normal. Then (2.5) is the *multinomial probit* model with general covariance structure. However, neither F nor F_i can be expressed in closed form; instead, equation (2.5) is usually written as a (J-1)-dimensional integral of the normal density function. In the special case where the random terms are identically and independently distributed (iid) with the univariate normal distribution, F is the product of J univariate normal cdfs, and we have the *iid probit* model, which still requires computation of a (J-1)-dimensional integral. For example, in the iid probit model for binary choice (J=2), (2.5) becomes

$$P_{1n} = \Phi\left(\frac{V_{1n} - V_{2n}}{\sigma}\right) \tag{2.6}$$

where Φ is the cumulative standard normal distribution function (a one-dimensional integral) and σ is the standard deviation of ε_{1n}-ε_{2n}. In equation (2.6), σ cannot be distinguished empirically from the scale of utility, which is arbitrary; for example, doubling σ has the same effect as doubling both V_1 and V_2. Hence it is conventional to normalize by setting σ=1.

The *logit* model (also known as multinomial logit or conditional logit) arises when the J random terms are iid with the extreme-value distribution (also known as Gumbel, Weibull, or double-exponential). This distribution is defined by

$$\Pr[\varepsilon_{jn} < x] = \exp(-e^{-\mu x}) \tag{2.7}$$

for all real numbers x, where μ is a scale parameter. Here the convention is to normalize by setting μ=1. With this normalization, McFadden (1973) shows that the resulting probabilities calculated from (2.5) have the logit form:

$$P_{in} = \frac{\exp(V_{in})}{\sum_{j=1}^{J} \exp(V_{jn})}. \tag{2.8}$$

This formula is easily seen to have the celebrated and restrictive property of *independence from irrelevant alternatives*: namely, that the odds ratio (P_{in}/P_{jn}) depends on the utilities V_{in} and V_{jn} but not on the utilities for any other alternatives. This property implies, for example, that adding a new alternative k (equivalent to increasing its systematic utility V_{kn} from $-\infty$ to some finite value) will not affect the relative proportions of people using previously existing alternatives. It also implies that for a given alternative k, the cross-elasticities $\partial \log P_{jn}/\partial \log V_{kn}$ are identical for all $j \neq k$: hence if the attractiveness of alternative k is increased, the probabilities of all the other alternatives $j \neq k$ will be reduced by identical percentages.

The binary form of (2.8), i.e. the form with J=2, is:

$$P_{in} = \frac{1}{1+\exp\left[-\left(V_{1n}-V_{2n}\right)\right]}.$$

If graphed as a function of (V_{1n}-V_{2n}), this equation looks quite similar to (2.6).

It is really the iid assumption (identically and independently distributed error terms) that is restrictive, whether or not it entails independence of irrelevant alternatives. Hence there is no basis for the widespread belief that iid probit is more general than logit. In fact, the logit and iid probit models have been found empirically to give virtually identical results when normalized comparably (Horowitz, 1980).[9] Furthermore, both probit and logit may be generalized by defining non-iid distributions. In the probit case the generalization uses the multivariate normal distribution, whereas in the logit case it can take a number of forms to be discussed later.

As for the functional form of V, by far the most common is linear in unknown parameters β. More general forms such as Box-Cox and Box-Tukey transformations are studied by Gaudry and Wills (1978). Note that even with V forced to be linear in *parameters*, it can easily be made

[9] Comparable normalization is accomplished by dividing the logit coefficients by $\pi/\sqrt{3}$ in order to give the utilities the same standard deviations in the two models. In both models, the choice probabilities depend on (β/σ_ε), where σ_ε^2 is the variance of each of the random terms ε_{in}. In the case of probit, the variance of ε_{1n}-ε_{2n}, $2\sigma_\varepsilon^2$, is set to one by the conventional normalization; hence $\sigma_\varepsilon^{PROBIT} = 1/\sqrt{2}$. In the case of logit, the normalization μ=1 in equation (2.7) implies that ε_{in} has standard deviation $\sigma_\varepsilon^{LOGIT} = \pi/\sqrt{6}$ (Hastings and Peacock, 1975, p. 60). Hence to make logit and iid probit comparable, the logit coefficients must be divided by $\sigma_\varepsilon^{LOGIT} / \sigma_\varepsilon^{PROBIT} = \pi/\sqrt{3} = 1.814$.

nonlinear in *variables* just by specifying new variables equal to nonlinear functions of the original ones. For example, the utility on mode i of a traveler n with wage w_n facing travel costs c_{in} and times T_{in} could be:

$$V_{in} = \beta_1 \cdot (c_{in} / w_n) + \beta_2 T_{in} + \beta_3 T_{in}^2 . \tag{2.9}$$

This is non-linear in travel time and in wage rate. If we redefine z_{in} as the vector of all such combinations of the original variables (z_{in} and s_n in eqn 2.4), the linear-in-parameters specification is simply written as

$$V_{in} = \beta' z_{in} \tag{2.10}$$

where β' is the transpose of column vector β.

2.2.2 Estimation

For a given model, data on actual choices, along with traits z_{jn}, can be used to estimate the unknown parameter vector β in (2.10) and to carry out statistical tests of the specification (i.e., tests of whether the assumed functional form of V and the assumed error distribution are valid). Parameters are usually estimated by maximizing the log-likelihood function:

$$L(\beta) = \sum_{n=1}^{N} \sum_{i=1}^{J} d_{in} \log P_{in}(\beta) \tag{2.11}$$

where N is the sample size. In this equation, d_{in} is the choice variable, defined as 1 if decision-maker n chooses alternative i and 0 otherwise, and $P_{in}(\beta)$ is the choice probability.

A correction to (2.11) is available for choice-based samples, i.e., those in which the sampling frequencies depend on the choices made. The correction simply multiplies each term in the second summation by the inverse of the sampling probability for that sample member (Manski and Lerman, 1977). This correction does not, however, make efficient use of the information on aggregate mode shares that it requires. Imbens and Lancaster (1994) show quite generally how to incorporate aggregate information to greatly improve the efficiency of disaggregate econometric models, while Berry, Levinsohn, and Pakes (1995) show that with assumptions about firm behavior it is even possible even to estimate the parameters of a micro-level model using only aggregate data.

One of the major attractions of logit is the computational simplicity of its log-likelihood function, due to taking the logarithm of the numerator in equation (2.8). With V linear in β, the logit log-likelihood function is globally concave in β, so finding a local maximum assures finding the global maximum. Fast computer routines to do this are widely available. In contrast, computing the log-likelihood function for multinomial probit with J alternatives entails computing for each member of the sample the (J-1)-dimensional integral implicit in equation (2.5). This has generally proven difficult for J larger than 3 or 4, despite the development of computational-intensive simulation methods (Train 2003).

It is possible that the likelihood function is unbounded in one of the coefficients, making it impossible to maximize. This happens if one includes a variable that is a perfect predictor of choice within the sample. For example, suppose one is predicting car ownership (yes or no) and wants to include among variables s_n in (2.4) a dummy variable for high income. If it happens that within the sample everyone with high income owns a car, the likelihood function increases without limit in the coefficient of this dummy variable. The problem is that income does too good a job as an explanatory variable: within this data set, the model exuberantly declares high income to make the alternative of owning a car infinitely desirable relative to not owning one. We know of course that this is not true and that a larger sample would contain counter-examples—even in the US, 1.5 percent of the highest-income households owned no car in 2001 (Pucher 2003). Given the sample we have, we might solve the problem by respecifying the model with more broadly defined income groups or more narrowly defined alternatives. Alternatively, we could postulate a *linear probability model*, in which probability rather than utility is a linear function of coefficients; despite certain statistical disadvantages, this model is able to measure the coefficient in question (Caudill, 1988) because there is a limit to how much income can affect probability.

2.2.3 Data

Some of the most important variables for travel demand modeling are determined endogenously within a larger model of which the demand model is just one component. The most common example is that travel times depend on congestion, which depends on amount of travel, which depends on travel times. Thus the actual use of a travel demand model may require a process of *equilibration* in which a solution is sought to a set of simultaneous relationships. An

elegant formulation of supply-demand equilibration on a congested network is provided in the remarkable study by Beckmann, McGuire, and Winsten (1956). Boyce, Mahmassani, and Nagurney (2005) provide a readable review of its history and subsequent impact.

With aggregate data, the endogeneity of travel characteristics is an important issue for obtaining valid statistical estimates of demand parameters. Fortunately, endogeneity can usually be ignored when using disaggregate data because, from the point of view of individual decision-making, the travel environment does not depend appreciably on that one individual's decisions. Nevertheless, measuring the values of attributes z_{in}, which typically vary by alternative, is more difficult than it may first appear. How does one know the traits that a traveler would have encountered on an alternative that was not in fact used?

One possibility is to use objective estimates, such as the *engineering values* produced by network models of the transportation system. Another is to use *reported values* obtained directly from survey respondents. Each is subject to problems. Reported values measure people's perceptions of travel conditions, which, even for alternatives they choose regularly, may be quite different from the measures employed in policy analysis or forecasting. People know even less about alternatives they do not choose. Hence even if reported values accurately measure the perceptions that determine choice, the resulting models cannot be used for prediction unless one can predict how a given change will alter those perceptions. Worse still, the reports may be systematically biased so as to justify the choice, thereby exaggerating the advantages of the alternative chosen and the disadvantages of other alternatives. The study by MVA Consultancy *et al*. (1987, pp. 159-163) finds such bias to be severe in a study of the Tyne River crossing in England. In this case the explanatory variables of the model are endogenous to the choice, which makes the estimated model appear to fit very well (a typical finding for studies using reported values) but which renders it useless for prediction.

Objective estimates of travel attributes, on the other hand, may be very expensive and not necessarily accurate. Even something as simple as the travel time for driving on a particular highway segment at a particular time of day is quite difficult to ascertain. Measuring the day-to-day variability of that travel time is even more difficult. Three recent studies in California have accomplished this, one by applying sophisticated algorithms to data from loop detectors placed

in the highway and two by using the floating-car method, in which a vehicle with a stopwatch is driven so as to blend in with the traffic stream.[10]

Ideally, one might formulate a model in which perceived attributes and actual choice are jointly determined, each influencing the other and both influenced by objective attributes and personal characteristics. This type of model most faithfully replicates the actual decision process. However, it is doubtful that the results would be worth the extra complexity unless there is inherent interest in perception formation for marketing purposes. For purposes of transportation planning, we care mainly about the relationship between objective values and actual choices. A model limited to this relationship may be interpreted as the reduced form of a more complex model including perceptions, so it is theoretically valid even though perception formation is only implicit. Hence the most fruitful expenditure of research effort is usually on finding ways to measure objective values as accurately as possible.

In a large sample, a cheaper way to compute objective values may be to assign values for a given alternative according to averages reported by people in the sample in similar circumstances who use that alternative. While subject to some inaccuracy, this at least eliminates endogeneity bias by using an identical procedure to assign values to chosen and unchosen alternatives.

The data described thus far measures *revealed preference* (RP) information, that reflected in actual choices. There is growing interest in using *stated preference* (SP) data, based on responses to hypothetical situations (Hensher, 1994). SP data permit more control over the ranges of and correlations among the independent variables by applying an appropriate experimental design (see for example Louviere, Hensher, and Swait, 2000). If administered in interviews using a portable computer, the questions posed can be adapted to information about the respondent collected in an earlier portion of the survey – as for example in the study of freight mode choice in India by Shinghal and Fowkes (2002). SP surveys also can elicit information about potential travel options not now available. It is still an open question, however, how accurately they describe what people really do.

It is possible to combine data from both revealed and stated preferences in a single estimation procedure in order to take advantage of the strengths of each (Ben-Akiva and

[10] Brownstone *et al*. (2003); Lam and Small (2001); Small, Winston, and Yan (2005).

Morikawa, 1990; Louviere and Hensher, 2001). So long as observations are independent of each other, the log-likelihood functions simply add. To prevent SP survey bias from contaminating inferences from RP, or more generally just to account for differences in surveys, it is recommended to estimate certain parameters separately in the two portions of the data: the scale factors μ for the two parts of the sample (with one but not both normalized), any alternative-specific constants, and any critical behavioral coefficients that may differ. For example, in the logit model of (2.8) and (2.12), one might constrain all parameters to be the same for RP and SP observations except for the scale, alternative-specific constants, and the first variable z_{1in}. Letting $\beta_2' z_{2in}$ represent the rest of $\beta' z_{in}$, adding superscripts for the parameters assumed distinct in the two data subsamples, and normalizing the RP scale parameter to one, the log-likelihood function (2.11) becomes the following:

$$L(\alpha,\beta,\mu^{SP}) = \sum_{n\in RP}\sum_{i=1}^{J} d_{in}\left\{\alpha_i^{RP} + \beta_1^{RP} z_{1in} + \beta_2' z_{2in} - \log\sum_{j=1}^{J}\exp(\alpha_j^{RP} + \beta_1^{RP} z_{1jn} + \beta_2' z_{2jn})\right\}$$

$$+ \sum_{n\in SP}\sum_{i=1}^{J} d_{in}\left\{\mu^{SP}\cdot(\alpha_i^{SP} + \beta_1^{SP} z_{1in} + \beta_2' z_{2in}) - \log\sum_{j=1}^{J}\exp\left[\mu^{SP}\cdot(\alpha_j^{SP} + \beta_1^{SP} z_{1jn} + \beta_2' z_{2jn})\right]\right\}$$

where (α,β,μ^{SP}) denotes the entire set of parameters shown on the right-hand side (excluding α_1^{RP} and α_1^{RP}, which can be normalized to zero). As before, the prime after a column vector indicates its transpose, so that it becomes a row vector. This expression is not as complicated as it looks: the first term in curly brackets is just the logarithm of the logit probability (2.8) for RP observations, while the second is the same thing for SP observations except utility V_{in} is multiplied by scale factor μ^{SP}.

2.2.4 Interpreting Coefficient Estimates

It is useful for interpreting empirical results to note that a change in $\beta' z_{in}$ in (2.10) by an amount of ± 1 increases or decreases the relative odds of alternative *i*, compared to each other alternative, by a factor exp(1)=2.72. Thus a quick gauge of the behavioral significance of any particular variable can be obtained by considering the size of typical variations in that variable, multiplied by its relevant coefficient – if the result is on the order of 1.0 or larger, such variations have large effects on the relative odds. In fact some authors prefer to provide this information by listing, in addition to or instead of the coefficient estimates, the marginal effect of a specified change in the

independent variable on the probabilities; a disadvantage of this measure is that it depends on the values of the variables and so makes comparisons across models more difficult.

The parameter vector may contain *alternative-specific constants* for one or more alternatives i. That is, the systematic utility may be of the form

$$V_{in} = \alpha_i + \beta' z_{in} . \tag{2.12}$$

Since only utility differences matter, at least one of the alternative-specific constants must be normalized (usually to zero); that alternative then serves as a "base alternative" for comparisons.

The constant α_i may be interpreted as the average utility of the unobserved characteristics of the i-th alternative, relative to the base alternative. In a sense, specifying these constants is admitting the inadequacy of variables z_{in} to explain choice; hence the constants' estimated values are especially likely to reflect circumstances of a particular sample rather than universal behavior. The use of alternative-specific constants also makes it impossible to forecast the result of adding a new alternative, unless there is some basis for a guess as to what its alternative-specific constant would be. Quandt and Baumol (1966) coined the term "abstract mode" to indicate the desire to describe a travel mode entirely by its objective characteristics, rather than relying on alternative-specific constants. In practice, however, this goal is rarely achieved.

Equation (2.12) is really a special case of (2.10) in which one or more of the variables Z are *alternative-specific dummy variables*, D^k, defined by D^k_{jn} =1 if $j=k$ and 0 otherwise (for each $j=1,\ldots,J$). (Such a variable does not depend on n.) In this notation, parameter α_i in (2.12) is viewed as the coefficient of variable D^i included among the z variables in (2.10). Such dummy variables can also be interacted with (i.e., multiplied by) any other variable, making it possible for the latter variable to affect utility in a different way for each alternative. All such variables and interactions may be included in z, and their coefficients in β, thus allowing (2.10) still to represent the linear-in-parameters specification. An SP experiment designed such that alternative-specific coefficients can be estimated for all attributes is sometimes called a "labeled" experiment.

The most economically meaningful quantities obtained from estimating a discrete-choice model are often ratios of coefficients, which represent marginal rates of substitution. By interacting the variables of interest with socioeconomic characteristics or alternative-specific constants, these ratios can be specified quite flexibly so as to vary in a manner thought to be *a*

priori plausible. A particularly important example is the marginal rate of substitution between time and money in the conditional indirect utility function, often called the *value of travel-time savings*, or *value of time* for short. It represents the monetary value that the traveler places on time savings, and is very important in evaluating the benefits of transportation improvements whose primary effects are to improve people's mobility. The value of time in the model (2.9) is

$$(v_T)_{in} \equiv -\left(\frac{dc_{in}}{dT_{in}}\right)_{V_{in}} \equiv \frac{\partial V_{in}/\partial T_{in}}{\partial V_{in}/\partial c_{in}} = \left(\frac{\beta_2 + 2\beta_3 T_{in}}{\beta_1}\right)\cdot w_n, \tag{2.13}$$

which varies across individuals since it depends on w_n and T_{in}.

As a more complex example, suppose we extend equation (2.9) by adding alternative-specific dummies, both separately (with coefficients α_i) and interacted with travel time (with coefficients γ_i):

$$V_{in} = \alpha_i + \beta_1 \cdot (c_{in}/w_n) + \beta_2 T_{in} + \beta_3 T_{in}^2 + \gamma_i T_{in} \tag{2.14}$$

where one of the α_i and one of the γ_i are normalized to zero. This yields the following value of time applicable when individual n chooses alternative i:

$$(v_T)_{in} = \left(\frac{\beta_2 + 2\beta_3 T_{in} + \gamma_i}{\beta_1}\right)\cdot w_n. \tag{2.15}$$

Now the value of time varies across modes even with identical travel times, due to the presence of γ_i. There is a danger, however, in interpreting such a model. What appears to be variation in value of time across modes may just reflect selection bias: people who, for reasons we cannot observe, have high values of time will tend to self-select onto the faster modes (MVA Consultancy et al., 1987, pp. 90-92). This possibility can be modeled explicitly using a random-coefficient model, described later in this chapter.

Confidence bounds for a ratio of coefficients, or for more complex functions of coefficients, can be estimated by standard approximations for transformations of normal variates. Specifically, if vector β is asymptotically normally distributed with mean b and variance-covariance matrix Σ, then a function $f(\beta\text{-}b)$ is asymptotically normally distributed with mean zero

and variance-covariance matrix $(\nabla f)\Sigma(\nabla f)'$, where ∇f is the vector of partial derivatives of f.[11] A more accurate estimate may be obtained by taking repeated random draws β^r from the distribution of β, which is estimated along with β itself, and then examining the resulting values $f(\beta^r)$. As an example, the 5th and 95th percentile values of those values define a 90 percent confidence interval for β. See Train (2003, ch. 9) for how to take such random draws.

2.2.5 Randomness, Scale of Utility, and Measures of Benefit

The variance of the random utility term in equation (2.4) reflects randomness in behavior of individuals or, more likely, heterogeneity among observationally identical individuals. Hence it plays a key role in determining how sensitive travel behavior is to observable quantities such as price, service quality, and demographic traits. Little randomness implies a nearly deterministic model, one in which behavior suddenly changes at some crucial switching point (for example, when transit service becomes as fast as a car). Conversely, if there is a lot of randomness, behavior changes only gradually as the values of independent variables are varied.

When the variance of the random component is normalized, however, the degree of randomness becomes represented by the inverse of the scale of the systematic utility function. For example, in the logit model (2.8), suppose systematic utility is linear in parameter vector β as in (2.10). If all the elements of β are small in magnitude, the corresponding variables have little effect on probabilities so choices are dominated by randomness. If the elements of β are large, most of the variation in choice behavior is explained by variation in the observable variables.

Randomness in individual behavior can also be viewed as producing variety, or *entropy*, in aggregate behavior. Indeed, it can be measured by the entropy-like quantity $-\Sigma_n\Sigma_j P_{jn}\log P_{jn}$, which is larger when the choice probability is divided evenly among the alternatives than when one alternative is very likely and others very unlikely. Anas (1983) derives the disaggregate logit model by maximizing an aggregate objective function that includes such an entropy term, subject

[11] Chow (1983: 182-3). The result requires asymptotic convergence of β at a rate proportional to the square root of the sample size. (That is, as the sample size N increases, the difference between the estimated and true values of β tends to diminish proportionally to $1/\sqrt{N}$.) In the simple case where $v_T=\beta_2/\beta_1$, it implies that the standard deviation σ_v of v_T obeys the intuitive formula: $(\sigma_v/v_T)^2 \cong (\sigma_1/\beta_1)^2 + (\sigma_2/\beta_2)^2 - 2\sigma_{12}/(\beta_1\beta_2)$, where σ_1 and σ_2 are the standard deviations of β_1 and β_2 and where σ_{12} is their covariance. Such an approximation requires that σ_1/β_1 and σ_2/β_2 be small, which in turn helps ensure that the variance of β_1/β_2 exists (it would not exist, for examle, if the mean of β_2 wre zero).

to constraints that guarantee consistency with observed aggregate shares and average values of characteristics. Similarly, Anderson et al. (1988) show that the aggregate logit model can be derived by maximizing a utility function for a representative traveler that includes an entropy term, subject to a consistency constraint on aggregate choice shares. Thus entropy is a link between aggregate and disaggregate models: at the aggregate level we can say the system tends to favor entropy or that a representative consumer craves variety, whereas at the disaggregate level we represent the same phenomenon as randomness in utility.

It is sometimes useful to have a measure of the overall desirability of the choice set being offered to a decision maker. Such a measure must account both for the utility of the individual choices being offered and for the variety of choices offered. The value of variety is directly related to randomness because both arise from unobserved idiosyncrasies in preferences. If choice were deterministic, i.e. determined solely by the ranking of V_{in} across alternatives i, the decision maker would care only about the traits of the best alternative; improving or offering inferior alternatives would have no value. But with random utilities, there is some chance that an alternative with a low value of V_{in} will nevertheless be chosen; so it is desirable for such an alternative to be offered and to be made as attractive as possible. A natural measure of the desirability of choice set J is the expected maximum utility of that set, which for the logit model has the convenient form:

$$E \max_{\mathrm{j}} (V_j + \varepsilon_j) = \mu^{-1} \log \sum_{j=1}^{J} \exp(\mu V_j) + \gamma \tag{2.16}$$

where γ=0.5772 is Euler's constant (it accounts for the nonzero mean of the error terms ε_j in the standard normalization). Here we have retained the parameter μ from (2.7), rather than normalizing it, to make clear how randomness affects expected utility. When the amount of randomness is small (large μ), the summation on the right-hand side is dominated by its largest term (let's denote its index by j*); expected utility is then approximately $\rho \cdot \log[\exp(V_{j*}/\rho)] = V_{j*}$, the utility of the dominating alternative. When randomness dominates (small μ), all terms contribute more or less equally (let's denote their average utility value by V); then expected utility is approximately $\mu^{-1} \cdot \log[J \cdot \exp(\mu V)] = V + \mu^{-1} \cdot \log(J)$, which is the average utility plus a term reflecting the desirability of having many choices.

Expected utility is, naturally enough, directly related to measures of consumer welfare. Small and Rosen (1981) show that, in the absence of income effects, changes in aggregate

consumer surplus (the area to the left of the demand curve and above the current price) are appropriate measures of welfare even when the demand curve is generated by a set of individuals making discrete choices. For a set of individuals n characterized by systematic utilities V_{jn}, changes in consumer surplus are proportional to changes in this expected maximum utility. The proportionality constant is the inverse of λ_n, the marginal utility of income; thus a useful welfare measure for such a set of individuals, with normalization μ=1, is:

$$W = \frac{1}{\lambda_n} \log \sum_{j=1}^{J} \exp(V_{jn}), \tag{2.17}$$

a formula also derived by Williams (1977). (The constant γ drops out of welfare comparisons so is omitted.) Because portions of the utility V_i that are common to all alternatives cannot be estimated from the choice model, λ_n cannot be estimated directly.[12] However, typically it can be determined from Roy's Identity:

$$\lambda_n = -\frac{1}{x_{in}} \cdot \frac{\partial V_{in}}{\partial c_{in}} \tag{2.18}$$

where x_{in} is consumption of good i conditional on choosing it among the discrete alternatives. In the case of commuting-mode choice, for example, x_{in} is just the individual's number of work trips per year (assuming income and hence welfare are measured in annual units). Expression (2.18) is valid provided that its right-hand-side is independent of i; when it is not, tractable approximations are available (Chattopadhyay, 2001).

2.2.6 Aggregation and Forecasting

Once we have estimated a disaggregate travel-demand model, we face the question of how to predict aggregate quantities such as total transit ridership or total travel flows between zones. Ben-Akiva and Lerman (1985, chap. 6) discuss several methods.

The most straightforward and common is *sample enumeration*. A sample of decision makers is drawn, each assumed to represent a subpopulation with identical observable

[12] If income y is included as an explanatory variable, it might be tempting to simply compute $\partial V_j/\partial y$ as a measure of λ. This is completely wrong because the indirect utility is strongly influenced by income, independently of which alternative is chosen, whereas V_j captures only the *relative* effects of income on utility of the various alternatives.

characteristics. (The estimation sample itself may satisfy this criterion and hence be usable as an enumeration sample.) Each individual's choice probabilities, computed using the estimated parameters, predict the shares of that subpopulation choosing the various alternatives. These predictions can then simply be added, weighting each sample member according to the corresponding subpopulation size. Standard deviations of forecast values can be estimated by Monte Carlo simulation methods.

One can simulate the effects of a policy by determining how it changes the values of independent variables for each sample member, and recomputing the predicted probabilities accordingly. Doing so requires that these variables be explicitly included in the model. For example, to simulate the effect of better schedule coordination at transfer points on a transit system, the model must include a variable for waiting time at the transfer points. Such a specification is called *policy-sensitive*, and its absence in earlier aggregate models was one of the main objections to the traditional travel-demand modeling framework. The ability to examine complex policies by computing their effects on an enumeration sample is one of the major advantages of disaggregate models.

Aggregate forecasts may display a sensitivity to policy variables that is quite different from a naive calculation based on a representative individual. For example, suppose the choice between travel by automobile (alternative 1) and bus (alternative 2) is determined by a logit model with utilities given by equation (2.9) with β_3=0. Then the probability of choosing bus travel is:

$$P_{2n} = \frac{1}{1+\exp\left[\left(\beta_1 / w_n\right)\cdot\left(c_{1n}-c_{2n}\right)+\beta_2\cdot\left(T_{1n}-T_{2n}\right)\right]}\,. \tag{2.19}$$

Suppose everyone's bus fare is c_2 and everyone's wage is w. Then

$$\frac{\partial P_{2n}}{\partial c_2} = \left(\beta_1 / w\right)\cdot P_{2n}\cdot\left(1-P_{2n}\right). \tag{2.20}$$

Now suppose half the population has conditions favorable to bus travel, such that P_{2n}=0.9; whereas the other half has P_{2n}=0.1. Aggregate bus share is then 0.5. The rate of change of aggregate bus share with respect to bus fare is, from (2.20), $(\beta_1/w)\cdot[½(0.9)(0.1)+½(0.1)(0.9)] = 0.09\cdot(\beta_1/w)$. But if we were to calculate it from (2.20) as though there were a single representative traveler with P_2=0.5, we would get $(\beta_1/w)(0.5)(0.5) = 0.25(\beta_1/w)$. This would

overestimate the true sensitivity by 178 percent. Again, the existence of variety reduces the actual sensitivity to changes in independent variables, in this case because there are only a few travelers (those with extreme values of $\varepsilon_{1n}-\varepsilon_{2n}$) who have a close enough decision to be affected.

McFadden and Reid (1976) derive an especially illuminating result illustrating this phenomenon in the case of a binary probit model where the independent variables are normally distributed in the population. They show that if a single individual's choice probability (2.6) is written in the form

$$P_1 = \Phi(\beta' z),$$

then the expected aggregate share $\overline{P}_1$ for the subpopulation represented by this individual is given by

$$\overline{P}_1 = \Phi\left(\frac{\beta'\bar{z}}{\sqrt{1+\sigma^2}}\right) \tag{2.21}$$

where $\bar{z}$ and σ^2 are the average of z and the variance of $\beta' z$, respectively, within this subpopulation. Once again, the existence of population variance reduces policy sensitivity and causes the naive calculation using an average traveler (equivalent to setting σ=0) to overestimate that sensitivity.

Equation (2.21) illustrates a danger in using aggregate models for policy forecasts. If an aggregate probit model fitting $\overline{P}_1$ to $\bar{z}$ were estimated, its coefficients would correspond to $\beta/\sqrt{(1+\sigma^2}$. If a policy being investigated changed σ, these coefficients would no longer accurately represent behavior under the new policy.

2.2.7 Specification Searches

Like most applied statistical work, travel demand analysis requires balancing completeness against tractability. The model that includes every relevant influence on behavior may require too much data to estimate with adequate precision, or it may be too complex to serve as a practical guide to policy analysis. A related problem, also common to most empirical work, is that the statistical properties of the model, such as standard errors of estimated coefficients, are valid only when the model's basic assumptions are known in advance to be correct. But in practice the researcher normally chooses a model's specification (i.e. its functional form and set

of included variables) using guidance from the same data as those from which its parameters are then estimated.

A good way to handle both problems is to base empirical models on an explicit behavioral theory. Rather than try out dozens of specifications to see what fits, one gives preference to relationships that are predicted by a plausible theory. For example, a specification like (2.9) would be chosen if there is good theoretical reason to think the value of time is proportional to the wage rate—a question explored later in this chapter.

Bayesian methods offer a more formal approach to using prior information or judgments when specifying empirical models. Instead of all-or-nothing decisions about model structure, they allow one to explicitly describe prior uncertainty and to calculate the manner in which prior beliefs need to be modified in light of the data. Such methods have recently been developed for parameter estimation in discrete choice models (Train 2003, ch. 12). Bayesian methods for model selection, in which the data (along with prior beliefs) determine explicit probabilities for competing model structures, are also available and could be usefully applied to transportation problems. See Berger and Pericchi (2001) for a good introduction.

2.2.8 Transferability

One of the goals of disaggregate travel-demand modeling is to describe behavioral tendencies that are reasonably general. This would enable a model estimated in one time and place to be used for another. The progress toward this goal has been disappointing, but some limited success has been achieved by making certain adjustments. Notably, the alternative-specific constants and the scale of the utility function are often found to be different in a new location, presumably because they reflect our degree of ignorance which may vary from one setting to another. Such adjustments can be made relatively inexpensively by using limited data collection in a new location or, in the case of alternative-specific constants, just by adjusting them to match known aggregate shares (Koppelman and Wilmot, 1982; Koppelman and Rose, 1985).

2.2.9 Ordered and Rank-Ordered Models

Sometimes there is a natural ordering to the alternatives that can be exploited to guide specification. For example, suppose one wants to explain a household's choice among owning no vehicle, one vehicle, or two or more vehicles. It is perhaps plausible that there is a single index

of propensity to own many vehicles, and that this index is determined in part by observable variables like household size and employment status.

In such a case, an *ordered response* model might be assumed. In this model, the choice of individual n is determined by the size of a "latent variable" $y_n^* = \beta' z_n + \varepsilon_n$, with choice j occurring if this latent variable falls in a particular interval $[\mu_{j-1},\mu_j]$ of the real line, where $\mu_0=-\infty$ and $\mu_J=\infty$. The interval boundaries $\mu_1,...,\mu_{J-1}$ are estimated along with β, except that one of them can be normalized arbitrarily if $\beta' z_n$ contains a constant term. The probability of choice j is then

$$P_{jn} = \Pr[\mu_{j-1} < \beta' z_n + \varepsilon_n < \mu_j] = F(\mu_j - \beta' z_n) - F(\mu_{j-1} - \beta' z_n) \qquad (2.22)$$

where $F(\cdot)$ is the cumulative distribution function assumed for ε_n. In the *ordered probit* model $F(\cdot)$ is standard normal, while in the *ordered logit* model it is logistic, i.e. $F(x) = \left[1 + \exp(-x)\right]^{-1}$. Thus probabilities depend entirely on a single index, $\beta' z_n$, calculated for individual n. When this index is strongly positive, all the terms $F(\mu_j-\beta' z_n)$ are small except for the last, $F(\mu_J-\beta' z_n) = F(\infty) = 1$, so the most likely choice will be alternative J. When the index is strongly negative, the most likely choice will be alternative 1. At intermediate values it becomes more likely that alternatives between 1 and J will be chosen. Note that all the variables in this model are characteristics of individuals, not of the alternatives, and thus if the latter information is available this model cannot easily take advantage of it.

In some cases the alternatives are integers indicating the number of times some random event occurs. An example would be the number of trips per month by a given household to a particular destination. For such cases, a set of models based on Poisson and negative binomial regressions is available (Washington, Karlaftis, and Mannering, 2003, ch. 10).

Sometimes information is available not only on the most preferred alternative, but on the individual's ranking of other alternatives. In this case, we effectively observe "choices" among numerous situations, including some where the most preferred alternative is hypothetically absent. Efficient use can be made of such data through the *rank-ordered logit* model analyzed by Beggs, Cardell, and Hausman (1981) and Hausman and Ruud (1987).[13] In the case where a complete ranking of J alternatives is obtained, the probability formula for rank-ordered logit is a product of J logit probability formulas, one for each ranked alternative, giving the probability of

[13] Rank-ordered logit is sometimes called "expanded logit" or "exploded logit." Beggs *et al.* call it "ordered logit," but that name is now usually reserved for an ordered response model as described here.

choosing that alternative from the set of itself and all lower-ranked alternatives. One may want to ignore the stated ordering among some low-ranked alternatives, or alternatively to estimate a separate scale factor for those choices, to allow for the possibility that a respondent pays less attention when answering questions about alternatives of little interest.

2.3 Examples of Disaggregate Models

Discrete-choice models have been estimated for nearly every conceivable travel decision, forming a body of research that cannot possibly be reviewed here.[14] In some cases, these models have been linked into large simultaneous systems requiring extensive computer simulation. An example is the system of models developed to analyze a proposal for congestion pricing in London (Bates *et al.* 1996).

In this section we present three very modest disaggregate models, each chosen for its compact representation of a behavioral factor that is central to urban transportation policy as analyzed in later chapters.

2.3.1 Mode Choice

Kenneth Train (1978, 1980) and colleagues have developed a series of models explaining automobile ownership and commuting mode, estimated from survey data collected before and after the opening of the Bay Area Rapid Transit (BART) system in the San Francisco area. Here we present one of the simplest, explaining only mode choice: the "naive model" reported by McFadden et al. (1977, pp. 121-123). It assumes choice among four modes: (1) auto alone, (2) bus with walk access, (3) bus with auto access, and (4) carpool (two or more occupants). The model's parameters are estimated from a sample of 771 commuters to San Francisco or Oakland who were surveyed prior to opening of the BART system.

Mode choice is explained by just three independent variables plus three alternative-specific constants. The three variables are: c_{in}/w_n, the round-trip variable cost (in US $) of mode i for traveler n divided by the traveler's post-tax wage rate (in $ per minute); T_{in}, the in-vehicle travel time (in minutes); and T_{in}^o, the out-of-vehicle travel time including walking, waiting, and

[14] For additional examples, see McCarthy (2001, ch. 3-4) and Small and Winston (1999).

transferring. Cost c_{in} includes parking, tolls, gasoline, and maintenance (Train, 1980, p. 362). The estimated utility function is:

$$V = \underset{(0.0054)}{-0.0412 \cdot c/w} \quad \underset{(0.0072)}{-0.0201 \cdot T} \quad \underset{(0.0070)}{-0.0531 \cdot T^o} \quad \underset{(0.26)}{-0.89 \cdot D^1} \quad \underset{(0.24)}{-1.78 \cdot D^3} \quad \underset{(0.25)}{-2.15 \cdot D^4} \qquad (2.23)$$

where the subscripts denoting mode and individual have been omitted, and standard errors of coefficient estimates are given in parentheses. Variables D^j are alternative-specific dummies.

This utility function is a simplification of (2.14) (with $\beta^3=\gamma^j=0$), except that travel time is broken into two components, T and T^o. Adapting (2.15), we see that the "value of time" for each of these two components is proportional to the post-tax wage rate, the proportionality constant being the ratio of the corresponding time-coefficient to the coefficient of c/w. Hence the values of in-vehicle and out-of-vehicle time are 49 percent and 129 percent of the after-tax wage. The negative alternative-specific constants indicate that the hypothetical traveler facing equal times and operating costs by all four modes will prefer bus with walk access (mode 2, the base mode); probably this is because each of the other three modes requires owning an automobile, which entails fixed costs not included in variable c. The strongly negative constants for bus with auto access (mode 3) and carpool (mode 4) probably reflect unmeasured inconvenience associated with getting from car to bus stop and with arranging carpools.

The model's fit could undoubtedly be greatly improved by including automobile ownership, perhaps interacted with $(D^1+D^3+D^4)$ to indicate a common effect on modes that use an automobile. However, there is good reason to exclude it because it is endogenous—people choosing one of those modes for other reasons are likely to buy an extra car as a result. This in fact is demonstrated by the more complete model of Train (1980), which considers both choices simultaneously. The way to interpret (2.23), then, is as a "reduced-form" model that incorporates implicitly the automobile ownership decision. It is thus applicable to a time frame long enough for automobile ownership to adjust to changes in the variables included in the model.

2.3.2 Trip-Scheduling Choice

One of the key decisions affecting congestion is the timing or scheduling of work trips. There is now a substantial body of empirical work on this subject, reviewed by Mahmassani (2000).

Although the scheduling decision is inherently continuous, most authors model it as a discrete choice among time intervals. There are two reasons for this: survey responses are

rounded off to a few even numbers, and disaggregate models can easily portray the complex manner in which travel time varies across possible schedules. Small (1982) estimates the choice among twelve possible five-minute intervals for work arrival time, using a set of auto commuters from the San Francisco Bay Area who have an official work-start time. The data set includes characteristics of the workers and a network-based engineering calculation of the travel time that each would encounter at each arrival time. Commuters are assumed to have full information; reliability of arrival is not considered except that the specification allows the commuter to avoid (through an estimated utility penalty) arriving just in time for work.

The utility specification postulates a linear penalty for arriving early, on the assumption that time spent before work is relatively unproductive; and a much larger linear penalty for arriving late, on the assumption that employer sanctions take hold with gradually increasing severity. Define *schedule delay*, S_D, as the difference (in minutes, rounded to nearest five minutes) between the arrival time represented by a given alternative and the official work start time. Define "Schedule Delay Late," *SDL*, as $\text{Max}\{S_D,0\}$ and "Schedule Delay Early," *SDE*, as $\text{Max}\{-S_D,0\}$. Define a "late dummy," *DL*, equal to one for the on-time and all later alternatives and equal to 0 for the early alternatives. Define *T* as the travel time (in minutes) encountered at each alternative.

The utility function estimated by Small (1982, Table 2, Model 1), with estimated standard errors in parentheses, is:

$$\begin{array}{lllll} V = & -0.106\cdot T & -0.065\cdot SDE & -0.254\cdot SDL & -0.58\cdot DL. \\ & (0.038) & (0.007) & (0.030) & (0.21) \end{array} \qquad (2.24)$$

This excludes two variables used to represent a tendency of respondents to round off answers to the nearest 10 or 15 minutes. More complex models are also estimated, in which the various penalties are nonlinear or depend upon such factors as the worker's family status or occupation, whether solo driver or carpooler, and how much flexibility for late arrivals the employer is said to allow.

Figure 2.1 shows utility function (2.24), divided by the coefficient of travel time. The marginal rates of substitution indicate that the commuter is willing to suffer an extra 0.61 minutes of congestion to reduce the amount of early arrival by one minute;[15] and 2.40 minutes of congestion to reduce late arrival by one minute, plus an extra 5.47 minutes congestion to avoid

[15] Calculated as 0.065/0.106=0.61.

any of the just-on-time or late alternatives. These turn out to be key parameters in models, to be presented in the next chapter, which describe equilibrium when congestion occurs in the form of queueing behind a bottleneck. They also can be used to formulate models of traveler response to network unreliability, as described by Bates et al. (2001).

These alternatives have a natural ordering; so why is the ordered response model not used? There are two reasons. First, as already noted, the ordered response model cannot take advantage of information that varies by alternative, such as travel time. Second, even accounting just for socioeconomic variables, there is no plausible combination of them that would exert a monotonic influence on the time of day; rather, there are likely to be some variables that favor peak times, others that favor times either earlier or later than that, others still that would affect the strength of preference for low travel times, and so forth.

2.3.3 Choice of Free or Express Lanes

Lam and Small (2001) analyze data from commuters with an option of paying to travel in a set of express lanes on a very congested freeway. The data set contains cross-sectional variation in the cost of choosing the express lanes because the toll depends on time of day and on car occupancy, both of which differ across respondents. Travel time also varies by time of day, fortunately in a manner not too highly correlated with the toll. The authors construct a measure of the unreliability of travel time by obtaining data on travel times across many different days, all at the same time of day. After some experimentation, they choose the median travel time (across days) as the best measure of travel time, and the difference between 90^{th} and 50^{th} percentile travel times (also across days) as the best measure of unreliability. This latter choice is based on the idea, documented in the previous subsection, that people are more averse to unexpected delays than to unexpected early arrivals.

The model explains a pair of related decisions: (1) whether to acquire a transponder (required to ever use the express lanes), and (2) which lanes to take on the day in question. A natural way to view these decisions is as a hierarchical set, in which the transponder choice is governed partly by the size of the perceived benefits of being able to use it to travel in the express lanes. As we will see in the next section, a model known as "nested logit" has been developed precisely for this type of situation, and indeed Lam and Small estimate such a model. As it happens, though, they obtain virtually identical results with a simpler "joint logit" model in

which there are three alternatives: (1) no transponder; (2) have a transponder but travel in the free lanes on the day in question; and (3) have a transponder and travel in the express lanes on the day in question. The results of this model are:[16]

$$\begin{aligned} V = \; & \underset{(0.411)}{-0.862 \cdot D^{tag}} \; \underset{(0.0058)}{+0.0239 \cdot Inc \cdot D^{tag}} \; \underset{(0.412)}{-0.766 \cdot ForLang \cdot D^{tag}} \; \underset{(0.853)}{-0.789 \cdot D^3} \\ & \underset{(0.138)}{-0.357c} \; \underset{0.056)}{-0.109 \cdot T} \; \underset{(0.048)}{-0.159 \cdot R} \; \underset{(0.046)}{+0.074 \cdot Male \cdot R} + (\text{other terms}). \end{aligned} \tag{2.25}$$

Here $D^{tag} \equiv D^2 + D^3$ is a composite alternative-specific dummy variable for those choices involving a transponder, or "toll tag"; its negative coefficient presumably reflects the hassle and cost of obtaining one. Getting a transponder is apparently more attractive to people with high annual incomes (*Inc*, in $1000s per year) and less attractive to those speaking a foreign language (dummy variable *ForLang*). The statistical insignificance of the coefficient of D^3, an alternative-specific dummy for using the express lanes, suggests that the most important explanatory factors are included explicitly in the model.

The coefficients on per-person cost c, median travel time T, and unreliability R can be used to compute dollar values of time and reliability. Here we focus on two aspects of the resulting valuations. First, reliability is highly valued, achieving coefficients of similar magnitudes as travel time. Second, men seem to care less about reliability than women; their value is only 53 percent as high as women's according to the point estimates,[17] although the difference (i.e. the coefficient of $Male \cdot R$) is not quite statistically significant even at a 10-percent significance level. Several studies of this particular toll facility have found women noticeably more likely to use the express lanes than men, and this formulation provides tentative evidence that the reason is a greater aversion to the unreliability of the free lanes.

2.4 Advanced Discrete-Choice Modeling

2.4.1 Generalized Extreme Value Models

[16] This is a partial listing of the coefficients in Lam and Small (2001), Table 11, Model 4b, with coefficients of T and R divided by 1.37 to adjust travel-time measurements to the time of the survey, as described on their p. 234 and Table 11, note *a*. Standard errors are in parentheses.

[17] The coefficient for women is -0.159, and that for men is -0.159+0.074 = -0.085.

Often it is implausible that the additive random utility components ε_j be independent, especially if important variables are omitted from the model's specification. This will make either logit or iid probit predict poorly.

A simple example is mode choice among automobile, bus transit, and rail transit. The two public-transit modes are likely to have many unmeasured attributes in common. Suppose a traveler initially has available only auto (j=1) and bus (j=2), with equal systematic utilities V_j so that the choice probabilities are each one-half. Now suppose we want to predict the effects of adding a type of rail service (j=3) whose measurable characteristics are identical to those for bus service. The iid models would predict that all three modes would then have choice probabilities of one-third, whereas in reality the probability of choosing auto would most likely remain near one-half while the two transit modes divide the rest of the probability equally between them. The argument is even stronger if we imagine instead that the newly added mode is simply a bus of a different color: this is the famous "red bus, blue bus" example.

The probit model generalizes naturally, as already noted, by allowing the distribution function in equation (2.5) to be multivariate normal with an arbitrary variance-covariance matrix. It must be remembered that not all the elements of this matrix can be distinguished (*identified*, in econometric terminology) because, as already noted, it is only the (J-1) utility differences that affect behavior.[18]

The logit model generalizes in a comparable manner, as shown by McFadden (1978, 1981). The distribution function is postulated to be *Generalized Extreme Value* (GEV), given by

$$F(\varepsilon_1,...,\varepsilon_J) = \exp\left[-G(e^{-\varepsilon_1},...,e^{-\varepsilon_J})\right]$$

where G is a function satisfying certain technical conditions. With this distribution, the choice probabilities are of the form

$$P_i = \frac{e^{V_i} \cdot G_i(e^{V_1},...,e^{V_J})}{G(e^{V_1},...,e^{V_J})} \tag{2.26}$$

where G_i is the i-th partial derivative of G. The expected maximum utility is

[18] The variance-covariance matrix of these utility differences has $(J\text{-}1)^2$ elements and is symmetric. Hence there are only $J(J\text{-}1)/2$ identifiable elements of the original variance-covariance matrix, less one for utility-scale normalization (Chu, 1981, pp. 58-59; Bunch, 1991).

$$E \max_j (V_j + \varepsilon_j) = \log G(e^{V_1},...,e^{V_J}) + \gamma \tag{2.27}$$

where again γ is Euler's constant.[19] Logit is the special case $G(y_1,...,y_J) = y_1+...+y_J$.

The best known GEV model, other than logit itself, is *nested logit*, also called *structured logit* or *tree logit* and first developed by Ben-Akiva (1974). McFadden (1981) discusses its theoretical roots and computational characteristics. In this model, certain groups of alternatives are postulated to have correlated random terms. This is accomplished by grouping the corresponding alternatives in G in a manner we can illustrate using the auto-bus-rail example, with auto the first alternative:

$$G(y_1, y_2, y_3) = y_1 + \left(y_2^{1/\rho} + y_3^{1/\rho}\right)^{\rho} . \tag{2.28}$$

In this equation, ρ is a parameter between 0 and 1 that indicates the degree of dissimilarity between bus and rail; more precisely, $1-\rho^2$ is the correlation between ε_1 and ε_2 (Daganzo and Kusnic, 1993). The choice probability for this example, computed from (2.26), may be written:

$$P_i = P(B_{r(i)}) \cdot P(i \mid B_r) \tag{2.29}$$

$$P(B_r) = \frac{\exp(\rho \cdot I_r)}{\sum_{s=1}^{2} \exp(\rho \cdot I_s)} \tag{2.30}$$

$$P(i \mid B_r) = \frac{\exp(V_i / \rho)}{\sum_{j \in B_r} \exp(V_j / \rho)} \tag{2.31}$$

where $B_1=\{1\}$ and $B_2=\{2,3\}$ are a partition of the choice set into groups; $r(i)$ indexes the group containing alternative i; and I_r denotes the *inclusive value* of set B_r, defined as the logarithm of the denominator of (2.31):

$$I_r = \log \sum_{j \in B_r} \exp(V_j / \rho) . \tag{2.32}$$

[19] This is demonstrated by Lindberg, Eriksson, and Mattsson (1995, p. 134). For a simple and elegant proof, see Choi and Moon (1997, p. 131).

When ρ=1 in this model, ε_2 and ε_3 are independent and we have the logit model. As $\rho \downarrow 0$, ε_2 and ε_3 become perfectly correlated and we have an extreme form of the "red bus, blue bus" example, in which auto is pitted against the better (as measured by V_i) of the two transit alternatives; in this case $\rho I_1 \to V_1$ and $\rho I_2 \to \max\{V_2, V_3\}$.

The model just described can be generalized to any partition $\{B_r, r=1,\ldots,R\}$ of alternatives, and each group B_r can have its own parameter ρ_r in equations (2.28)-(2.32), leading to the form:

$$G(y_1,\ldots,y_J) = \sum_r \left(\sum_{j \in B_r} y_j^{1/\rho_r} \right)^{\rho_r} . \tag{2.33}$$

This is the general two-level nested logit model. It has choice probabilities (2.29)-(2.32) except that the index s in the denominator of (2.30) now runs from 1 to R. Like logit, it can be also derived from an entropy formulation (Brice, 1989). The groups B_r can themselves be grouped, and those groupings further grouped, and so on, giving rise to even more general "tree structures" of three or more levels.

As in the logit model, the inclusive value is a summary measure of the overall desirability (expected maximum utility) of the relevant group of alternatives. This fact gives the "upper-level" probability (2.30) a natural interpretation as a choice among groups, taking the logit form with I_r playing the role of the independent variable and ρ_r its coefficient. As an example of an application of this interpretation, the inclusive value of a set of locations can serve as a measure of accessibility in models of destination choice (Ben-Akiva and Lerman, 1979). The welfare measure for the two-level nested logit model is, from (2.27), (2.32), and (2.33):

$$W = \frac{1}{\lambda} \log \sum_r \exp(\rho_r \cdot I_r) \tag{2.34}$$

where again λ is the marginal utility of income.

In nested logit, $\{B_r\}$ is an exhaustive partition of the choice set into mutually exclusive subsets. Therefore equation (2.31) is a true conditional probability, and the model can be estimated sequentially: first estimate the parameters (β/ρ) from (2.31), use them to form the inclusive values (2.32), then estimate ρ from (2.30). Each estimation step uses an ordinary logit log-likelihood function, so it can be carried out with a logit algorithm. However, this sequential method is not statistically efficient, nor does it produce consistent estimates of the standard

errors of the coefficients (Amemiya, 1978). Several studies show that maximum-likelihood estimation, although computationally more difficult, gives more accurate results (Hensher, 1986; Daly, 1987; Brownstone and Small, 1989).[20]

There is some experience with other GEV models. Most of them generalize (2.33) by not requiring the subsets B_r to be mutually exclusive. Small (1987) defines subsets each encompassing two or more alternatives that lie close to each other on some ordering. Chu (1981) and Koppelman and Wen (2000) study models in which the subsets B_r include all possible pairs of alternatives:

$$G(y_1,...,y_J) = \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} \left(y_j^{1/\rho_{jk}} + y_k^{1/\rho_{jk}} \right)^{\rho_{jk}} . \tag{2.35}$$

Such a model has the same number of estimable parameters in its variance-covariance matrix as multinomial probit (again, the arbitrary scale requires one more normalization), and so might have comparable generality. In practice this model seems to be easier to estimate than multinomial probit.

Nevertheless, estimation of GEV models is often difficult because of the highly nonlinear manner in which ρ_r enters the equation for choice probabilities. When the true model is GEV but differs only moderately from logit, a reasonable approximation can be estimated using two steps of a standard logit estimation routine, a procedure that appears to be considerably more stable than maximum likelihood estimation of the exact GEV model (Small, 1994).

A different direction for generalizing the logit model is to maintain independence between error terms while allowing each error term to have a unique variance. This is the heteroscedastic extreme value model of Bhat (1995); it is a random-utility model but not in the GEV class, and its probabilities cannot be written in closed form so require numerical integration. Other extensions of the logit model are described by Koppelman and Sethi (2000).

2.4.2 Combined Discrete and Continuous Choice

In many situations, the choice among discrete alternatives is made simultaneously with some related continuous quantity. For example, a household's choice of type of automobile to own is

[20] If maximizing the log-likelihood function is numerically difficult, one can start with the sequential estimator and carry out just one step of a Newton-Raphson algorithm toward maximization; this yields a statistically efficient estimate and seems to work well in practice (Brownstone and Small, 1989).

closely intertwined with its choice of how much to drive. Estimating equations to explain usage, conditional on ownership, creates a *sample selection bias* (Heckman, 1979): for example, people who drive a lot are likely to select themselves into the category of owners of nice cars, so we could inadvertently overstate the independent effect of nice cars on driving. A variety of methods are available to remove this bias, as described in Train (1986, chap. 5), Mannering and Hensher (1987), and Washington *et al.* (2003, ch. 12).

The essence of the problem can be illustrated within an example of binary choice: that of owning a new or used automobile, denoted j=1 or 2. Each type of car has a fixed measurable quality level Q_j that we can assume is higher for new cars, i.e. $Q_1>Q_2$. For example, Q could be the number of safety features offered from a particular list, or simply an alternative-specific dummy variable equal to 1 for a new car. Let us suppose that the decision of how much to drive depends on car quality and income Y, as follows:

$$x = \beta_0 + \beta_Q Q + \beta_Y Y + u \tag{2.36}$$

where Y is income and u is a random error term. For simplicity we have omitted the subscript n denoting the individual in the sample. Car quality can be written in terms of the choice variable d_{1n} defined earlier, as follows (again omitting subscript n):

$$Q = d_1 Q_1 + (1 - d_1) Q_2 . \tag{2.37}$$

Substituting (2.37) into (2.36) makes explicit the dependence of the usage decision (x) on the ownership decision (d_1).

Suppose also that the ownership decision depends on some set of observable variables X, which could include Y and (Q_1-Q_2):

$$\begin{aligned} &d_1 = 1 \text{ if } U_1 > U_2, \quad 0 \text{ otherwise;} \\ &U_1 - U_2 = \beta_X' X + \varepsilon . \end{aligned} \tag{2.38}$$

This equation defines a binary probit model if ε is assumed normal, binary logit if ε is assumed logistic.

Selection bias is present if u and ε are correlated, which is likely because unobservable factors may affect both usage and the relative desirability of a new car. (An example of such a factor is how much this individual likes listening to a high-quality car stereo.) If u is correlated with ε, it is also correlated with the car-type indicator d_1 and therefore with car quality Q via

(2.37). This biases the estimated coefficients in (2.36), especially β_Q, because Q is endogenous there.

If we can find an exogenous proxy for Q, we can use it instead and solve the problem. This can be accomplished using the following two-step procedure proposed by Heckman (1979).

Step 1 consists of estimating a reduced-form version of (2.38). In this stage, the endogenous variable x is replaced by the exogenous variables that are postulated to determine it: namely, Q_1, Q_2, and Y. Thus the ownership decision is modeled as some function of X, Q_1, Q_2, and Y. Unfortunately theory does not provide definitive guidance on the functional form for these variables, but usually some experimentation will produce a satisfactory fit. In this step, all that matters is that we obtain a reasonably good predictor of the probability $\hat{P}_1$ that the person will choose a new car. For convenience, let Z represent all the variables (and transformations of them, if any) used in this reduced form, and ε_R be the error term, so that the utility difference is

$$U_1 - U_2 = \beta_Z Z + \varepsilon_R .$$

From the estimated coefficient vector $\hat{\beta}_Z$, we can compute a predicted probability $\hat{P}_1$ of choosing a new car, equal to $\Phi(\hat{\beta}_Z Z)$ if the model is probit or $\left[1+\exp(-\hat{\beta}_Z Z)\right]^{-1}$ if the model is logit.

Step 2 consists of estimating a variant of (2.36) that is purged of endogeneity. There are two alternative strategies for doing this:

Step 2 Version (a): Replace Q by an exogenous predictor $\hat{Q}$.

We look for an unbiased estimate of Q that does not use the observed ownership choice, d_1, as does (2.37). There are at least three possibilities, which are Methods II, I, and III of Train (1986, p. 90):

(i) Compute $\hat{Q}$ as $E(Q) \equiv \hat{P}_1 \cdot Q_1 + (1-\hat{P}_1)\cdot Q_2$.

(ii) Compute $\hat{Q}$ as the predicted value from an auxiliary regression of observed Q on all the exogenous variables of the system, namely Z. (Note this method does not actually require that Step 1 be carried out.)

(iii) Compute $\hat{Q}$ from an auxiliary regression as in (ii) with $E(Q)$, calculated as in (i), as an additional variable in the regression. This procedure is more statistically efficient than

either (i) or (ii) because it incorporates data on actual choices (via the process for computing $\hat{P}_1$) as well as on variables *Z*.

Method (iii) is probably the best choice in most cases, although like (ii) it requires one to specify arbitrarily the exact functional form of the auxiliary regression.

*Step 2 Version (b): Add a "correction term" to the error term in (*2.36*) to make it independent of u.*

One way to look at selection bias is that observed *Q* is conditional on the individual's ownership decision. Therefore using *Q* as a variable in (2.36) would be appropriate if (2.36) could be transformed into an equation describing usage *conditional on* ownership. This can be done by making its error term conditional on ownership. If *u* is assumed to be normal, as is usual, the required transformation is accomplished by subtracting the conditional expectation of a normal variable, given its link to ownership via (2.38), from *u*; the remaining error term is can be assumed independent of *Q* and so (2.36) is purged of selectivity bias. A recent example of use of this technique is West's (2004) model of automobile type choice and amount of use.

That conditional expectation can be computed explicitly for binary probit and logit models.[21] We write the new term to be added to (2.36) as γC where γ is a parameter to be estimated and *C* is a "correction variable" computed from the results of Step 1. The estimated value of γ will give us information about the correlation between *u* and ε, which we denote by ρ. It is this correlation that causes the problem, so we can test for selection bias by testing whether γ is different from zero.

Table 2.1 gives formulas for the correction variable $C = d_1C_1+(1-d_1)C_2$; it also shows how parameter γ is related to correlation ρ. In this table, Φ denotes the probability distribution function of the standard normal distribution, and ϕ its derivative (i.e. the normal density function); σ_u is the standard deviation of *u*; and $\hat{P}_2 = 1 - \hat{P}_1$. Sometimes data are lacking on people

[21] Supposedly it can be done for multinomial logit model as well, but it is extremely complex. Dubin and McFadden (1984) specify a usage model conditional on a single choice, *i*; the include *J*-1 correction terms, and so estimate *J*-1 correlations (between *u* and ε_j, $j \neq i$). However they do not discuss how to pool the data with observations of individuals who choose other alternatives.

making choice j=2, in which case the correction factor is simply C_1 and the usage equation is estimated on just the subsample of new car owners.[22]

Table 2.1. Selectivity Correction Terms $\gamma \cdot (D^1 C_1 + D^2 C_2)$

Model	*Correction Variable*		*Coefficient*
	C_1	C_2	γ
Probit	$\frac{\phi(\hat{\beta}_Z Z)}{\hat{P}_1}$	$-\frac{\phi(\hat{\beta}_Z Z)}{\hat{P}_2}$	$\rho\sigma_u$
Logit	$-\left[\frac{\hat{P}_2 \ln \hat{P}_2}{1-\hat{P}_2} + \ln \hat{P}_1\right]$	$\left[\frac{\hat{P}_1 \ln \hat{P}_1}{1-\hat{P}_1} + \ln \hat{P}_2\right]$	$\left(\sqrt{6}/\pi\right)\cdot \rho\sigma_u$

What this procedure does is add correction γC_1 for those individuals in the sample who chose a new car, and γC_2 for the others. Note that in each row, C_1 is positive and C_2 is negative. Thus if ρ is positive, indicating that people choosing new cars are likely to drive more for unobserved reasons, the adjustment indicates that ε has a positive expected value for those individuals who choose new cars, and a negative expected value for those who choose used cars. The extent of the adjustment is determined by estimated coefficient γ, which is positively related to ρ as shown in the last column.

More elaborate systems of equations can be handled with the tools of *structural equations modeling*. These methods are quite flexible and allow one to try out different patterns of mutual causality, testing for the presence of particular causal links. They are often used when large data sets are available describing mutually related decisions. Golob (2003) provides a review.

2.4.3 Disaggregate Panel Data

[22] Sign conventions vary in the literature. In the probit case, some references replace $\hat{\beta}_Z Z$ by the equivalent quantity $\Phi^{-1}(\hat{P}_1)$, where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function.

Just as with aggregate data, data from individual respondents can be collected repeatedly over time. A good example is the Dutch Mobility Panel, in which travel-diary information was obtained from the same individuals (with some attrition and replacement) at ten different times over the years 1984-1989. The resulting data have been widely used to analyze time lags and other dynamic aspects of travel behavior (Van Wissen and Meurs, 1989).

The methods described earlier for aggregate panel data are applicable to disaggregate data as well. In addition, attrition becomes a statistical issue: over time, some respondents will be lost from the sample and the reasons need not be independent of the behavior being investigated. The solution is to create an explicit model of what causes an individual to leave the sample, and to estimate it simultaneously with the choice process being considered. Pendyala and Kitamura (1997) and Brownstone and Chu (1997) analyze the issues involved.

2.4.4 Random Parameters and Mixed Logit

In the random utility model of (2.4)-(2.5), randomness in individual behavior is limited to an additive error term in the utility function. Other parameters, and functions of them, are deterministic: that is, the only variation in them is due to observed variables. Thus for example, the value of time defined by (2.13) varies with observed travel time and wage rate but otherwise is the same for everyone.

Experience has shown, however, that parameters of critical interest to transportation policy vary among individuals for reasons that we do not observe. Such reasons could be missing socioeconomic characteristics, personality, special features of the travel environment, and data errors. These, of course, are the same reasons for the inclusion of the additive error term in utility function (2.4). So the question is, why not also include randomness in the other parameters?

The only reason is tractability, and that has largely been overcome by advances in computing power. Boyd and Mellman (1980) and Cardell and Dunbar (1980) showed how one could allow a parameter in the logit model to vary randomly across individuals. The idea is to specify a distribution, such as normal with unknown mean and variance, for the parameter in question; the overall probability is determined by embedding the integral in (2.5) within another integral over the density function of that distribution. Subsequently, this simple idea was generalized to allow for general forms of randomness in all the parameters – even alternative-specific constants, where further randomness might seem redundant yet it proves a simple way to

produce correlation patterns like those in GEV without the complexity of the GEV probability formulas. Such models are tractable because the outer integration (over the distribution defining random parameters) can be performed using simulation methods based on random draws, while the inner integration (that over the remaining additive errors ε_{jn}) is unnecessary because, conditional on the values of random parameters, it yields the logit formula (2.8). The model is called *mixed logit* because the combined error term has a distribution that is a mixture of the extreme value distribution with the distribution of the random parameters.

The mixed logit model is simple to write out. Using the logit formulation of (2.8) and (2.10), the choice probability conditional on random parameters is

$$P_{in|\beta} = \frac{\exp(\beta' z_{in})}{\sum_j \exp(\beta' z_{jn})}. \tag{2.39}$$

Let $f(\beta|\Theta)$ denote the density function defining the distribution of random parameters, which depends on some unknown "meta-parameters" Θ (such as means and variances of β). The unconditional choice probability is then simply the multi-dimensional integral:

$$P_{in} = \int P_{in|\beta} \cdot f(\beta \mid \Theta) d\beta . \tag{2.40}$$

Integration by simulation consists of taking R random draws β^r, r=1,…,R, from distribution $f(\beta|\Theta)$, calculating $P_{in|\beta}$ each time, and averaging over the resulting values:

$$P_{in}^{sim} = (1/R)\sum_{r=1}^{R} P_{in|\beta}^{r} .$$

Doing so requires, of course, assuming some trial value of Θ, just as calculating the usual logit probability requires assuming some trial value of β. Under reasonable conditions, maximizing the likelihood function defined by this simulated probability yields statistically consistent estimates of the meta-parameters Θ. Details are provided by Train (2003).

Brownstone and Train (1999) demonstrate how one can shape the model to capture anticipated patterns by specifying which parameters are random and what form their distribution takes – in particular, whether some of them are correlated with each other.[23] In their application,

[23] The following simplified explanation is adapted from Small and Winston (1999).

consumers state their willingness to purchase various makes and models of cars, each specified to be powered by one of four fuel types: gasoline (G), natural gas (N), methanol (M), or electricity (E). Respondents were asked to choose among hypothetical vehicles with specified characteristics. A partial listing of estimation results is as follows:

$$V = -0.264\cdot[p/\ln(inc)] + 0.517\cdot range + (1.43+7.45\phi_1)\cdot size + (1.70+5.99\phi_2)\cdot luggage + 2.46\phi_3\cdot nonE + 1.07\phi_4\cdot nonN + (other\ terms)$$

where p (vehicle price) and *inc* (income) are in thousands of dollars; the *range* between refueling (or recharging) is in hundreds of miles; *luggage* is luggage space relative to a comparably sized gasoline vehicle; *nonE* is a dummy variable for cars running on a fuel that must be purchased outside the home (in contrast to electric cars); *nonN* is a dummy for cars running on a fuel stored at atmospheric pressure (in contrast to natural gas); and ϕ_1-ϕ_4 are independent random variables with the standard normal distribution. All parameters shown above are estimated with enough precision to easily pass tests of statistical significance.

This model provides for observed heterogeneity in the effect of *price* on utility, since it varies with *income*. It provides for random coefficients on *size* and *luggage*, and for random constants as defined by *nonE* and *nonN*. This can be understood by examining the results term by term.

The terms in parentheses involving ϕ_1 and ϕ_2 represent the random coefficients. The coefficient of *size* is random with mean 1.43 and standard deviation 7.45. Similarly, the coefficient of *luggage* has mean 1.70 and standard deviation 5.99. These estimates indicate a wide variation in people's evaluation of these characteristics. For example, it implies that many people actually prefer less luggage space namely, those for whom ϕ_2<-1.70/5.99; presumably they do so because a smaller luggage compartment allows more interior room for the same size of vehicle. Similarly, preference for vehicle size ranges from negative (perhaps due to easier parking for small cars) to substantially positive.

The terms involving ϕ_3 and ϕ_4 represent random alternative-specific constants with a particular correlation pattern, predicated on the assumption that groups of alternatives share

common features for which people have idiosyncratic preferences – very similar to the rationale for nested logit. Each of the dummy variables *nonE* and *nonN* is simply a sum of alternative-specific constants for those car models falling into a particular group. The two groups overlap: any gasoline-powered or methanol-powered car falls into both. If the coefficients of ϕ_3 and ϕ_4 had turned out to be negligible, then these terms would play no role and we would have the usual logit probability conditional on the values of ϕ_1 and ϕ_2. But the coefficients are not negligible, so each produces a correlation among utilities for those alternative in the corresponding group. For example, all cars that are not electric share a random utility component $2.46\phi_3$, which has standard deviation 2.46; this is in addition to other random utility components including ε_{in} in (2.4), which as we have seen has standard deviation equal to $\pi/\sqrt{6}=1.28$ by normalization – a requirement for (2.39) to be valid. Thus the combined additive random term in utility,[24] $(\varepsilon_{in}+2.46\phi_{3n}\cdot nonE_i+1.07\phi_{4n}\cdot nonN_i)$, exhibits correlation across those alternatives *i* representing cars that are not electric and, by similar argument involving ϕ_4, across those alternatives representing cars that are not natural gas. Those alternatives falling into both *nonE* and *nonN* are even more highly correlated with each other. Note that because the distributions of ϕ_3 and ϕ_4 are centered at zero, this combined random term does not imply any overall average preference for or against various types of vehicles; such absolute preferences are in fact included in *other terms*.

The lesson from this example is that mixed logit can be used not only to specify unobserved randomness in the coefficients of certain variables, but also to mimic the kinds of correlation patterns among the random constants for which the GEV model was developed. Indeed, McFadden and Train (2000) show that it can closely approximate virtually any choice model based on random utility. The model described above acts much like a GEV model with overlapping nests for alternatives in groups *nonE* and *nonN*, and with random parameters for *size* and *luggage*. It is probably easier to estimate than such a nested logit model, especially if one is already committed to random parameters. Even more complicated error structures can be accommodated within this framework, for example one designating repeated observations from a given individual (Small,

[24] As Brownstone and Train point out, the terms in ϕ_1 and ϕ_2 also may be viewed as part of an additive random utility term, but one that is not a constant, i.e. it depends on values of observed variables.

Winston, and Yan, 2005) or spatial correlation related to geographical location (Bhat and Guo, 2004).

In principle, the mixing idea can be applied to any choice model, not just logit, in order to randomize its parameters. Indeed, it happens that the multinomial probit model was first developed with a random-parameters formulation (Hauman and Wise, 1978), a fact that has caused some confusion about the relationship between probit and logit. There may be cases where it is easier to estimate a random-parameters multinomial probit than a mixed logit model, but usually it is harder because one needs to simulate not only the explicit integral in (2.40) but also the integral that, for probit, is part of the definition the conditional choice probability $P_{in|\beta}$.

2.5 Activity Patterns

A more fundamental approach to the demand for travel would be to explain the entire structure of decision-making about what activities to undertake in what locations. This idea has proven difficult to translate into workable models that use available data, but important progress has been made. For example, some surveys now elicit multi-day diaries describing all activities and travel undertaken during a period of time. Descriptive statistics on activity patterns have been compiled showing surprising similarities across nations including US, UK, Japan, Canada, and The Netherlands (Timmermans *et al.*, 2002).

Ettema and Timermans (1997), Ben-Akiva and Bowman (1998), and Bhat and Koppelman (1999) provide reviews of activity-based models. There are two main classes. Econometric models extend the basic framework of this chapter to deal with additional choice dimensions such as trip frequency, destination, and type and duration of activities undertaken. Simulation models may also utilize a utility-maximization choice framework, but they emphasize more the enumeration of feasible activities based on various constraints such as that relating the starting and ending times of each trip to its duration.

One significant advance has been to model an entire tour (a round trip visiting one or more destinations in sequence) as an object of choice. The problem is that this quickly leads to enormous numbers of possible alternatives, especially when one considers a daily schedule containing several possible tours. Bowman and Ben-Akiva (2001) improve tractability by breaking the overall decision about the daily schedule into parts, including a primary tour type,

secondary tour type(s), and destinations and modes of travel for each tour. Such a model lends itself to a structured choice model such as nested logit. Illustrating the difficulty of designing realistic models, the authors acknowledge that the results in their example are able to explain only a small part of variations in observed activity patterns.

Few if any formal models have been able to account for flexibility in both the times of day and the locations at which activities take place, both of which are fundamental to describing the trips connecting them. Furthermore, to fully understand the processes generating travel, one needs to model the substitution between in-home and out-of-home activities, which adds further to the sheer number of possibilities to consider.

As an example of what can be accomplished with such models, Shiftan and Suhrbier (2002) utilize one of the best data sets for activity analysis – a 1994 household survey in Portland, Oregon – to analyze several policies classified as "travel demand management." One result is illustrative. A policy to encourage telecommuting is predicted to reduce long-distance work trips to downtown Portland. But it *increases* the number of short tours as people make special-purpose trips for activities that previously were handled as part of a tour from home to work and back. Other studies of telecommuting have found only a very small net reduction in travel (Choo, Mokhtarian, and Salomon, 2005). More generally, many types of telecommunication appear to be complements to, rather than substitutes for, travel (Plaut, 1997).

2.6 Value of Time and Reliability

Among the most important quantities inferred from travel demand studies are the monetary values that people place on saving various forms of travel time or improving the predictability of travel time. The first, loosely known as the *value of time* (VOT), is a key parameter in cost-benefit analyses that measure the benefits brought about by transportation policies or projects. The second, the *value of reliability* (VOR), also appears important, but accurate measurement is a science in its infancy. The benefits or losses due to changes in time and reliability are normally captured as part of consumer surplus, for example that given by (2.17), so long as they are part of the demand model. However, it is often enlightening to separate them explicitly.

2.6.1 Theory of Value of Time

The most natural definition of value of time is in terms of compensating variation. The value of saving a given amount and type of travel time by a particular person is the amount that person could pay, after receiving the saving, and be just as well off as before. This amount, divided by the time saving, is that person's average value of time saved for that particular change. Aggregating over a class of people yields the *average value of time* for those people in that situation. The limit of this average value, as the time saving shrinks to zero, is called the *marginal value of time*, or just "value of time;" by definition, it is independent of the amount of time saving despite confusion on this subject.[25]

Value of time may depend on many aspects of the trip-maker and of the trip itself. To name just a few, it depends on trip purpose (e.g. work or recreation), demographic and socio-economic characteristics, time of day, physical or psychological amenities available during travel, and the total duration of the trip. There are two main approaches to specifying a travel-demand model so as to measure such variations. One is known as *market segmentation*: the sample is divided according to criteria such as income and type of household, and a separate model is estimated for each segment. This has the advantage of imposing no potentially erroneous constraints, but the disadvantage of requiring many parameters to be estimated, with no guarantee that these estimates will follow a reasonable pattern. The second approach uses theoretical reasoning to postulate a functional form for utility that determines how VOT varies. This second approach is pursued here.

A useful theoretical framework builds on that of Becker (1965), in which utility is maximized subject to a time constraint. Becker's theory has been elaborated in many directions; here, we present ideas developed mainly by Oort (1969) and DeSerpa (1971), adapting the exposition of MVA Consultancy et al. (1987).

Let utility U depend on consumption of goods G, time T_w spent at work, and times T_k spent in various other activities k. We can normalize the price of consumption to one. Utility is

[25] It is sometimes claimed that the average value of time savings diminishes rapidly as the time savings shrink to zero, which would imply a very low marginal rate. But these claims are based on *ad hoc* empirical specifications, and ignore the fact that travel patterns are in constant flux so the time saving from one particular policy cannot long be distinguished from other sources of differences in trip times. Studies based on consistent definitions have not found such dependencies (MVA Consultancy *et al*., 1987, pp. 65-68), and theory refutes the alleged rationale for them (Mackie, Jara-Díaz, and Fowkes, 2001).

maximized subject to several constraints. First, there is the usual budget constraint involving unearned income Y and earned income wT_w, where w is the wage rate. Second, a time constraint requires that time spent on all activities equal total time available, $\overline{T}$. Finally, the nature of certain activities (such as travel) imposes a minimum $\overline{T}_k$ on time T_k spent in activity k. (We will consider as an extension the possibility that T_w is also constrained.)

This problem can be solved by maximizing the following Lagrangian function with respect to G, T_w, and $\{T_k\}$:

$$L = U\left(G, T_w, \{T_k\}\right) + \lambda \cdot \left[Y + wT_w - G\right] + \mu \cdot \left[\overline{T} - T_w - \sum_k T_k\right] + \sum_k \phi_k \cdot \left[T_k - \overline{T}_k\right], \qquad (2.41)$$

where λ, μ, and $\{\phi_k\}$ are Lagrangian multipliers that indicate how tightly each of the corresponding constraints limits utility. The first-order condition for maximizing (2.41) with respect to one activity time T_k is

$$U_{Tk} - \mu + \phi_k = 0 \qquad (2.42)$$

while that with respect to T_w is

$$U_{Tw} + \lambda \cdot \left[w + T_w \cdot (dw/dT_w)\right] - \mu = 0\,, \qquad (2.43)$$

where subscripts on U indicate partial derivatives. We have allowed for a nonlinear compensation schedule by letting w depend on T_w.

We can denote the value of utility at the solution to this maximization problem by V, the indirect utility function; it depends on Y, $\overline{T}$, wage schedule $w(T_w)$, and minimum activities times $\{\overline{T}_k\}$. The rate at which utility increases as the k-th minimum-time constraint is relaxed is given by its Lagrange multiplier, ϕ_k; the increase with respect to unearned income is λ. Hence the marginal value of time for the k-th time component is their ratio:

$$v_T^k \equiv \left(\frac{\partial Y}{\partial \overline{T}_k}\right)_V = \frac{\phi_k}{\lambda}\,. \qquad (2.44)$$

Those activities for which the minimum-time constraint is not binding, i.e. those for which ϕ_k=0, are called by DeSerpa *pure leisure activities*. The others, which presumably include most travel, are *intermediate activities*.

Equations (2.42)-(2.44) imply:

$$v_T^k = \frac{\mu - U_{Tk}}{\lambda} = w + T_w \cdot \frac{dw}{dT_w} + \frac{U_{Tw}}{\lambda} - \frac{U_{Tk}}{\lambda}\,. \qquad (2.45)$$

This equation decomposes the value of travel-time savings into the opportunity cost of time that could be used for work, μ/λ, less the value of the marginal utility of time spent in travel. The opportunity cost is both pecuniary (the first two terms after the last equality) and nonpecuniary (the third term, which could be positive or negative).

Most of the theoretical literature assumes that the wage is fixed, in which case equation (2.45) gives the result noted by Oort (1969): the value of time exceeds the wage rate if time spent at work is enjoyed relative to that spent traveling, and falls short of it if time at work is relatively disliked. This is a fundamental insight into how the value of time, even for non-work trips, depends on conditions of the job. It suggests a modeling strategy that interacts variables believed to be related to compensation and work enjoyment with those measuring time or cost. In addition, we might expect v_T^k to rise with total trip time because the total time constraint in (2.41) will bind more tightly, causing μ, the marginal utility of leisure, to rise.

2.6.2 Empirical Specifications

The most common situation for measuring values of time empirically is one where a discrete choice is being made, such as among modes or between routes. To clarify how the general theory just presented corresponds to empirical specifications, assume that there is only one pure leisure activity, k=0, and that the other activities are all mutually exclusive travel activities, each consisting of one trip. We can also add travel cost $c_k\delta_k$ to the budget constraint, where δ_k is one if activity k is chosen and zero otherwise. The indirect utility function has the same derivatives with respect to exogenous variables c_k and $\overline{T}_k$ as does the Lagrangian function:

$$\frac{\partial V}{\partial c_k} = -\lambda\delta_k ; \qquad \frac{\partial V}{\partial \overline{T}_k} = -\phi_k\delta_k .$$

Equivalently, the conditional indirect utility functions needed for a discrete-choice model satisfy:

$$\frac{\partial V_k}{\partial c_k} = -\lambda ; \qquad \frac{\partial V_k}{\partial \overline{T}_k} = -\phi_k , \tag{2.46}$$

which imply that our definition of value of time in (2.44) is identical to that in (2.13):

$$v_T^k \equiv \frac{\phi_k}{\lambda} = \frac{\partial V_k / \partial \overline{T}_k}{\partial V_k / \partial c_k} .$$

Note also that the first of equations (2.46) is identical to (2.18) since here there is assumed just one trip per time period. (It is easy to generalize this model to allow for an endogenously chosen number of trips per time period.)

Our theory provides some guidance about how to specify the systematic utilities V_k in a discrete choice model. Suppose, for example, one believes that work is disliked (relative to travel) and that its relative marginal disutility is a fixed fraction of the wage rate. Suppose further that the wage rate is fixed, so the second term in (2.45) disappears. Then (2.45) implies that the value of time is a fraction of the wage rate, as for example with specification (2.9) with β_3=0. Alternatively, one might think that work enjoyment varies nonlinearly with the observed wage rate: perhaps negatively due to wage differentials that compensate for working conditions, or perhaps positively due to employers' responses to an income-elastic demand for job amenities. Then (2.45) implies that value of time is a nonlinear function of the wage rate, which could suggest using (2.9) with a non-zero term β_3 or with additional terms involving cost divided by some other power of the wage.

Train and McFadden (1978) demonstrate how specific forms of the utility function in (2.41) can lead to operational specifications for the conditional indirect utility function with a desired relationship between value of time and wage rate. For example, in (2.9), we could have achieved the same relationship by multiplying the time-related variables by wage rather than by dividing the cost by wage; but doing so would imply a different underlying function $U(\cdot)$.

2.6.3 Extensions

Several extensions to the theory just presented are interesting. First, consider work-hour constraints. People are not always free to change the amount of time they spend at work, perhaps because they are locked into a particular job with fixed hours or because there are few jobs offered with the work hours that they prefer. To some extent this is handled by allowing w to depend on T_w. We could represent a stricter constraint by adding a term $\phi_w \cdot \left[T_w - \overline{T}_w\right]$ to (2.41), where ϕ_w is another Lagrangian multiplier whose sign indicates whether this person would prefer fewer (ϕ_w>0) or more (ϕ_w<0) hours at the job. This modification adds a term ϕ_w/λ to the value of time as given by (2.45). Possibly a positive value is suggested by the finding of MVA Consultancy et al. (1987, pp. 149-150) that people who are required to work extra hours at short notice have 15-20 percent higher values of travel time than other workers.

Another extension is suggested by Jara-Díaz (2000, 2003). Suppose goods consumption requires using an amount of leisure time proportional to those goods. This constraint is represented by modifying the term $\phi_0 \cdot \left[T_0 - \overline{T}_0\right]$ in the last summation in (2.41), where subscript 0 indicates the leisure activity; it now becomes $\phi_0 \cdot \left[T_0 - \ell \cdot G\right]$, where ℓ is the unit time requirement for consumption. The constraint is binding if $\phi_0>0$. (This modification involves a modification of the definition of leisure, which previously was defined, following DeSerpa, by the condition $\phi_0=0$.) While this modification does not alter the formulas derived for value of time, it does change the meaning of λ (the marginal utility of income) in those formulas. Previously, $\lambda=U_G$, as is easily seen by writing the first-order condition for maximizing (2.41) with respect to G. But this modified leisure constraint introduces a new term in that first-order condition, resulting in $\lambda=U_G-\phi_0\cdot\ell$. Since λ appears in the denominator of expressions for value of time, this change would tend to raise the value of time if the constraint is binding. This may be viewed as yet another model producing the "harried leisure class" postulated by Linder (1970) as a byproduct of rising productivities.

A different extension is considered by De Borger and Van Dender (2003). Suppose the time required for travel depends on the amount of time worked, so that $\overline{T}_k = t_k \cdot T_w$ for fixed parameter t_k. This might happen for a number of reasons: secondary workers entering or leaving the work force, part-time workers changing the number of days worked, or part-time workers having to accept more distant jobs in order to increase time worked. In that case, (2.43) acquires the additional term $-\phi_k t_k$, and all terms on the right-hand side of (2.45) are divided by $(1+t_k)$. Thus greater commuting time *decreases* the value of time – opposite to the effect noted earlier from a rising marginal utility of leisure – because it reduces the hourly wage rate net of commuting cost. De Borger and Van Dender suggest in numerical simulations that the effect can be quite large and causes unexpected results: for example, reducing congestion can cause the value of time to rise so much that total travel cost actually increases.

Other theoretical extensions show that value of time can depend on tax rates (Forsyth, 1980) and on scheduling considerations (Small, 1982).

2.6.4 Value of Reliability

It is well known that uncertainty in travel time, which may result from congestion or poor adherence to transit schedules, is a major perceived cost of travel (e.g., MVA Consultancy et al., 1987, pp. 61-62). This conclusion is supported by attitudinal surveys (Prashker, 1979), and perhaps by the frequent finding that time spent in congestion is more onerous than other in-vehicle time.[26] How can this aversion to unreliability be captured in a theoretical model of travel?

One approach, adapting Noland and Small (1995), is to begin with the model of trip-scheduling choice presented in equation (2.24). Dividing utility by minus the marginal utility of income, we can write this model in terms of trip cost, in a conventional notation that we will use extensively in the next chapter:

$$C(t_d, T_r) = \alpha \cdot T + \beta \cdot SDE + \gamma \cdot SDL + \theta \cdot DL \tag{2.47}$$

where $\alpha \equiv v_T/60$ is the per-minute value of travel time, β and γ are per-minute costs of early and late arrival, and θ is a fixed cost of arriving late. The functional notation $C(t_d,T_r)$ is to remind us that each of the components of trip cost depends on the departure time, t_d, and a random (unpredictable) component of travel time, $T_r \geq 0$. Our objective is to measure the increase in expected cost C due to the dispersion in T_r, given that t_d is subject to choice by the traveler. Letting C^* denote this expected cost after the user chooses t_d optimally, we have

$$C^* = \underset{t_d}{Min}\, E\left[C(t_d, t_r)\right] = \underset{t_d}{Min}\left[\alpha \cdot E(T) + \beta \cdot E(SDE) + \gamma \cdot E(SDL) + \theta \cdot P_L\right] \tag{2.48}$$

where E denotes an expected value taken over the distribution of T_r, and where $P_L \equiv E(DL)$ is the probability of being late. This equation can form the basis for specifying the reliability term in a model like (2.25). It captures the effect of travel time uncertainty upon expected schedule delay costs, but may omit other reasons why uncertainty could cause disutility.

To focus just on reliability, let's ignore the dynamics of congestion for now by assuming that E(T) is independent of departure time. Remarkably, the optimal value of t_d then does not depend on the distribution of T_r, provided that its probability density is everywhere finite. To find this optimal departure time, let $f(T_r)$ be this probability density function, T_f the travel time when T_r=0, and t^* the desired arrival time at the destination. The next to last term in the square brackets of (2.48) can then be written as

[26] See for example MVA Consultancy et al. (1987), p. 149; Small, Noland, Chu, and Lewis (1999); and Hensher (2001).

$$\gamma \cdot E(SDL) = \gamma \cdot E(t_d + T_r - \tilde{t} \mid T_r > \tilde{t} - t_d)$$
$$= \gamma \cdot \int_{\tilde{t}-t_d}^{\infty} (t_d + T_r - \tilde{t}) \cdot f(T_r) dT_r$$

where $\tilde{t} \equiv t^* - T_f$ is the time the traveler would depart if T_r were equal to zero with certainty. Differentiating yields:

$$\frac{d}{dt_d} \gamma \cdot E(SDL) = 0 + \gamma \cdot \int_{\tilde{t}-t_d}^{\infty} \left[\frac{d}{dt_d} (t_d + T_r - \tilde{t}) \cdot f(T_r) \right] dT_r = \gamma P_L^*$$

where P_L^* is the optimal value of the probability of being late.[27] Similarly, differentiating the term involving β in (2.48) yields $-\beta \cdot (1 - P_L^*)$. Finally, differentiating the last term yields -θf^0 where $f^0 \equiv f(\tilde{t} - t_d^*)$ is the probability density at the point where the traveler is neither early nor late. Combining all three terms and setting them equal to zero gives the first-order condition for optimal departure time:

$$P_L^* = \frac{\beta + \theta f^0}{\beta + \gamma}. \tag{2.49}$$

In general this does not yield a closed-form solution for t_d^* because f^0 depends on t_d^*. However, in the special case θ=0, it yields $P_L^* = \beta/(\beta + \gamma)$, a very intuitive rule for setting departure time that is noted by Bates *et al.* (2001, p. 202). The rule balances the aversions to early and late arrival.

The cost function itself has been derived in closed form for two cases: a uniform distribution and an exponential distribution for T_r. In the case of a uniform distribution with range b, (2.49) again simplifies to a closed form:

$$P_L^* = \frac{\beta + (\theta / b)}{\beta + \gamma}.$$

[27] The term "0" in this equation arises from differentiating the lower limit of integration: $-[d(\tilde{t} - t_d)/dt_d] \cdot [(t_d + T_r - \tilde{t}) \cdot f(T_r)]_{T_r = \tilde{t} - t_d} = 1 \cdot 0 = 0$.

The value of C^* in this case is given by Noland and Small (1995) and Bates *et al.* (2001). In the special case θ=0, it is equal to the cost of expected travel time, $\alpha \cdot E(T)$, plus the following cost of unreliability:

$$v_R = \left(\frac{\beta\gamma}{\beta+\gamma} \right) \cdot \frac{b}{2} \ . \tag{2.50}$$

The quantity in parentheses is a composite measure of the unit costs of scheduling mismatch, which plays a central role in the cost functions considered in the next chapter. Thus (2.50) indicates that reliability cost derives from the combination of costly scheduling mismatches and dispersion in travel time.

More generally, the last two terms in (2.48) are potentially important if $\gamma > \beta$ or if θ is large, conditions that are in fact true according to the empirical findings in (2.24). These terms are sensitive to the values of $E(SDL)$ and P_L, which depend especially on the shape of the distribution of T_r in its upper ranges, since this determines the likelihood that T_r takes a high enough value to make the traveler late. Thus we might expect the expected cost of unreliability to depend more on this part of the distribution (its "upper tail") than on other parts.

Equation (2.50) applies equally to the expected cost of schedule mismatches on a transit trip, under the common assumption that people arrive at a transit stop at a steady rate, if b is reinterpreted as the headway between transit vehicles.[28] Although under that interpretation v_R is proportional to expected waiting time, it is *not* a representation of waiting-time cost but rather must be added to it. In the case where the transit headway is itself uncertain, or where the vehicle might be too full to accommodate another passenger, the derivation of reliability cost for transit becomes much more complicated (Bates *et al*., 2001).

2.6.5 Empirical Results

Research has generated an enormous literature on empirical estimates of value of time, and a much smaller one on value of reliability. Here we rely mainly on reviews of this literature by others.

[28] This is pointed out by Wardman (2004, p. 364), who attributes the point to unpublished work by John Bates.

Waters (1996) reviews 56 value-of-time estimates from 14 different nations. Each is stated as a fraction of the gross wage rate. Focusing on those where the context is commuting by automobile, he finds an average ratio of VOT to wage rate of 48 percent, and a median ratio of 42 percent. He suggests that "a representative [VOT] for auto commuting would be in the 35 to 50 percent range, probably at the upper end of this range for North America." Consistent with this last statement, both Transport Canada (1994, sect. 7.3.2) and US Department of Transportation (1997) currently recommend using a ratio of 50 percent for personal travel by automobile.

Reviewing studies for the UK, Wardman (1998, Table 6) finds an average VOT of £3.58/hour in late 1994 prices, which is 52% of the corresponding wage rate.[29] Mackie *et al*. (2003), reviewing a larger set of UK studies, recommend best hourly values for VOT of £3.96 for commuting and £3.54 for other trips at 1997 prices; their average is 51% of the relevant wage rate.[30] Gunn (2001) find that Dutch values used in 1988, differentiated by level of household income, track well various British results for a similar time. However, Gunn reports that there was a substantial unexplained downward shift in the profile for 1997 – a phenomenon possibly resulting from better amenities in vehicle. Another Dutch study – using a novel methodology in which the "choice" is job termination rather than mode or route – finds a ratio of VOT to wage rate of one-third for shorter commutes (less than one hour round trip) and two-thirds for longer ones, for an average of "almost half" (Van Ommeren, Van den Berg, and Gorter, 2000). A French review by the Commissariat Général du Plan (2001, p. 42) finds VOT to be 77 and 42 percent of the wage for commuting and other urban trips, respectively, for an average of 59 percent. Finally, a Japanese review suggests using 2,333 yen/hour for weekday automobile travel in 1999, which was 84 percent of the wage rate.[31]

There is considerable evidence that value of time rises with income but less than proportionally, which makes the expression of VOT as a fraction of the wage rate, as above,

[29] Mean gross hourly earnings for the UK were £6.79 and £7.07/hour in spring 1994 and 1995, respectively. Source: UK National Statistics Online (2004, Table 38).

[30] Mean gross hourly earnings in 1997 were £7.42/hour, from same source as previous footnote.

[31] Japan Research Institute Study Group on Road Investment Evaluation (2000), Table 3-2-2, using car occupancy of 1.44 (p. 52). Average wage rate is calculated as cash earnings divided by hours worked, from Japan Ministry of Health, Labour and Welfare (1999).

somewhat less attractive. An earlier set of studies in England found that although members of the highest-income group had incomes more than three times those of the lowest group, their values of time were only 30 to 40 percent higher.[32] Similarly, Mackie *et al.* (2003) find that values of time for the highest of three broad income groups are 1.5 to 2.4 times those for the lowest group. The easiest way to summarize this issue is in an elasticity of value of time with respect to income. Wardman (2001, p. 116), using a formal meta-analysis, finds that elasticity to be 0.51 when income is measured as gross domestic product per capita; with a larger sample he obtains 0.72 (Wardman, 2004, p. 373), and he is part of a group that recommends using an elasticity of 0.8 (Mackie *et al.*, 2003). These elasticities could be subject to a downward bias if there is indeed a downward trend, independent of income, as suggested by Gunn.

Wardman's (2001) meta-analysis is especially useful for tracking the effects of various trip attributes on value of time. For example, there is a 16 percent differential between value of time for commuting and leisure trips, and considerable differences across modes, with bus riders having a lower than average value and rail riders a higher than average value – possibly due to self-selection by speed.

Most important, walking and waiting time are valued much higher than in-vehicle time – a universal finding conventionally summarized as 2 to 2-1/2 times as high. Wardman actually gets a considerably smaller differential, namely a ratio of 1.62, which is quite precisely estimated; nevertheless Mackie *et al*. (2003) recommend using a ratio of 2.0. There is considerable dispersion in the reported estimates of these relative valuations, especially in the relative value of waiting time (MVA Consultancy et al., p. 130). This may indicate that the disutility of transfers (which entail waiting as well as other possible difficulties) is quite variable, and suggests a payoff from research into the sources of this variation.

A number of studies have been carried out using Chilean data. Munizaga et al. (2004), using an innovative model that combines choices of activities and travel modes by residents of Santiago, obtain average VOT equal to 46% and 67% of the wage rate for middle and upper income groups, respectively.

SP data often yield considerably smaller values of time than RP data. For example, Hensher (1997) and Calfee and Wintson (1998) obtain values using SP surveys of car commuters

[32] MVA Consultancy et al., pp. 133-135, 150, 152.

of 19 percent and 20 percent, respectively, of the wage rate.[33] Brownstone and Small (2005) take advantage of three data sets, all from "high occupancy toll lane" facilities in southern California, that obtained RP and SP data from comparable populations, in some cases from the same individuals. They find that SP results for VOT are one-third to one-half the corresponding RP results,[34] the latter being 50-90 percent of the wage rate. One possible explanation for this difference is hinted at by the finding, from other studies of these same corridors, that people overestimate the actual time savings from the toll roads by roughly a factor of two; thus when answering SP survey questions, they may indicate a per-minute willingness to pay for *perceived* time savings that is lower than their willingness to pay for *actual* time savings. If one wants to use a VOT for purposes of policy analysis, one needs it to correspond to actual travel time since that is typically the variable considered in the analysis. Therefore if RP and SP values differ when both are accurately measured, it is the RP values that are relevant for most purposes.

From this evidence, it appears that the value of time for personal journeys is almost always between 20 and 90 percent of the gross wage rate, most often averaging close to 50 percent. Although it varies somewhat less than proportionally with income, it is close enough to proportional to make its expression as a fraction of the wage rate a good approximation and more useful than expression as an absolute amount. (This is not to prejudge whether it may be desirable to use a constant absolute amount in cost-benefit analysis for political or distributional reasons, a subject we consider in chapter 5.) There is universal agreement that value of time is much higher for travel while on business, generally taken as 100 percent of total compensation including benefits. The value of walking and waiting time for transit trips is probably 1.6 to 2.0 times that of in-vehicle time, not counting some context-specific disutility of having to transfer from one vehicle to another.

Several studies have applied mixed logit to measure variation from unobserved sources in the disutility of time and reliability. Hensher (2001) allows for random coefficients of three types of travel time, using SP data on New Zealand commuters, resulting in standard deviations of

[33] This statement is based on Calfee and Winston's summary of the average over the entire sample (p. 91), and on Hensher's Table 3.7 (p. 274), panel for "private commute," using his preferred VOT of $4.35/hour.

[34] See their Table 1, rows 4-5, 13-14

VOT equal to 41-58 percent of the corresponding mean VOT.[35] The California studies reviewed by Brownstone and Small (2005) measure heterogeneity as the inter-quartile range (75th minus 25th percentile values) of the distribution of VOT or VOR with that measure, unobserved heterogeneity in VOT (i.e., that due just to random coefficients) is 55-125 percent of median VOT in two cases using RP data, and 144 percent in one case using SP data.

There has been far less empirical research on value of reliability. Almost all of it has been based on SP data, for at least two reasons: it is difficult to measure unreliability in actual situations, and unreliability tends to be correlated with travel time itself. However, a few recent studies have had some success with RP data. One key development is to measure unreliability as a property of the upper percentiles of the distribution of travel times, as suggested by the theory discussed earlier. It turns out that such a measure is less correlated with travel time than is a symmetric measure like standard deviation, because the upper-percentile travel times (i.e., travel times that occur only rarely) tend to arise from incidents such as accidents or stalled vehicles. The occurrence of such incidents is closely correlated to congestion, but the delays they cause are less so because the effects of the incident persist long after it occurs.

Bates *et al.* (2001) review several SP studies of car travel that define unreliability as the standard deviation of travel time. Those that they deem most free of methodological problems produce a value of reliability (VOR), expressed in units of money per unit increase in that standard deviation, on the order of 0.8 to 1.3 times the value of time (VOT). Brownstone and Small (2005) review studies in which unreliability is defined as the difference between the 90th and 50th percentile of the travel-time distribution across days, or some similar measure. In those studies also, VOR tends to be of about the same magnitude as VOT. One of those studies, using data from the high-occupancy toll (HOT) lane on State Route 91 in the Los Angeles region, finds that roughly two-thirds of the advantage of the HOT lane to the average traveler is due to its lower travel time and one-third is due to its higher reliability.[36]

If reliability is not controlled for in studies of value of time, the estimated VOT may include some aversion to unreliability to the extent that time and unreliability are correlated. Nevertheless, the studies reviewed by Brownstone and Small (2005) obtain high VOT for

[35] This statement is based on Hensher's Table 3, Model 3a, the lower panel showing values of time in which the cost coefficient is that on a variable measuring the toll.

[36] An updated version of that study is Small, Winston, and Yan (2005).

automobile users even when simultaneously measuring VOR. In the case of the value of waiting time for public transit, the bias is measured explicitly by (2.50) under the interpretation stated earlier, in which (2.50) is proportional to expected waiting time $b/2$.

Turning to freight transportation, it is clear that values of time and reliability are important, but empirical evidence is sparse and definitions inconsistent. Most studies use SP methodology and many involve mostly involving inter-city travel. De Jong (2000) provides a recent review of studies, which suggests that for countries like The Netherlands, where a high proportion of travel is urban, values of time are quite high, consistent with conventional theory that travel time for road-freight vehicles it is viewed similar to business time and includes some inventory value for equipment and payload. Kawamura (2000), Wigan *et al.* (2000), and Fowkes *et al.* (2004) provide some evidence on values of both travel time and reliability.

2.7 Conclusions

All tractable approaches to travel-demand analysis are based upon greatly simplified portrayals of travel behavior. This is necessary because the purposes of travel and the variety of choices available make travel choices so complex. As a result, distinct or even mutually contradictory analytical approaches may each provide useful information for particular circumstances, and the sophisticated planner will want to understand a wide variety of approaches.

Both aggregate and disaggregate models can be instructive, the choice between them depending in any given instance on availability of micro data and on how important it is to have an explicit representation of individual decision-making processes. Many of the problems plaguing the traditional planning process are not inherent in aggregate models, but rather in simplifications that obscure important feedback effects. Disaggregate models, even if they have not always improved forecasting accuracy, have performed well in many circumstances and have enabled researchers to undertake new and sophisticated types of policy analysis. They have also enriched our understanding of how variability affects travel behavior, and they have given new insight into aggregate measures of attractiveness, accessibility, and welfare.

The theory of time allocation is well developed and permits us to rigorously address conceptual issues concerning value of time and reliability. Despite uncertainty, a consensus has developed over many of the most important empirical magnitudes for values of time, permitting

them to be used confidently in benefit assessment. Another decade should bring similar consensus to value of reliability.

LIST OF BOOKMARKS—CHAPTER 2

Original eqn #	New eqn #	bookmark name
2.1	2.1	
2.2	2.2	aggbinlog
2.3	2.3	aggmnl
2.4	2.4	Ujn
2.5	2.5	Pin
2.6	2.6	P1n
2.7	2.7	PrEV
2.8	2.8	Prlogit
2.9	2.9	VINex
2.10	2.10	VIN
2.11	2.11	LIK
2.12	2.12	VINalpha
2.13	2.13	VT
2.14	2.14	VINex2
2.15	2.15	VT2
2.21	2.17	W
2.22	2.18	Roy
2.23	2.19	P2n
2.24	2.20	dP2ndc2
2.26	2.21	McfReid
	2.22	OrderedProb
2.37	2.23	TrainModel
2.38	2.24	TripTim
	2.25	VLamSmall
2.28	2.26	PiGEV
	2.27	EMaxUGEV
2.29	2.28	GNLBinary
2.30	2.29	PiNL
2.31	2.30	PiNLupperBin
2.32	2.31	PiNLlower
2.33	2.32	Incval
	2.33	GNL
2.35	2.34	WNL
2.36	2.35	
	2.36	AutoUse
	2.37	QDef
	2.38	AutoOwn
	2.39	PinConditional
	2.40	PinUnconditional

2.39	2.41	Lagrangian
2.40	2.42	UTk
2.41	2.43	UTw
	2.44	vTkDef
2.42	2.45	vTk
	2.46	dVdck
	2.47	SDCost
	2.48	SDCostStar
	2.49	PLStar
	2.50	CStar

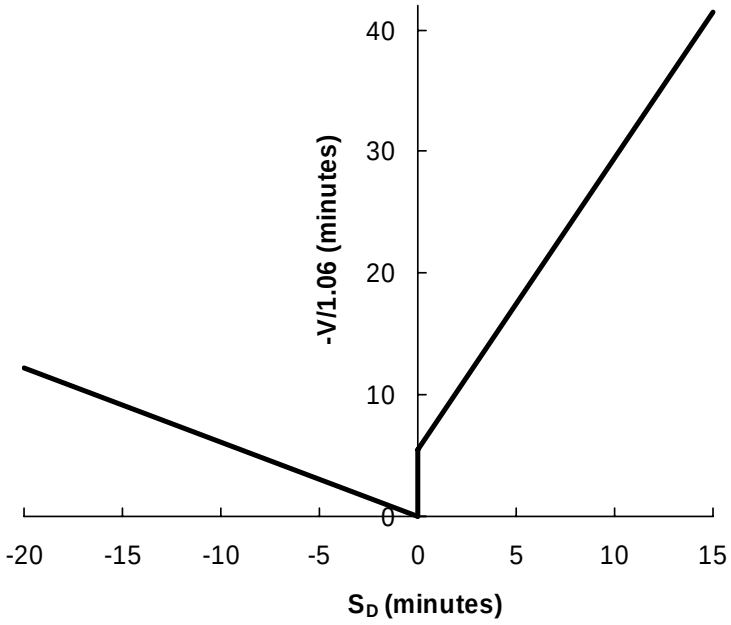


Figure 2.1 Disutility of Schedule Delay

3. COSTS

Having surveyed the demand for travel in the previous chapter, we now turn to its supply, that is, the conditions determining how much travel of various types can be accomplished at what prices. Just as travel demand is multidimensional, so a full analysis of the supply of transportation involves many facets, including multiple outputs, complex price structures, dimensions of service quality, and alternative forms of industrial organization.

It is useful to separate supply analysis into different parts. The first part, the subject of this chapter, consists of describing the technologies and factor supplies faced by transportation providers, information that is usefully summarized as cost functions. Other parts, in Chapters 4 through 6, consist of pricing, investment, and strategic decisions; these analyses involve transportation providers' economic behavior, resulting market outcomes, and normative criteria by which policy makers might like to influence those outcomes.

Because service quality is so important to the demand for transportation, we must also include it in any supply analysis. One way to do so is to define quality dimensions for each output. This is conceptually natural, but cumbersome. Another way is to view consumers as part of the production process, as in Becker's (1965) theory of household production; the level of service quality, like any other productive input, is then determined by conditions for efficient production. This approach, adopted here, treats user-supplied inputs, such as time, as though purchased in markets at prices equal to the values that are determined from demand analysis. In so doing, it moves such user inputs from the demand side to the supply side of the analysis and embeds them directly into cost functions.

Knowledge of cost functions enables us to answer questions about the relative efficiency of various types of transportation and about the relative importance of various parts of the production process such as capital, user time, operator wages, public facilities, and even unintended spillovers to nonusers. The discussion begins in the next section with basic cost concepts. Section 3.2 then surveys our knowledge of cost functions for public transit service. Sections 3.3 through 3.5 do the same for highway transportation, with an emphasis on private automobiles and congestion; these sections provide a variety of models of congestion and synthesize our knowledge of key quantitative parameters affecting the social cost of automobile

transportation. Section 3.6 briefly compares the average costs of particular types of trips by various private and public modes.

3.1 The Nature of Cost Functions

The literature on transportation cost contains much confusion that can be avoided by using standard economic concepts and terminology. Useful reviews include Jara-Díaz (1982), Braeutigam (1999), and Pels and Rietveld (2000). What follows is our own synthesis.

General Definitions

The basic description of technology is the *production function*, which describes the relationship between outputs and inputs:

$$F(q, x; \theta) = 0 \tag{3.1}$$

where q and x are vectors of outputs and inputs, respectively, and θ is a vector of parameters which may include service-quality descriptors. (Alternatively, services of different quality may be considered as different outputs in the vector q.)

The *cost function* for a given producer is the minimum cost of producing output vector q, given the production function and the supply relations for inputs. Usually, these supply relations are assumed to consist of a fixed price vector w, in which case the problem becomes minimizing input expenditures $w'x$ subject to the technology constraint (3.1). The solution, if unique, determines an optimal input vector x^*; the resulting minimum cost, $w'x^*$, depends on q, w, and θ so is written as the cost function $C(q,w;\theta)$. If instead the supply relations are represented by more elaborate factor-supply equations, w in what follows must be reinterpreted as a vector of parameters fully describing those factor equations.

If all inputs are included in x, including those that can be varied only over a long time period, we obtain a *long-run cost function*, which will be denoted $\widetilde{C}$ when the distinction is relevant. If instead one or more inputs are held fixed during the minimization, the resulting cost is called a *short-run cost function*. Typically, the fixed input is a measure of capital stock, say x_n;

its fixed value $\bar{x}_n$ becomes another argument of the resulting cost function, which we may write as $C(q, w; \theta, \bar{x}_n)$. By definition,

$$\tilde{C}(q, w; \theta) = \min_{x_n} C(q, w; \theta, x_n) . \tag{3.2}$$

Either the short- or long-run cost function may approach a positive constant C^0 as $q \to 0$. If so, C^0 is called the *fixed cost* and $C - C^0$ the *variable cost*. A short-run cost function always contains a fixed cost because it includes the carrying cost of fixed capital (e.g. $w_n \bar{x}_n$); the rest of the short-run cost is called *operating cost*, since it characterizes ongoing operations. Operating cost may contain a fixed component, for example maintaining the air supply in a subway tunnel or the paint on an automobile garaged outdoors. Fixed cost should not be confused with *sunk cost*, a dynamic concept which expresses irreversibility in starting a business: for example, the marketing analysis and initial advertising campaign that might initiate a new transit service. A fixed operating cost can be eliminated by closing down the service entirely, whereas a sunk cost cannot.

Letting C denote either a short- or long-run cost function, we may define marginal cost with respect to output q_i as $mc_i = \partial C / \partial q_i$. One can show from (3.2) that, as follows from the envelope theorem, the long-run marginal cost is equal to the short-run marginal cost with $\bar{x}_n$ set to x_n^*; this implies that if capital stock is optimal, the cost of producing a small increment of output is the same whether or not capital stock is allowed to vary.

Economies of Scale

Interest often centers on the degree of *scale economies*, s, which summarizes how fast costs rise with respect to output(s). If output q is a scalar, s is defined simply as the inverse of the output-elasticity of cost: letting $ac = C/q$ be average cost,

$$s \equiv \frac{ac}{mc} = \frac{C}{q \cdot (\partial C / \partial q)} . \tag{3.3}$$

If $mc < ac$ so that $s > 1$ (equivalently, if ac is falling in q), we have *economies of scale*. The opposite case ($s < 1$) is *diseconomies of scale*; and $s = 1$ defines a situation of *neither economies nor diseconomies of scale* or, more simply, *neutral scale economies*. Because a short-run cost

function has a larger fixed cost than the corresponding long-run cost function, it is more likely to show scale economies.[1]

If a firm sells output q at a price equal to its marginal cost, revenue is

$$R = q \cdot (\partial C / \partial q) = C / s \,. \tag{3.4}$$

Hence revenue will exactly cover total cost if there are neutral scale economies (s=1); scale diseconomies will produce a profit, while scale economies will produce a deficit. This observation makes it clear that an analysis of scale economies has significant implications for the financial terms at which marginal-cost pricing can take place, which makes it of interest in the study of regulation, competition, and public pricing.

This relationship between cost coverage and economies of scale generalizes readily to many outputs. Following Bailey and Friedlaender (1982), define s by the last equality in (3.3) but with the denominator reinterpreted as the inner product between output vector q and the gradient vector of the cost function. (This version of s is a measure of *multi-product scale economies*.) Then (3.4) again holds under marginal-cost pricing. In this case s can be related to a combination of individual-product scale economies and *economies of scope*, which measure the extent to which it is cheaper to produce several products within the same firm rather than in separate firms.

Definition of Outputs

The definitions just given can be made operational only by simplifying the complex production processes encountered in real life. For example, a transit agency does many things, only a few of which can be measured and analytically manipulated as outputs; it draws on many resources, only a few of which find their way into formal analysis as inputs. There is no one correct set of definitions; what matters is that the definitions chosen to study a particular phenomenon facilitate understanding and prediction.

---

[1] If the input price vector w is constant, then scale economies and diseconomies are equivalent to *increasing returns to scale* and *decreasing returns to scale*, respectively, in the production function (3.1). These terms refer to whether production rises more or less than proportionally when all inputs are increased together by the same proportion. As a result, scale economies and returns to scale are often treated as synonymous. However, a rising or falling supply price of a factor input can upset this relationship.

For transport cost analysis, it is useful to consider two classes of output. One, which we can call *final* or *demand-related outputs*, measures the quantity and/or extent of trips taken. This type of output corresponds to the variables studied in travel-demand analysis. A complete cost analysis would distinguish all the various kinds of trips produced, such as trips from Central London to Heathrow Airport during the afternoon rush hour. In practice, final outputs are usually aggregated in some manner for tractability – expressed for example as total passenger trips, revenue passengers (the number of distinct fares paid), unlinked passenger trips (the number of passenger boardings of distinct vehicles), passenger-miles, vehicle-miles, or even total revenues (a valid output measure if the fare structure is held constant in the analysis).

From the point of view of the transportation provider, however, final outputs are not under its control in the same way that, say, the number of chairs produced is under the control of a furniture manufacturer. No one would analyze a furniture manufacturer by counting as its output the number of its chairs that are occupied at any moment. Similarly, the transit firm may be more interested in the cost of producing the potential for trips — as measured, for example, by vehicle-miles, vehicle-hours, or seat-miles of service. We may consider such measures to be *intermediate outputs*, because they are combined with user time to produce the final outputs; they are also called *supply-related outputs*. Intermediate outputs are sometimes bought and sold as intermediate goods, for example when a public transit agency contracts to pay a private firm for a particular amount and type of bus service on a particular route, while the agency itself undertakes to use its marketing abilities to convert this service into actual trips taken.

Whether one measures cost functions in terms of final or intermediate outputs depends upon the purpose of the analysis. A study of the technical efficiency of firms' production would use intermediate outputs, whereas a study of the effectiveness of the firms' service offerings and marketing policies would use final outputs. One may also include both in a multi-output analysis.

Implicit in the definition of a cost function for producing *final* outputs is a decision rule for choosing *intermediate* outputs. For example, determining the minimum cost of producing passenger trips along a given bus route entails finding the cost-minimizing headway (the time interval between buses). This suggests a two-step strategy for analyzing transit service. In the first step, a cost function is defined in terms of intermediate outputs such as vehicle-miles, vehicle-hours, and peak vehicles in service. In the second step, a model is constructed to represent optimal choice of intermediate outputs, given the environment and final output

demands. A description of this environment might include the length of a corridor, the area from which it draws patronage, densities of trip origins and destinations, and possible methods by which passengers can access the system and reach their final destinations. This two-step model makes explicit the optimization of intermediate outputs, and thereby makes it possible to analyze a firms' operating policies as well as its technical production process.

Whatever the type of outputs considered, care should be taken when aggregating them into a manageable number of empirical measures. A pragmatic way of handling multiple outputs parsimoniously in cost functions is to choose aggregate measures of output (e.g. vehicle-miles) while allowing the function to depend also on descriptors of the operating environment (e.g. traffic speeds). It is especially important to retain the distinction between expanding the *density* of output, for example by adding more vehicle-miles on a given route network, and expanding the *spatial scale* of output, for example by extending service to new suburban locations. The former often allows more intense use of equipment, thereby lowering average cost—a form of scale economies called *economies of density*. In contrast, extending service to new locations may or may not involve scale economies: if it does they are called *economies of size*. Many transportation industries have been found to have economies of density but not of size (Braeutigam, 1999). However, some have argued that the usual forms of network aggregation defining economies of size are ambiguous and do not correspond to useful policy questions (Basso and Jara-Díaz, 2006).

Methods of Measurement

There are at least three general approaches to empirically measuring cost functions. The *accounting* approach examines the budgetary accounts of one or more enterprises, adjusts as needed to match economic concepts of opportunity cost, and then attributes specific accounts to specific outputs. The *engineering* approach builds a production function from technical descriptions of the production process and adds information about input prices. The *statistical* approach infers how cost varies with levels of output and other variables by observing the costs actually incurred in many different situations: for example, over many time periods or over a cross section of firms. The generality of the statistical approach has been greatly enlarged by techniques, pioneered by Spady and Friedlaender (1978), for estimating flexible functional forms such as the trans-log function (which is quadratic in logarithms of all variables). There is

considerable overlap among these three approaches, and any given study may make use of more than one.

External, Social, and Full Costs

Recent years have seen increased attention to costs that are borne not by the providing agency or the individual users of a given service, but by other parties. Examples abound: air pollution, noise, ground-water contamination, and wildlife disruption, to name a few. Such costs are called "external" because they fall on parties who are not part of the decision resulting in that cost (for example, a decision to fill a gas tank and thereby increase the volume of wholesale fuel deliveries with their attendant risks of spills). Those parties might be people who have themselves made similar decisions, but unless they do so as part of a collective (e.g. a tour operator), it can be presumed for the most part that each person disregards such external effects when deciding on travel arrangements.

Social or *full costs* are the total costs including any external costs. We define the *marginal social cost* (*msc*) of a particular travel movement (such as a vehicle-mile in an automobile) as the derivative of social cost with respect to that movement. The *msc* therefore includes both the marginal private cost as defined from a private cost function and the *marginal external cost* (*mec*), i.e. the effect of that movement on other parties. If private cost operates with neither economies nor diseconomies of scale, then marginal private cost equals average private costs. If furthermore the externality is fully mutual, in the sense that external costs are borne entirely by other travelers making the same kind of decision, then the average private cost is the same as average social cost (*ac*), so that $mec = msc - ac$. Congestion is usually modeled this way.

With a mutual externality, costs do not divide up neatly between the perpetrators and the recipients of the externality, because they are the same people. Therefore measures of "total external cost," for example obtained by multiplying *mec* by quantity, are not easy to interpret and generally not very useful. This is especially true because many externalities are only partly mutual. For example, carbon monoxide emissions from motor vehicles tend to remain close to the highway so their damage is borne partly by the parties producing it (drivers on that highway) and partly by third parties (pedestrians or nearby residents). As another example, motor vehicle injuries involve a complex mix of private costs, mutual external costs, and external costs borne

by non-motorists. Note that for the determination of *mec*, relevant for the formulation of efficient (tax) policies, it is immaterial whether or not the externality is mutual.

Numerous studies have attempted to quantify social costs, sometimes identifying also which are external and which are marginal to a particular movement. See, for example, Murphy and Delucchi (1998), Litman (2005), Nash *et al.* (2003), and the papers in Greene, Jones, and Delucchi (1997). Section 3.4.6 incorporates results of many such studies.

3.2 Cost Functions for Public Transit

This section examines some of the many attempts to measure the cost of providing bus or rail transit service. We adopt the two-step strategy described earlier: first we analyze the cost of producing intermediate outputs, then we use the results in explicit optimization models of the production of final outputs. The first three subsections that follow are mainly about the first step, describing three approaches to measuring cost functions; the last subsection covers the second step.

3.2.1 Accounting Cost Studies

Accounting cost studies seek to determine the relation between cost and intermediate outputs by examining cost accounts of transit agencies. Studies using this approach usually assume that cost is a linear function of a few measures of intermediate outputs such as route-miles *RM*, peak vehicles in service *PV*, vehicle-hours *VH*, and vehicle-miles *VM*:

$$C = c_1 \cdot RM + c_2 \cdot PV + c_3 \cdot VH + c_4 \cdot VM \ . \tag{3.5}$$

This approach involves "fully allocated costs," in the sense that all cost items are allocated to one and only one of the outputs: there are no fixed costs and no joint costs.[2]

[2] By joint cost we mean one that depends on the level of two or more outputs and cannot be disaggregated into output-specific costs. Button (1993) divides these into two types: "joint costs" where the outputs must unavoidably be produced together (e.g. train travel in opposite directions), and "common costs" where it is an economic choice to produce them together (e.g. carrying passengers and cargo on the same train).

We can use the outputs in this model to distinguish between economies of density and of size, by noting that *RM* is a measure of network size. the two types of scale economies described earlier. If all four outputs are expanded together, cost rises by the same percentage; so the cost function shows neither economies nor diseconomies of size. If route-miles are held fixed, however, cost rises less than proportionally to a simultaneous increase in the other three outputs (assuming $c_1>0$); so there are economies of density.

Many accounting studies were developed during the 1970s and 1980s due to the need for more fine-tuned cost information as part of policy developments involving deregulation and privatization of transit (Savage, 1988, 1989). Table 3.1 compares the results of two such studies that use equation (3.5). They attempt to provide figures that are comparable across modes, although only one (Boyd, Asher, and Wetzler 1973, 1978) includes any infrastructure for bus (an exclusive busway). Each study has a different strength: Allport (1981) draws from the accounts of a single agency (in Rotterdam) providing three types of transit, thereby eliminating some sources of difference in comparing across modes; whereas Boyd *et al*. draw from many transit agencies in Canada, U.S., and Mexico, thereby providing a more representative sample. From both studies, it is clear that capital costs for rail vehicles are much higher than for buses. The Allport study suggests that, for Rotterdam, any cost advantage of light rail (running on city streets) over heavy rail (with fully separated right of way) is confined primarily to lower capital and maintenance costs for its tracks; it is uncertain whether even this is a real cost advantage or a failure to account for the opportunity cost of public street space.

Naturally the cost of providing transit service depends on the balance of peak and offpeak service. The model of equation (3.5) portrays this dependence in two ways. First, peak service obviously determines the value of PV . Second, peak operations require a larger number of vehicle-hours VH to provide the same frequency of service because peak congestion slows those vehicles. In addition, one would expect driver costs per vehicle-hour to differ between peak and offpeak service , because peak periods are too short to constitute a full workday and therefore result in unproductive time and/or overtime pay for full-time drivers. To represent this, it is common to divide vehicle-hours into *base service*, VH_b, and *peak service*, VH_p, the former representing service at a constant rate over most of the day (including peak hours) and the latter representing additional service during peak hours only. (One could distinguish night or weekend

service as well, but we forego that complication.) This suggests the following modification of (3.5):

$$C = c_1 \cdot RM + c_2 \cdot PV + c_b \cdot VH_b + c_p \cdot VH_p + c_4 \cdot VM \ .$$

Analysis of staffing requirements in British and U.S. transit agencies has suggested to several authors that the ratio c_p/c_b is about 2.0 for bus systems;[3] empirical estimates range from 1.1 to 2.5.[4] Simulation studies of driver schedules, given work rules and overtime pay rates, suggest an approximate value of c_p/c_b=1.5 for typical conditions.[5] This figure is probably the best estimate currently available.

Since the extra cost incurred in peak service depends on work rules, it might be reduced through labor negotiations. Chomitz and Lave find that work-rule changes, especially hiring part-time drivers, could reduce total bus operating costs modestly, in most cases between 3 and 8 percent.

Abbas and Abd-Allah (1999) demonstrate how the cost accounts of a transit agency can be dissected to allocate costs among output categories, using accounts of the primary public transit provider for Cairo, Egypt. To aid in allocation, they use information about the activity (operation, maintenance, or administration) and travel mode(s) to which a given cost item pertains. Some of their results are summarized in Table 3.2. It is notable that nearly every type of unit cost is much higher for full-size bus than for minibus, and higher still for tram (streetcar). These differences are reduced but not eliminated if we divide by average capacity, as shown in the second panel. The aggregate percentages shown in the third panel portray a surprisingly high proportion of costs depending on peak vehicles in service, which effectively are fixed in the short run; this result is attributed by the authors to "overstaffing" of the transit system, the extent

---

[3] Boyd et al., 1978; Mohring, 1979; Jansson, 1980, pp. 57-58; Cervero, 1982, p. 70.

[4] These include: 1.1 for Albany, New York (Reilly, 1977); 1.1 or 1.9 for eastern San Francisco Bay Area (Cervero, 1982; Small, 1983a, p. 36); 1.3 for Los Angeles County (Cervero, 1982); and 2.5 for a number of British operators (McClenahan et al., 1978). The latter study also provided a ratio (relative to weekday base period) of 1.1 for Saturday and 1.4 for Sunday.

[5] Chomitz and Lave (1984). Their simulations also suggest a typical value for c_b/c_3 of 0.82. These statements are based on cases B and E of their Table 1, p. 67, using the column "13/10," which they indicate is most typical of work rules.

of which is suggested by the figures in the last row. (However, Egyptian wage rates are far lower than those in highly developed nations, so it is appropriate that capital equipment be used more intensively.)

3.2.2 Engineering Cost Studies

Engineering cost studies use detailed engineering information to construct cost functions in a 'bottom-up' fashion. The classic study by Meyer, Kain, and Wohl (1965) is a masterful example of engineering costing, supplemented by accounting and statistical methods. They estimate cost functions for several forms of public transit, as well as for automobile travel. The authors specify in great detail the characteristics of each mode, including engineering specifications and lifetimes for the physical infrastructure and vehicles, precise operational characteristics such as headways and station dwell times, and prices for all components. Many of these parameters are specified as functions of passenger volume and urban residential density. Some of the costs are estimated from firms' accounts and others from statistical analysis, but most come from actual price quotes, for example the prices of vehicles.

Meyer, Kain, and Wohl's cost estimates for highway construction are discussed in Section 3.5.3, and their overall results comparing costs for different modes are considered in Section 3.6.

3.2.3 Statistical Cost Studies

Statistical cost studies pool information from various transit agencies and/or time periods and use statistical inference to estimate the parameters of cost functions. These studies permit relaxing the assumption of linearity in cost functions, and so are especially useful for their results on scale economies.

Viton (1980b) uses translog functions (i.e., functions which are quadratic in the logarithms of the variables) to estimate a short-run operating-cost function for rapid rail operations, with vehicle-miles as output, using annual data for seven North American agencies in the years 1970-1980. Cost is specified as a function of output, input prices, and fixed capital stock. Because track length is fixed, the ratio of average to marginal cost is a measure of the degree of economies of density. The results are firm-specific, but generally suggest a U-shaped average cost curve, with strong economies of density for some smaller agencies (maximum

s=2.04) and strong diseconomies for some large ones (*s*=0.30 for New York).[6] The diseconomies found for New York, Chicago, and Philadelphia are evidence of congested operations on a too-small system of tracks. According to Viton's estimated model, the median agency's short-run average operating cost, in 2003 prices, is $5.03 per vehicle-mile.[7] Savage (1997) distinguishes between economies of density and size, finding that 22 US light and heavy rail systems operate with strong economies of density but close to neutral economies of size.

Turning to bus providers, Viton (1981b) estimates a short-run cost function on a 1975 cross-section of 54 U.S. city bus systems, using vehicle-miles as output. He then uses the results, along with engineering estimates of capital costs, to construct long-run costs under optimal capital utilization. The results again indicate a U-shaped average cost function, but a much flatter one than for rail: he finds mild scale economies for small firms (maximum *s*=1.16 for the smallest) and mild scale diseconomies for large firms (*s*=0.87 for the largest, Chicago). Long-run average cost, restated in 2003 prices, ranges from $2.73 to $4.65 per vehicle-mile.[8] Viton finds that most bus providers have a fleet that is considerably larger than the one he computes as optimal. Button and O'Donnell (1985), using 55 British bus agencies and passenger revenues as output, similarly find mild scale economies (up to 1.43) for small firms and diseconomies (down to 0.89) for large firms.[9]

Berechman (1993) and De Borger and Kerstens (2000) review these and other statistical studies, reaching the following conclusions. First, for bus providers, statistical evidence supports the conclusion of Viton (1981b): intermediate outputs such as vehicle-miles are produced with a mildly U-shaped relationship between average cost and output. Second, producing final outputs such as passenger trips is much more likely to entail scale economies. Third, rail systems exhibit much greater variability in scale economies; this is especially true in the short run because their capital stock may be too large or too small for the current operations. Fourth, however, there

---

[6]From Viton (1980b, Table 2, p. 251). We have transformed his definition of economies of density (*ED*) to ours according to *s*=1/(1-*ED*).

[7]From Viton (1980b, Table 3, p. 252), updated using the transportation component of the consumer price index.

[8]From Viton (1981b, Table V, p. 300), transforming his *SCE* to our *s* by *s*=1/(1-*SCE*).

[9]From Button and O'Donnell (1985, Fig. 1, p. 75), using the same definitional transformation as with Viton.

seems to be a bias toward operating with a larger than optimal capital stock, possibly due to incentives built into capital subsidy programs.

Wunsch (1996) provides a nice example of how knowledge from previous studies can guide statistical specification so as to get the most from a limited data set. Wunsch compiles cost data from a cross section of 178 separate operating agencies throughout western and northern Europe. Rather than estimate flexible functional forms from this rather small data set, he uses earlier work to justify the assumptions that (a) there are no scale economies in producing convoy-miles *VM* (where a convoy is one or more vehicles operated by a single driver, e.g. a train or bus); (b) labor costs follow a linear form, something like (3.5), and are proportional to the local wage rate; and (c) non-labor costs are allocable entirely to convoy-miles. The variant of (3.5) used for labor cost is:

$$LC^j = \left[c_{1a}^j \cdot TM^j + c_{1b}^3 \cdot Stations^j + c_3 \cdot VH^j + \left(c_{4a}^j + c_{4b}^j \cdot n^j\right) \cdot VM^j\right] \cdot \left(w / \overline{w}\right) \qquad (3.6)$$

where j=1,2,3 is a modal indicator representing bus, streetcar, or subway; *TM* is the number of track-miles in the streetcar or subway system (0 for bus); *Stations* is the number of subway stations (0 for bus and streetcar); n is the capacity of a convoy in persons (measured as square meters of floor space divided by four); w is the local wage rate; and $\overline{w}$ is the average wage rate over the sample. The study does not distinguish between peak and offpeak service. Because an agency's bus and streetcar operations, and sometimes its subways, are consolidated into a single account, this equation is estimated by simultaneously estimating one equation for total labor cost, $LC = \Sigma_j S^j LC^j$ (where S^j is the j-th modal share in convoy-miles), and another just for subway cost, LC^3, where separate observations on subway cost are available (16 agencies).[10] In order to avoid heteroscedasticity (differing variances across observations), equation (3.6) is divided by VM^j before aggregation and estimation.

[10] For 76 operating agencies, labor costs are not segregated from capital costs in the data. In these cases the left-hand side of (3.6) is replaced by total cost and an additional term $c_{1c}^j \cdot VM^j$, not multiplied by wage, is added to the equation to represent capital cost, with c_{1c}^j estimated.

An example will help clarify how to interpret the coefficients. Parameter c_3 is the cost of a vehicle-hour of service at wage rate $\overline{w}$; its estimate of 598 BF (Belgian francs) per convoy-hour, compared to $\overline{w}$=520 BF per hour, suggests that labor costing the same as 1.15 hours of driver time is required for every incremental convoy-hour of service provided. Furthermore, the relative size of $c_3 \cdot VH^j$ compared to the other terms in (3.6) determines the labor-cost elasticity with respect to speed, i.e., the percentage change in labor cost brought about by a one percent increase in speed. If all labor cost were proportional to *VH*, this elasticity would be -1; the actual elasticities estimated for average conditions are very different: -0.392 for bus, -0.121 for streetcar, and -0.047 for subway. These values are important in knowing how congestion affects transit costs, but their absolute values may be underestimates because of the assumption that all non-labor cost (for those properties where it could be isolated) is independent of speed.

Wunsch's results can also be used to estimate scale economies in providing vehicle capacity – which are the root source of economies of density. Thus, the results help us understand what passenger densities are required for rail modes (streetcar, subway, or a mixture of the two which Wunsch calls "light rail') to be cheaper than bus. As it happens, the coefficients c_{4b}^j of person-miles of capacity are estimated to be nearly identical for all transit modes, about 0.9 US cents per potential person-mile (at 2003 prices); i.e., the *incremental* operating cost of expanding passenger capacity is very similar among the three modes. However, due to the other terms, the *average* costs of passenger capacity decline with capacity, in a manner that turns out to be nearly identical for the two rail modes but very different for bus: namely, average capacity costs start higher but decline more rapidly in the case of rail (i.e., streetcar or subway). The capacity where they cross is about 400 potential passengers per convoy, which means the operating costs of rail modes are lower than bus if there is enough passenger density to require trains holding 400 or more people. This capacity is greater than that observed for any streetcar system, causing Wunsch to conclude that, in terms of operating costs, "streetcars do not fill a significant gap between buses and underground rail" (p. 171). Of course, to complete the comparison, we need to consider capital costs as well, which we take up in Sections 3.5 and 3.6.

3.2.4 Cost Functions Including User Inputs

As already noted, travelers must supply some inputs, especially their time, as part of producing final outputs such as trips. We illustrate here with public transit users, and in Sections 3.4-3.5 with users of private vehicles.

Transit users spend time accessing the system, waiting for vehicles, riding in vehicles, possibly transferring between vehicles, and getting to final destinations. This section considers just waiting time. The consequences of including waiting time as an input to the production of trips are dramatic, and similar consequences would follow from including time spent walking or transferring. Specifically, Mohring (1972) shows that when waiting-time costs are included, transit service is subject to strong economies of density in producing final outputs, even if such economies are absent for producing intermediate outputs.

We can demonstrate this proposition with a simplified version of Mohring's model for peak-period bus transit on a single route.[11] The measure of final output, q, is the number of passengers per peak hour on the route. It is produced using two inputs. First is the intermediate good V defined as vehicles passing a given bus stop per peak hour, produced at unit cost c_p. Second is a user-supplied input, aggregate waiting time per peak hour W, valued at unit cost α^W. Suppose average waiting time per passenger, W/q, is equal to half the headway, $1/V$. Aggregate costs to the bus agency and to the users, respectively, are then:

$$C_B = c_p V \; ; \;\; C_W = \frac{\alpha^W q}{2V} .$$

We choose V to minimize the sum of these costs, subject to a constraint imposed by bus capacity:

$$q \leq NV$$

where N is the total number of passengers a bus can pick up and drop off as it travels the entire route. That is, $N=n\ell/d$ where n is the physical capacity of the bus (maximum number of passengers at any time); d is the average passenger's trip length; and ℓ is the length of the route. Letting λ be the Lagrangian multiplier of the constraint, the first-order condition is:

---

[11] See Mohring (1976), pp. 145-146.

$$c_p - \frac{\alpha^W q}{2V^2} - N\lambda = 0 .$$

There are two possible solutions. If λ=0, indicating that buses are not full, the solution is (with a star denoting optimized choices):

$$V^* = \sqrt{\frac{\alpha^W}{2c_p}} \cdot \sqrt{q} \;; \qquad W^* = \frac{q}{2V^*} = \sqrt{\frac{c_p}{2\alpha^W}} \cdot \sqrt{q}$$

$$C_B^* = c_P \cdot V^* = \sqrt{\frac{\alpha^W \cdot c_p}{2}} \cdot \sqrt{q} \;; \qquad C_W^* = \alpha^W \cdot W^* = \sqrt{\frac{\alpha^W \cdot c_p}{2}} \cdot \sqrt{q} .$$

Two properties of this solution are worth noting. First, the optimal bus frequency V^* is proportional to the square root of the passenger density q; this is known as the *square root rule* for operating policy. Second, the cost function is also proportional to $\sqrt{q}$, which gives it economies of scale (i.e., of density): specifically, s=2. This implies that a marginal-cost price will not cover the total cost incurred by the transit provider. In fact, the optimal fare is zero, which is the difference between the marginal cost $\partial C/\partial q$ and the value of the inputs supplied by users, C_W/q. The intuition here is simple: if buses are not full, then it costs nothing to take another passenger. The aggregate subsidy is equal to the total cost of waiting time, C_W^*.

If λ>0, indicating the capacity constraint is binding, the solution is $V^*=q/N$, $W^*=q/(2\cdot V^*)=N/2$, $C_B^*=c_p\cdot q/N$, and $C_W^*=\alpha^W\cdot N/2$. Over the range of output for which this solution holds, the total cost function is linear in output and has fixed cost C_W^*. It thus again exhibits density economies, namely $s=1+(C_W^*/C_B^*)$; now s is greater, the greater are waiting-time costs compared to operating costs.[12]

So whether or not the constraint is binding, there are economies of density because either operating costs or waiting costs grow less than proportionally as output expands.

---

[12] Derivation: $s = ac / mc = [(C_B^*+C_W^*)/q] / (c_p/N) = (C_B^*+C_W^*)/ C_B^*$. Note however that $s\leq 2$ in an optimum, because otherwise we would have $C_W^* > C_B^*$, implying that a marginal increase in frequency V would reduce waiting costs by more (namely by C_W/V) than that it would raise the cost to the bus agency (namely by $c_p=C_B/V$) — contrary to the assumption that the agency is minimizing the sum of these costs.

Analogous models can be constructed to show how the transit operator could respond by increasing route density instead of, or in addition to, frequency along a route.[13] In this case it is savings in walking time as well as waiting time that account for increasing returns. Because there are now two ways the agency can save user cost by offering more service, optimal vehicle-miles offered grows more rapidly with passenger density—specifically, with its two-thirds power. As vehicle-miles are expanded, half of the increased service is configured so as to reduce waiting costs and the other half to reduce walking costs.

There are many ways this model can be made more realistic. We could consider offpeak travel as a separate output. We could consider the width of the peak period to be variable, and take into account the effect that peak broadening would have on parameter c_p (Kraus and Yoshida, 1999). We could take into account the effect on average speed of additional passengers boarding or leaving the vehicle, thereby obtaining a positive optimal fare even when buses are not full. Mohring (1972) and Kraus (1991) show that this last effect can be quite important. We could recognize that bus capacity is not absolute but rather influences the degree of crowding and probability of the first bus to arrive being too full to enter.

We could also allow bus capacity to be endogenous, chosen as part of overall cost minimization, and thereby estimate optimal bus size. This approach is adopted by Jansson (1980), Glaister (1986), and Nash (1988), who conclude that the optimal bus size is much smaller than the actual size in typical situations in Sweden and Britain. There is some evidence from partial bus deregulation in Britain that small firms using small buses do, in fact, find a niche when allowed to do so.

The fact that economies of density result from treating waiting time as a cost provides a fundamental insight into public transportation modes. These modes depend on matching a set of desired trips, each at a particular time and place, to available vehicles. Similar results hold for airlines (Douglas and Miller, 1974) and taxicabs (Frankena and Pautler, 1986). The insight does not depend upon a literal view of waiting time, but applies to any disutility created by infrequent service, for example deviations from most desired arrival times. The practical consequences of

[13] See Jansson (1980) and Nash (1988).

this insight depend on the precise service arrangements. If intermingling of services by more than one firm causes the user to care only about the firms' combined service frequency, the economies of density are industry-wide and firms confer externalities on one another. If, on the other hand, the user has to pre-commit to one firm, the economies are firm-specific and create a natural monopoly.

3.3 Highway Travel: Congestion Technology

The importance of the automobile in urban travel patterns has created great interest in how best to cope with the various costs that it imposes. This question can be addressed by defining and measuring cost functions for motor vehicles on highways. Doing so facilitates pricing and investment analyses, which are the central contributions of economics to public policy in this area. For example, questions about optimal pricing or privately owned highways can be addressed by applying standard economic tools to carefully defined cost functions. The use of cost functions also makes precise what it really costs society to undertake a particular kind of trip by motor vehicle.

We therefore analyze the costs of highway travel in this and the next two sections. This section presents the pure technology of highway congestion, a subject brought squarely into transportation analysis by Beckmann, McGuire and Winsten (1956). Because it is so crucial to urgent policy questions, we provide considerable detail. We also argue that the static model used in the standard economic analysis of congestion is not fully satisfactory, and present a dynamic model that is tractable for the analyses of the following sections. Section 3.4 then derives short-run cost functions, *i.e.*, those for fixed road capacity, and confronts them with demand functions to characterize short-run equilibrium. The approach in that section is to incorporate user time directly as a cost, thereby making the congestion technology an integral part of the cost function; it also reviews empirical evidence on the magnitudes of short-run variable costs. Section 3.5 adds information about infrastructure costs in order to compute long-run cost functions.

3.3.1 Fundamentals of Congestion

Highway congestion arises from many causes. Traffic forms queues at signals. Cars entering from side streets wait for gaps in traffic on a main highway. Cars traveling behind slower vehicles on two-lane roads must wait for gaps in oncoming traffic before passing.

We begin with uniform, stationary-state congestion on a homogeneous highway without traffic signals. When many vehicles try to use the highway simultaneously, the resulting high *density D* (number of vehicles per unit of distance) forces them to slow down for safety reasons, thereby reducing average vehicle *speed S*. One way to depict congestion, then, is as a functional relationship $S(D)$. An example is shown in quadrant *a* of Figure 3.1, in mirror-image form in which D increases toward the left.

FIGURE 3.1

We are also interested in traffic *flow* or *volume V*, defined as the number of vehicles passing a given point per unit time. Traffic flow is identically equal to the product of speed and density:

$$V \equiv D \cdot S\,, \tag{3.7}$$

which is consistent with its units of measure: vehicles/hour ≡ (vehicles/mile) · (miles/hour). Unless stated otherwise, we normalize V and D with respect to road width, so that "vehicles" becomes a shorthand for "vehicles/lane" in these definitions.[14]

Given identity (3.7), the congestion technology can be expressed equivalently as a functional relationship between *any* two of the variables V, D, and S. One is the *speed-flow relationship* $S(V)$ shown in quadrant *b*; it is defined over the region $V \in [0, V_K]$, where V_K is the per-lane *capacity* of the highway. As seen in the figure, the relation is double-valued; we refer to

[14] A more precise term than "vehicles" is "passenger-car equivalents," a measure that combines vehicles of different sizes and acceleration capabilities, each with a weight indicating its contribution to congestion. See Krammes and Crowley (1986). In some situations it is more accurate to assume that slow vehicles effectively occupy an entire lane, and treat the highway as two separate roadways (possibly interacting if heavy vehicles are allowed to pass each other); see OECD (1983), chap. IV.

the upper branch as *congested* or *normally congested* and the lower branch as *hypercongested*.[15] The third possible relationship, that between V and D, is called by Haight (1963, pp. 69-73) the *fundamental diagram of traffic flow*; it is shown (rotated clockwise by 90 degrees) in quadrant *c* of Figure 3.1. Haight shows that flow first rises and later falls as density increases from zero, as depicted in the diagram.

Figure 3.1 shows diagrammatically how one can derive any of these three relationships from any of the others. The unused quadrant, in the lower left, is simply a diagonal line to equate values of density on the left horizontal axis and the lower vertical axis. The figure also shows the density D_m and speed S_m that correspond to maximum flow V_K. All other allowed flow levels can result from either a congested or a hypercongested speed and density. For example, a zero flow prevails when there are no vehicles on the road (D=0), leading to the free-flow speed S_f; or when density reaches the value, known as the *jam density* D_j, that reduces speed to zero — this situation is shown at the origin of quadrant *b* and at the points marked D_j in quadrants *a* and *c*.

If these variables are defined over a very small region of time and space, the relationships shown in Figure 3.1 are instantaneous ones. Aggregate relationships can be built from them by relating the variables at neighboring times and places. This approach is used to build detailed computer models of a real facility (Coombe, 1989). However, much of the economic literature has used the instantaneous speed-flow relationship to analyze aggregate performance of an entire highway. This may work well for situations where conditions change only slowly over time and space, but in other situations the instantaneous flow past a single point may be quite different from the economic demand for travel on the highway as reflected in the number of vehicles attempting to enter it. We discuss this problem more completely in Section 3.4.

3.3.2 Empirical Speed-Flow Relationships

[15]Engineers often call the upper branch of the speed-flow relationship "free flow" and the lower branch "congested flow." In our terminology, which is more common in economics, "free flow" refers instead to the limiting condition as $V \rightarrow 0$ on the upper branch. We use the terms "congested" and "normally congested" interchangeably, depending on whether an explicit distinction with "hypercongested" is needed.

We begin with some empirical evidence concerning the fundamental diagram, and some extensions of it that have been found necessary to portray observed data.

Instantaneous Relationships

There is uncertainty about the true shapes of the curves in Figure 3.1 in the neighborhood of maximum flow V_k and density D_m. Figure 3.2 illustrates why. Fig. 3.2a plots observations of flow versus occupancy, a measure of density, for the Queen Elizabeth Way in Toronto. (Thus it is depicting the curve in Figure 3.1c after rotation by 90 degrees.) Although the two branches of the flow-density curve are reasonably well defined, the middle portion connecting them is obscured by scatter; and it is not clear whether the branches meet, i.e., that the relationship is continuous.

Similarly, the speed-density data plotted in Figure 3.2b, from the Santa Monica Freeway in Los Angeles, appear to reflect two distinct regimes that are connected, not by a continuous curve, but by a region where the relationship is only vaguely defined.

FIGURE 3.2: (a) Original Fig. 3.2; (b) Original Fig. 3.3a

At least two explanations have been put forth for the dispersion of observations where flow is close to capacity. Both posit two distinct flow regimes, one congested and the other hypercongested. One explanation is a measurement problem: since speed, flow, and density are in practice measured over a finite span of space and time, a given observation may inadvertently average two points from the two different regimes. Another explanation is that many intermediate-density observations correspond to disequilibrium conditions during transition between the congested and hypercongested regimes (Hall, Allen, and Gunter, 1986). Compounding matters is that in many data sets there is a paucity of observations at flows near capacity. This is presumably due to bottlenecks upstream or downstream from the point in question. Indeed, empirical estimates of speed-flow relationships are strongly influenced by nearby bottlenecks (Branston, 1976; Hall and Hall, 1990).

Despite these difficulties, many empirical estimates are reported in the literature; Hall (2002) provides a review. Some illustrative examples are described here.

Of historical interest is Greenshields' (1935) linear speed-density relationship, estimated on a two-lane, two-way road:

$$S = S_m \cdot \left[1 - \left(D / D_j\right)\right].$$

Applying identity (3.7) yields a parabolic speed-flow relationship. Later studies revealed that this parabolic shape is less accurate for larger highways, where instead speeds often remain constant, or nearly so, over a substantial range of flow levels.

Not surprisingly, the functional form used in the estimation may strongly affect the shape of the estimated function, even when the same data are used. For example, two separate research groups have fit speed-flow curves like that in Figure 3.1c to the same data from a four-lane section of the Washington (D.C.) Beltway. Boardman and Lave (1977) get the following result:[16]

$$V = 2490. - 0.523 \cdot \left(S - 35.34\right)^2$$

where V is in vehicles per hour per lane. The two solutions to this quadratic equation in S represent congestion and hypercongestion. Inman (1978) obtains:[17]

$$V^{(2.95)} = 3.351 \cdot 10^9 - 231.4 \cdot \left(S - 7.2\right)^{(4.06)}$$

where, for any quantity X, $X^{(a)}$ denotes the Box-Cox transformation of X, defined as $(X^a - 1)/a$. The Boardman-Lave and Inman curves are both plotted in Figure 3.3, along with the scatter of data points. It appears that the Boardman-Lave curve represents the data more faithfully; but neither curve captures well the previously mentioned tendency, obvious from the raw data, for expressway speed in the region of normal congestion to remain nearly constant throughout most of the range of volume-capacity ratios.[18]

---

[16] This is their preferred equation (20), p. 350, converted to units of vehicles per lane per hour by setting V=5q.

[17] This is calculated from Inman's equation (2), p. 23, with parameter estimates from the top row of his Table 1, p. 25. The units and the unreported constant X were supplied by Inman, private conversation, Sept. 1989; they are N=V/500, G=S/10, and X=36.488. The Inman curve is undefined for S < 7.2 mi./hr., and can show no backward-bending portion unless the exponent on the right-hand side is constrained to be precisely an even integer, as it is for the special case represented by the Boardman-Lave formula.

FIGURE 3.3: Original Fig. 3.4

Figure 3.4 shows two more recent "official" speed-flow relations, which do account for this phenomenon: the COBA11 and Highway Capacity Manual (HCM2000) formulations in the UK and US, respectively.[19] Both portray normal congestion, not hypercongestion, and do so with two joined segments. HCM2000 joins a completely flat segment with a segment defined by a power function, with continuous derivative; whereas COBA11 joins two linear segments with a discontinuous slope. The parameters depend on highway geometry, the share of heavy vehicles, and other factors. Also shown in the figure are dotted lines representing two other regimes posited by Hall, Hurdle, and Banks (1992). One represents queue discharge within a bottleneck, and the other flow within a queue behind the bottleneck. In Section 3.4, we shall see that these regimes and their descriptions are consistent with emerging views on the hypercongested branch of the speed-flow relationship.

Figure 3.4

Space-Averaged Relationships

The relationship holding at a single time and place is not by itself useful for economic analysis of congestion, because it does not relate service quality for entire trips to the number of people attempting to travel. On real highways, queues form behind bottlenecks and traffic volumes vary over time and place. One way to take such features into account is through formal network models, in which a speed-flow relationship applies to each link and users choose the resulting quickest routes (Marcotte and Nguyen, 1998). Here we consider simpler approaches, namely averaging over space or time.

---

[18]This tendency is also noted by Banks (1989) for San Diego, and by Hall and Hall (1990) for Toronto, in both cases for short uniform stretches of highway unaffected by a bottleneck.

[19] See UK Department for Transport (2002), US Transportation Research Board (2000), and Smith, Hall, and Montgomery (1996).

Keeler and Small (1977) use observations on three expressways in the San Francisco Bay Area to estimate quadratic functions (like Boardman and Lave's) relating speed and volume-capacity ratio V/V_k, each averaged over a long stretch of highway. One resulting equation is:[20]

$$\frac{V}{V_K} = 0.8603 - 0.001923 \cdot (S - 45.68)^2 .$$

In contrast to the instantaneous speed-flow curves, Keeler and Small's averaged curves show, on the upper (congested) branch, a substantial negative slope over the entire range of average vehicle flow. This reflects the fact that as traffic is added to a real highway, non-uniformities in the highway design or in demand patterns cause minor slowdowns even when the average volume-capacity ratio is well below one.

Clearly, the exact nature of an aggregate speed-flow relationship depends on the extent and nature of heterogeneity, so a curve fitted for one highway is unlikely to generalize.This is all the more true for streets and arterial highways subject to congestion at signalized intersections, whose technology is entirely different from that of an unobstructed highway. Here there are even more sources of heterogeneity including signal timing, turn lanes, intersection geometry, and on-street parking. Smeed (1968, p. 34) reports the following relationship, attributed to Wardrop, for city streets in London:

$$\frac{V}{w} = 68 - 0.13 \cdot S^2$$

where w is width of road in feet. This relationship does not have a hypercongested branch, but rather approaches zero speed, implying infinite density, as $V \rightarrow 68w$.

For some purposes, it may be more useful to model speeds and flows over areas rather than along a single roadway. Ardekani and Herman (1987) use time-lapse aerial photography, combined with ground measurement of volumes, to estimate the following relationship between averaged values within the central area of Austin, Texas:[21]

---

[20]This is Keeler and Small's equation (12), p. 11, for the Eastshore Freeway, rewritten to make transparent the maximum V/V_K ratio and the maximum speed.

[21]Their equation (8), with values n=1.58 and v_m=60/1.95 as indicated just above their equation (6) on p. 6.

$$S = 18.38 \cdot \left[1 - (0.01D)^{1.239}\right]^{2.58} \tag{3.8}$$

where D is vehicle density per lane. They verify separately that (3.7) holds, to a good approximation, for their space-averaged quantities; this enables (3.8) to be converted into a speed-flow curve, for which maximum flow occurs at 9.1 miles per hour and which does include a hypercongested branch.

Time-Averaged Relationships

The space-averaged relationships just described cannot tell us what happens when demand exceeds the maximum flow. In such situations, speed depends not only on contemporaneous flow but on past flows, usually via queuing. Time-averaged speed-flow functions incorporate such time dependence by relating average speed over a specified period to the average vehicle inflow over that period.

Two functional forms that allow for traffic flow above capacity are in common use. The first specifies travel delay as a power function of the volume-capacity ratio:

$$T = T_f \cdot \left[1 + a \cdot (V / V_K)^b\right] \tag{3.9}$$

where T denotes travel time per mile (the inverse of speed). This function has been used in many economic models of congestion such as Vickrey (1963), with parameter b typically assumed to be between 2.5 and 5. With parameter values a=0.15 and b=4, it is known as the Bureau of Public Roads (BPR) function, used widely in U.S. transportation planning.[22] With values a=0.2 (freeways) or 0.05 (arterials), and b=10, it is known as the "updated BPR function" (denoted BPR-U below), derived by Skabardonis and Dowling (1996) to approximate the speed-flow functions in the 1994 Highway Capacity Manual, a predecessor to HCM2000 discussed above. While the potentially unlimited flow permitted by (3.9) might seem unrealistic, if it is used sensibly the traffic will be limited by other factors such as total demand or upstream bottlenecks.

---

[22] See U.S. Bureau of Public Roads (1964). The BPR function has been incorporated into the Urban Transportation Planning Process computer software to describe a single link in a network.

One drawback of (3.9) is that it does not account for how long traffic exceeds capacity. This disadvantage is remedied in a duration-dependent function derived by Small (1983a) to express the average travel time over a peak period of fixed duration P, when peak-period inflow V is at a uniform rate and delay results from queuing behind a single bottleneck with a constant capacity V_K. It yields a piecewise-linear relationship:[23]

$$T = \begin{cases} T_f & \text{if } V \le V_K \\ T_f + \frac{1}{2} \cdot P \cdot (V / V_K - 1) & \text{if } V > V_K . \end{cases} \tag{3.10}$$

Both equations (3.9) and (3.10), when estimated using nonlinear least-squares with simulation-based data points reported by Dewees (1978), fit these data surprisingly well.[24] Similarly, Small (1983a, pp. 32-33) finds that (3.10) approximates the pattern of travel times during the afternoon peak period on an eleven-mile stretch of freeway in the San Francisco Bay area.

Akçelik (1991) develops a travel-time function that is smooth, like (3.9), and that also approaches linearity for very high flows, like (3.10). It introduces a "delay parameter" J_a that is motivated by stochastic queuing models with random arrivals. The function is:

$$T = T_f + 0.25P \cdot \left[\left(\frac{V}{V_K} - 1 \right) + \sqrt{\left(\frac{V}{V_K} - 1 \right)^2 + \frac{8 J_a V / V_K}{V_K P}} \right], \tag{3.11}$$

which has (3.10) as a special case when J_a=0.[25] Akçelik's function seems to produce reasonable results when used in network models (Dowling, Singh, and Cheng, 1998).

Figure 3.5 compares the relationships depicted by (3.9), (3.10), and (3.11) by plotting normalized average speed (relative to S_f) versus normalized inflow (relative to V_K) for P =1 hour,

---

[23] Cassidy and Bertini (1999) find empirical evidence that discharge rates from bottlenecks fall after queue formation and then partially recover. The constant flow assumption nevertheless appears a reasonable approximation to observed behavior.

[24] The data are from Toronto arterials. We used nonlinear least squares with the nine data points reported by Dewees. Estimated parameters for equation (3.9), with V_K arbitrarily set to 1000 veh/hr, are T_f=2.48 min, a=0.102, and b=4.08. Estimated parameters for equation (3.10) are T_f=3.07 min, P=14.44 min, and V_K=1357 veh/hr.

[25] Akçelik proposes the following parameters $\{V_K, S_f, J_a\}$ for different types of roads, where S_f=L/T_f (mi/hr) and L is the length of the road: freeway {2000,75,0.1}; uninterrupted arterial {1800,62,0.2}; interrupted arterial {1200,50,0.4}; interrupted secondary {900,37,0.8}; high-friction secondary {600,25,1.6}.

V_K=2000 veh/hr, J_A=0.1, and two different values of T_f, one corresponding to a one-mile long road with top speed 75 mi/hr and the other to a 75-mile long road with the same top speed. The BPR, BPR-U, and PL curves represent BPR, updated BPR, and piecewise-linear functions, respectively, the latter from (3.10).

FIGURE 3.5

For the shorter road, the speed-flow curves derived from Small's and Akçelik's models are similar over most of the range considered, but diverge for inflows near V_K. For the longer road, the two functions become graphically indistinguishable; this is because the absolute travel delay in these functions, arising from bottleneck queueing, is unaffected by the length of the road and thus is diluted when averaging over the longer road. The BPR speed-flow functions, by contrast, do not depend on road length. The BPR functions are less steep than the two time-dependent functions for the short road, but steeper for the long road. These differences reflect the focus of the BPR functions on flow congestion along the entire length of the roadway.

Can such formulations be generalized to a dense street network, such as the downtown of a large city? Olszewski and Suchorzewski (1987) discuss ways to define the capacity of a downtown street network in Warsaw, Poland. May, Shepherd, and Bates (2000) go further by defining a matrix of origin-destination demands on a simulated network and multiplying it by a series of scalars to represent increasing demand. They find that in terms of averaged flow and speed on the network itself, a backward-bending speed-flow relation as depicted in Figure 3.1b still applies, suggesting the existence of hypercongestion for the area as a whole. But if they account also for queuing times on approaches to the network, they find that trip travel times increase monotonically with demand, yielding average speeds like the curves in Figure 3.5. This could suggest a modeling approach for downtown areas in which flow greater than capacity creates both hypercongestion and bottleneck queuing, an approach developed by Small and Chu (2003) as discussed later.

3.3.3 Dynamic Congestion Models

Queuing at a Bottleneck

Although the functions just discussed explicitly incorporate queuing, they take demanded flow as given and as constant over the period of interest. A dynamic formulation with queuing behind a bottleneck can deal better with extreme congestion. Furthermore, we can safely restrict attention to deterministic queuing — the kind we have been discussing — because stochastic queuing, which arises due to random fluctuations in the traffic stream, accounts for only a small fraction of travel delays (Newell, 1971, p. 125).

Let $V_a(t)$ be the volume of traffic arriving at time t at a point bottleneck of capacity V_K, or at the queue behind it if there is one. Let $V_b(t)$ be the volume passing through the bottleneck. Flows V_a and V_b are often called "arrivals" and "departures" (at and from the queue), respectively, in the queuing literature; but vice versa in the bottleneck literature, where V_a often represents departure from home and V_b arrival at work (both on the assumption that travel times upstream and downstream of the bottleneck are zero). To avoid confusion, we call V_a *queue entries* and V_b *queue exits* — even when the queue is of zero length.

Let $N(t)$ be the number of vehicles stored in the queue. It is common to ignore the physical length of highway required to store them or, equivalently, to consider the queue to be "vertical" rather than horizontal. Suppose we can ignore the reduction of speed from congestion so long as inflow is less than capacity V_K. Then the following kinked performance function relates queue-exits to queue-entries:

$$V_b(t) = \begin{cases} V_a(t) & \text{if } \mathrm{V_a}(t) \le V_K \text{ and } N(t) = 0 \\ V_K & \text{otherwise}. \end{cases} \tag{3.12}$$

The number of vehicles N in the queue, if any, changes depending on the difference between inflow and outflow:

$$\dot{N}(t) = V_a(t) - V_b(t), \tag{3.13}$$

where the dot denotes a time derivative (possibly one-sided).

Consider the typical case where the entry rate starts low, builds, then decreases, always remaining finite. Let t_q be the time when $V_a(t)$ first equals capacity V_K. Then $N(t)=0$ for $t \le t_q$ and has a right derivative at t_q given by $V_a(t_q)-V_K$; the queue builds, then shrinks, changing at rate $V_a(t)-V_K$ until it finally disperses at some time $t_{q'}$ defined by:

$$N(t_{q'}) \equiv \int_{t_q}^{t_{q'}} (V_a(t) - V_K)\mathrm{d}t = 0 .$$

With a first-in, first-out queuing discipline, each vehicle entering the back of the queue at time t must wait for $N(t)$ vehicles to pass through the bottleneck before it can pass through. This causes a *queuing delay*, for a driver entering the queue at t, equal to:

$$T_D(t) = \frac{N(t)}{V_K} = \int_{t_q}^{t} \left(\frac{V_a(t')}{V_K} - 1 \right) \mathrm{d}t', \quad t_q \le t \le t_{q'} . \tag{3.14}$$

We shall also refer to T_D as *travel-time delay* or simply *travel delay*.

An example, taken from Newell (1987), is shown in Figure 3.6. The two curves show cumulative queue-entries and queue-exits as functions of time, so that their slopes represent entry and exit flows V_a and V_b. When V_a exceeds V_K, a queue develops, and $N(t)$ can be found as the vertical distance between cumulative entries and exits. The queuing delay $T_D(t)$, for the driver entering the queue at t, is given by the horizontal difference between cumulative entries and exits.

FIGURE 3.6

Consider the special case of a fixed peak period for input flows , with incoming traffic constant at V_a during the time interval $[t_p , t_{p'}]$ and zero outside it. If a queue forms, it begins at time t_q=t_p. These equations then yield the following queuing delay, for $t \in [t_p , t_{p'}]$:

$$T_D(t) = \begin{cases} 0 & \text{if } V_a \le V_K \\ \left[(V_a / V_K) - 1 \right] \cdot (t - t_p) & \text{if } V_a > V_K . \end{cases} \tag{3.15}$$

The average travel delay is:

$$\bar{T}_D(t) = \frac{1}{(t_{p'} - t_p)} \int_{t_p}^{t_{p'}} T_D(t)dt$$

$$= \begin{cases} 0 & \textit{if } V_a \le V_K \\ \frac{1}{2}(t - t_p) \cdot \left[(V_a / V_K) - 1 \right] & \textit{if } V_a > V_K . \end{cases}$$

Adding a free-flow travel time T_f per unit distance for the journey yields equation (3.10) for $P=(t_{p'}-t_p)$. Note that the period of outflows lasts longer than the period of inflows P, and extends from t_p to $t_p + P\cdot(V_a/V_K)$.

We shall refer to this congestion technology, where travel delays result exclusively from vertical queuing at a bottleneck, as "pure" bottleneck congestion. Its consequences for equilibrium queues and optimal pricing are considered in Section 3.4 and Chapter 4.

Analysis of Shock Waves

While much economic modeling congestion has considered pure bottleneck congestion as just described, some has begun to take advantage of more sophisticated dynamic models. We provide here a brief survey based in part on Lindsey and Verhoef (2000).

Probably the most famous is the *hydrodynamic* or *kinematic model*, developed by Lighthill and Whitham (1955) and Richards (1956) and therefore known as the LWR model; see Daganzo (1997) for a review. It is a *continuum* model in that traffic characteristics V, D, and S are assumed to be continuous functions of location x and time t, in a manner similar to physical models of fluids (i.e., liquids and gases). Three essential assumptions are the following. First, a relationship $S(D)$ holds between speed and density, as shown in Figure 3.1a, even under non-stationary conditions. Second, identity (3.7) holds everywhere, thus defining functions $V(D)$ and $V(S)$. And third, vehicles are neither created nor destroyed along the road, resulting in the following *conservation* or *continuity equation*:[26]

$$\frac{\partial V(t,x)}{\partial x} + \frac{\partial D(t,x)}{\partial t} = 0 \ . \tag{3.16}$$

[26] Equation (3.16) can be understood as follows. Let ΔN denote the change in the number of cars between two locations x and $x+\Delta x$ over the time interval from t to $t+\Delta t$. If Δt is small enough, the flow rates at the two locations can be treated as constant over time; letting ΔV denote the difference in V between the two locations, we see that the number of vehicles between x and $x+\Delta x$ builds up if incoming flow exceeds outgoing flow, i.e. it builds at rate $-\Delta V$. Thus $\Delta N = -\Delta t\cdot\Delta V$. If Δx is small enough, density at the two locations can be treated as equal; letting ΔD denote the change in density over time, we see that the number of vehicles in this space of length Δx changes in proportion to the change in density: $\Delta N = \Delta x\cdot\Delta D$. Because vehicles are conserved, these two expressions for ΔN should be equal, in turn implying $\Delta V/\Delta x + \Delta D/\Delta t = 0$. When the discrete increments become infinitesimal, (3.16) is obtained.

The resulting dynamics can be described in terms of *shock waves*, which are disturbances to traffic caused by traffic lights, accidents, lane reductions, or other discrete variations. Specifically, a shock wave is the moving boundary between two stationary states, i.e., it is the moving point at which vehicles leave one state and enter another. For example, consider an upstream steady state with $V_u=S_u \cdot D_u$, and a downstream steady state with $V_d=S_d \cdot D_d$. The location dividing these two states will propagate along the highway at some speed S_w, the value of which is to be determined. Upstream traffic catches up with the boundary at relative speed S_u-S_w, and hence enters the shock wave at rate $D_u \cdot (S_u-S_w)$. Similarly, downstream traffic leaves the boundary at relative speed S_d-S_w, and hence exits the wave at rate $D_d \cdot (S_d-S_w)$. Conservation of vehicles requires these rates to be equal, implying:

$$S_w = \frac{V_u - V_d}{D_u - D_d} \,. \tag{3.17}$$

It can be shown that a forward-moving shock wave can never travel faster than the traffic that carries it, provided the function $V(D)$ is concave and crosses the origin.[27]

Finding a solution for the LWR model is tedious when traffic inflow varies continuously over time (Newell, 1988). Therefore a number of simplified models have been formulated. Indeed, the model of pure bottleneck congestion can be regarded as one of these, in which the shock wave, defined as the back of the queue, travels at speed $S_w=0$ (since the queue is assumed to have no spatial extent). Equivalently, according to (3.17), the queue density D_d is infinite.

Agnew (1977) and Mahmassani and Herman (1984) propose a simplification involving *instantaneous propagation*, by which changes in a road's inflow rate V_i immediately affect its outflow rate V_o. With N denoting the number of vehicles on a finite stretch of road, Agnew's model is summarized by the differential equation of state:

$$\dot{N} = V_i - V_o(N),$$

[27] After rotating the $V(D)$ curve of Figure 3.1c by 90°, S_w between two states can be found geometrically as the slope of the straight line connecting these states. The traffic speed in a state is given by the slope of the ray from the origin through that state. Under the stated assumptions, S_w is then always smaller than S_u and S_d.

where the function $V_o(N)$ has the same general shape as $V(D)$ in Figure 3.1c. This model is usually interpreted as implying that density is uniform along the entire road at every instant, so that shock waves propagate at an infinite speed. This has the unrealistic implication that drivers adjust speeds in response to changes in upstream traffic conditions.

A different simplification, by Henderson (1974) and Chu (1995), adopts the opposite assumption of *no propagation*: a vehicle's speed – assumed constant during the trip – is determined as a function only of the flow at one point in space and time. That point is either the entry of the vehicle onto the road (Henderson) or the exit of the vehicle from the road (Chu). This formulation therefore does not consider possible interactions between adjacent vehicles whose departure times are different, despite the fact that the distance between them may be changing during the trip because they travel at different speeds. Hence, there is no propagation of shock waves.

Mun (2002) uses the LWR model to examine the total travel time for traffic entering an otherwise uniform road interrupted by a bottleneck (i.e. a region with lower capacity). The regions on either side of the bottleneck have different capacities and therefore different speed-density functions $S(D)$. When traffic exceeds the bottleneck capacity, a queue builds up and exits the bottleneck at a rate equal to bottleneck capacity. Given an exogenous inflow $V_a(t)$ upstream of the bottleneck, volumes and densities can be found both upstream of the queue and within it; (3.17) determines the rate at which the back of the queue moves, allowing one to compute the queue length at every instant of time. Total trip time is the sum of time traversing the distance to the back of the queue plus the time spent in the queue itself. Mun shows that the model reduces to the Henderson model when inflow never exceeds capacity, and to the deterministic queuing model of (3.14) — augmented by a constant free-flow travel time over the entire length of the road — when the upstream section has sufficient capacity that its speed-flow curve can be approximated by a constant speed.

Car-Following Models

Another approach is to allow for continuous-time and continuous-space traffic dynamics, like LWR, but treat traffic itself as consisting of discrete vehicles whose behavior is specified. A car-following equation stipulates how the motion of vehicle n+1 (the 'follower') depends on the motion of vehicle n (the "leader"). Usually the dependent variable is acceleration, and it depends

on the distance between the leader and follower and on their speed difference, as in the General Motors model described by May (1990):

$$\ddot{x}_{n+1}(t+\delta) = \frac{a \cdot \left[\dot{x}_{n+1}(t+\delta)\right]^m}{\left[x_n(t) - x_{n+1}(t)\right]^l} \cdot \left[\dot{x}_n(t) - \dot{x}_{n+1}(t)\right] \ , \tag{3.18}$$

where x denotes location, $\dot{x}$ is speed, $\ddot{x}$ is acceleration, δ is a reaction time, and a, l and m are non-negative parameters.

Under stationary traffic conditions, car-following models imply a relationship between speed and density (the inverse of vehicle spacing) that is consistent with the relationships we have considered above. For example, May shows that if m=0 and l=1, integrating (3.18) over t yields:

$$\dot{x}_{n+1} = a \cdot \log(x_n - x_{n+1}) + C_0 \, ,$$

with C_0 denoting a constant of integration. Equivalently,

$$S = a \cdot \log(C_1/\mathrm{D}) \, , \tag{3.19}$$

where $D_j = a \cdot \log[C_0]$ is the jam density, as can be seen by setting S=0 in (3.19). This equation reproduces a speed-density relationship that was proposed by Greenberg (1959); it suffers from the disadvantage that free-flow speed is infinite. Other parameter values for (3.18) have been found to correspond to other macroscopic models (Hall, 1990).

Verhoef (2001, 2003) proposes a simpler car-following model which is even more obviously a dynamic extension of a steady-state model. Verhoef takes the function $S(D)$ to represent the behavior of the follower, where again D is the inverse of vehicle spacing:

$$\dot{x}_{n+1}(t) = S(D) \, ; \qquad \mathrm{D} = \left[x_n(t) - x_{n+1}(t)\right]^{-1} \ . \tag{3.20}$$

With $S(D)$ a non-decreasing function, this model reproduces the basic behavioral assumption in the more complex model of (3.18), namely that driver n+1 accelerates when driving slower than driver n. Verhoef finds the model to be quite tractable and useful for examining stability of steady-state equilibria under conditions of hypercongestion.

Surprisingly, economists have rarely considered the behavioral motivations underlying the relationship between density and speed that underlies most congestion models. If high density causes traffic to slow down, this must somehow reflect decisions by individual drivers,

presumably decisions trading off speed against safety. Rotemberg (1985) proposed precisely such a tradeoff within a steady-state car following framework. Verhoef and Rouwendal (2004) show that such a model can produce a locus of equilibrium outcomes that forms a backward-bending speed-flow relation just like the one in Figure 3.1b. But with this economic trade-off identified, it is now possible to examine the desirability of regulation that might change individual driver behavior. In fact, both of these papers find that for any given flow level, economic efficiency (taking into account the drivers' own evaluations of accident costs) could be improved by inducing drivers to go faster than the equilibrium level with individual choice. This is because an individual driver, considering his or her own optimal speed and vehicle spacing, ignores the effect this has on the accident risks (directly) and travel times (indirectly, after adjustments) experienced by other drivers. Interestingly enough, these two considerations both work in the same direction: for a given flow, an increase in speed implies a decrease in density, as shown by (3.7), which in turn reduces accident risks for other drivers.

3.3.4 Congestion Modeling: A Conclusion

Our review, even though selective, reveals a varied menu of approaches to modeling congestion.[28] Most economic analysis has used just two of these: the static speed-flow curve and the dynamic deterministic bottleneck model. Furthermore, researchers have barely begun to describe the behavior that underlies congestion technology or to identify externalities in that behavior.

Researchers face a difficult tradeoff between tractability and realism. The economic literature has mostly sided with tractability, producing many valuable insights but with limited applicability. There could be a considerable payoff from incorporating more realistic engineering models into economic analysis. Examples will be encountered in Chapter 4 as we discuss dynamic congestion tolls and hypercongestion.

---

[28] Other interesting phenomena include path-dependency of observed speed-density combinations (Zhang, 1999), multiple-class traffic (Hoogendoorn and Bovy, 1998), and what appear to be spontaneous phase transitions, analogous to transitions between liquids and gases, observed in simulation results by Kerner and Rehborn (1997) and others. The relevance of the latter to observed traffic is disputed by Daganzo, Cassidy, and Bertini (1999).

3.4 Highway Travel: Short-Run Cost Functions and Equilibrium

The congestion models we have described can be used to formulate cost functions for highway travel. In this section we deal with short-run models, which we define as describing a highway with fixed capacity (certain other types of capital, such as vehicles and parking facilities, may be variable). Section 3.5 considers long-run cost functions.

In order to simplify the exposition, we assume that everyone has an identical value of time, denoted in this chapter by α. The average cost c on a defined length of road then consists of monetary expenses c_{00} (like fuel consumption and maintenance), the cost of free-flow travel time $\alpha \cdot T_f$, the cost of travel delays $\alpha \cdot (T - T_f)$, and, in some models, the cost of undesirable schedules c_S (to be defined). The first two cost components make up the travel cost in absence of congestion, c_0; the latter two give the congestion-related cost c_g. Thus:

$$c = c_0 + c_g = \left[c_{00} + \alpha \cdot T_f\right] + \left[\alpha \cdot \left(T - T_f\right) + c_S\right]. \tag{3.21}$$

We ignore for simplicity any dependence of money costs on congestion.

We assume that average cost c is borne entirely by the user; it is sometimes called the *generalized cost* because it indicates the monetary value of the resources supplied by an individual taking a trip. The related concept of *generalized price*, denoted by p, adds to c any applicable tolls and taxes.

3.4.1 Stationary-State Congestion on a Homogeneous Road

We begin with stationary-state congestion on a single homogeneous road with identical users. Simple as this set-up may seem, the resulting model has proven capable of creating great confusion. We therefore provide a detailed discussion.

We define a stationary state as a situation where traffic flow V is constant over time and space and is equal to the rates at which trips are started and ended. Thus the situation could in principle last indefinitely long. In reality, of course, traffic congestion undergoes rapid changes; stationary-state models abstract from such changes, and their practical usefulness is limited for this reason. Their advantage is that they are basically static and therefore relatively simple.

The key simplification is to recognize that with V constant and equal to the inflow and outflow rates, it can represent both quantity demanded (by users) and quantity supplied (according to the congestion technology) at a given average cost c. Following Walters (1961), we might picture the situation as resulting from the interaction of a demand curve $V=V_D(c)$ or its inverse, $d(V)$, and a supply curve $c(V)$. When congestion is described by a speed-flow function $S(V)$ on a road of length L, and there are no scheduling costs, this supply curve takes the form

$$c(V) = c_{00} + \alpha \cdot T(V) = c_{00} + \alpha \cdot L / S(V) \ . \tag{3.22}$$

So long as we stay on the normally congested portion of $S(V)$, this supply curve is rising and leads to conventional equilibrium results. We will need to keep in mind, however, that although the user is assumed to perceive c as both average and marginal private cost, marginal social cost will be different unless $c(V)$ is constant. This is because $c(V)$ incorporates a *technological externality*: a direct technological dependence of one person's average travel cost on the travel decisions of others. We describe the consequences of this in the pricing analysis of the next chapter.

When we examine the hypercongested portion of $S(V)$, we run into trouble with Walters' interpretation. For one thing, the existence of hypercongestion implies that the average cost depicted by (3.22) is not single-valued — in fact, it does not fit the formal definition of a cost function, which is the *minimum* cost of producing a given output. Furthermore, when we confront it with the inverse demand function $d(V)$, as in Figure 3.7, we can get as many as three different candidate equilibria, whose properties have engendered considerable controversy (Verhoef, 1999; Small and Chu, 2003). The normally congested equilibrium, denoted x in the figure, resembles a standard economic market equilibrium with a downward sloping demand and an upward sloping supply. But for the two hypercongested equilibria, y and z, the "supply curve" slopes downward. Intuition warns that there is something peculiar here. How should one interpret a situation where an increase in traffic inflow produces faster travel and thus a lower average cost?

FIGURE 3.7

Conventional stability analysis of the candidate equilibria is inconclusive: x is stable for both price and flow perturbations, y for flow perturbations only, and z for price perturbations only.[29] Thus whichever type of perturbation is taken as the criterion for stability, the model produces two candidate equilibria in the case of the demand curve shown. If we insist that an equilibrium should be stable against both types of perturbations, we would reject both hypercongested candidate equilibria; but then we must acknowledge that for a higher demand curve like $d_1(V)$, there is no stable equilibrium.

One difficulty with conventional stability analysis is that the perturbations considered involve simultaneous changes in the flow rates *into* and *along* the road, which is physically impossible. It therefore seems more appropriate to consider perturbations of the inflow rate, treating flow levels along the road as endogenous. Doing so introduces the concept of dynamic stability: can a given stationary state arise as the end state following some transitional phase initiated by a change in the inflow rate?

Verhoef (2001) examines dynamic stability using the car-following model (3.20), allowing for vertical queuing before the entrance when inflows cannot be physically accommodated on the road. He finds that the entire hypercongested branch of the $c(V)$ curve in Figure 3.7 is dynamically unstable.[30] The locus of dynamically stable stationary states turns out to be the curve shown as $c_{stat}(V)$ in Figure 3.8; it follows the normally congested part of $c(V)$ and rises vertically once volume reaches capacity, just as with deterministic queuing. This generates a new stationary state, x', which is dynamically stable. This state involves a maximum flow on

---

[29] See Verhoef (1999). When applying convention stability analysis, an equilibrium is stable for flow perturbations if a small increase in flow leads to average cost c(V) above inverse demand $d(V)$, inducing users to reduce their inflow. An equilibrium is stable for price perturbations if a small increase in "price" (average cost) leads to excess supply, i.e. it leads to a "price" where the supply curve is to the right of the demand curve. In conventional markets this would cause suppliers to reduce the price level; but here the "supplier" is a congestion technology rather than a profit-motivated firm, making this stability criterion a questionable one.

[30] The same result occurs with the LWR model when discontinuous changes in traffic conditions (V, D and S) are ruled out. The intuition is that, from equation (3.16), the shock wave between two hypercongested stationary states always travels at a negative speed. This is because, with $V(D)$ downward-sloping between both states, V_u-V_d and D_u-D_d must have opposite signs. Therefore a change in inflow can never cause a transition between two hypercongested stationary states, or, indeed, from the maximum-flow state to any hypercongested state: the boundary to the new state can travel only backward so can never enter the road.

the road, a constant-length queue before its entrance with cost c_q', and rates of queue-entries and queue-exits both equal to the capacity of the road. It does not involve hypercongestion on the road itself; rather, hypercongestion exists only within the entrance queue, consistent with the terminology of Figure 3.4. Furthermore, the flow rate and speed inside the queue are irrelevant to total trip time, making the economic properties of the model independent of the shape of the hypercongested portion of the speed-flow curve.

FIGURE 3.8

Thus the true supply curve for stationary-state traffic, $c_{stat}(V)$, is everywhere rising and intersects any downward-sloping demand curve exactly once. It has two distinct regimes, one of them vertical; but it is smooth and may sometimes be approximated by a power function based on (3.9) (which we originally derived as a time-averaged relationship):

$$c(V) = c_{00} + \alpha T_f \cdot \left[1 + a \cdot \left(V / V_K\right)^b\right]. \tag{3.23}$$

Our views on the dynamic instability of hypercongestion are not undisputed. McDonald, d'Ouville, and Liu (1999) provide empirical results that appear to involve hypercongestion for sustained periods of time in absence of a downstream bottleneck. Furthermore, alternative solutions to the questions raised by the conventional diagram of Figure 3.7 have been proposed. Else (1981), Hills (1993), and Ohta (2001) try to solve the problem by using traffic density, number of travelers on the road, or the total number of trips (not expressed per unit of time) as the relevant argument in static inverse demand and average cost functions. In our view, these non flow-based quantities do not give a meaningful economic measure of aggregate stationary-state output. The total number of trips is not even defined for stationary state traffic until a time period for measurement is specified – in which case the measure becomes flow-based after all. Furthermore, traffic density is an aggregate measure of the proportion of road space occupied at a given point in time, not of the number of trips taken over an interval of time; a demand function defined over density would therefore assume that the good demanded is not the completion of trips but rather the occupation of road space. Tell that to your average harried commuter! We conclude that traffic flow is the appropriate output measure for stationary-state

analyses, while the total number of trips is appropriate for time-averaged or dynamic models that specify an applicable time period.

3.4.2 Time-Averaged Models

Cost models using the time-averaged congestion functions described earlier avoid thee problems. They can accommodate temporary inflows greater than capacity, yet are single-valued and look much like the cost function of Figure 3.8. For fixed time period P, the time-averaged inflow volume V has a simple interpretation as quantity demanded: namely, it is the number of trips divided by P.

Figure 3.9 compares the cost functions derived from two different time-averaged speed-flow relationships, (3.10) and (3.11). The stationary-state average cost function $c_{stat}(V)$ of Figure 3.8 is shown for comparison. Both of the time-averaged functions become steeper for higher P; specifically, as V increases both functions approach a straight line that itself approaches the vertical axis as $P\rightarrow\infty$. This increasing similarity between time-averaged cost functions and the stationary-state function, as the time period becomes indefinitely large, makes intuitive sense. Yet the correspondence is imperfect in both cases: at flows below capacity, the piecewise linear function allows for no congestion, while the Akçelik function C_{AK} may allow for too much since, for high enough values of P, it will cross the C_{stat} curve and exhibit arbitrarily high travel times even when $V<V_K$.

FIGURE 3.9

Furthermore, the time-averaged static models have some inherent weaknesses. First, it is not clear how to measure P from observed traffic patterns, which fail to adhere to the assumption of a constant flow occurring only over a well-defined period .[31] Second, P is set exogenously, but in reality will vary with traffic conditions and policies. Third, the assumed exogenous inflow

[31] A naïve but understandable choice of P, defined by the instants that queuing begins and ends, would imply a time-averaged inflow V equal to capacity V_K. This choice would produce travel delay equal to zero according to the piecewise-linear function and $(1/2)^{½}\cdot(PJ_a/V_K)^{½}$ according to Akçelik's function.

rate is unlikely to be consistent with any rational demand behavior. All three problems are solved by formulating dynamic models that endogenize departure times, to which we turn in the next subsection.

3.4.3 Dynamic Models with Endogenous Scheduling

The dynamic congestion technologies discussed in Section 3.3 allow construction of dynamic equilibrium models, in which departure times (and therefore peak duration) are endogenous and travel delays vary continuously over time. A common assumption in such models is that travelers choose an optimal schedule for their trip by trading off travel time cost against schedule-delay cost, as in the demand model of Section 2.3.2. Average cost, as defined in (3.21), then includes a part c_S due to schedule delay. It could also include a part due to unreliability, but current theoretical models have not incorporated that separately.

We treat here the case where scheduling costs arise from deviations between an individual's actual and desired arrival time at work, following the notation of equation (@2.47) with θ=0. (This case applies best to the morning peak period; modeling the afternoon peak would presumably require assuming a desired departure time from work — a case that has received far less analysis.) Recall that the per-minute costs of early and late arrival are β and γ, respectively. Then the travel cost for an individual departing from home at time t is:

$$c(t) = c_{00} + \alpha \cdot T(t) + c_S(t); \quad c_S(t) = \begin{cases} \beta \cdot \left(t_d - t - T(t)\right) & \text{if } t + T(t) \le t_d \\ \gamma \cdot \left(t + T(t) - t_d\right) & \text{if } t + T(t) > t_d, \end{cases} \tag{3.24}$$

where t_d is the desired arrival time at work and $T(t)$ the travel time incurred when departing at t. Of course, $T(t)$ and therefore $c(t)$ depend also on capacity and perhaps on past, current, or even future traffic levels (the latter possibility arising with instantaneous propagation).

We take the simplest dynamic congestion technology discussed in Section 3.3.3, namely pure bottleneck congestion, where congestion occurs solely through vertical queuing behind a bottleneck. It is convenient to assume one traveler per vehicle and to define *average congestion cost* by subtracting the constant $c_{00}+\alpha T_f$, i.e.,

$$c_g(t) \equiv c(t) - c_{00} - \alpha \cdot T_f .$$

For convenience, we set free-flow travel time T_f to zero. Then $T(t)$ is equal to the travel delay $T_D(t)$ defined in (3.15), except that the time t_p when congestion begins is now endogenous and denoted by t_q.

Because desired schedules are defined in terms of arrival time at work, it is convenient to focus on the time t' when a traveler exits the queue. Given that T_f=0, this is also that traveler's arrival time at the work. If there were no congestion, the rate at which travelers depart from the queue would be simply the distribution of desired work-arrival times, which we denote by $V_d(\cdot)$. We can now work backward to find queue-entry rate $V_a(t)$ (i.e., the rate of arrivals at the vertical queue, or equivalently the rate of departures from home) that is consistent with travelers' independent scheduling decisions.

If $\text{Max}_t V_d(t) \leq V_K$, there is no queuing or schedule delay, and the entry and exit rates are both equal to the desired rate, $V_d(t)$. If capacity is insufficient, however, people must trade off queuing delay against schedule delay in choosing their queue-entry times, which will imply a certain aggregate queue-entry time pattern $V_a(t)$. The resulting equilibrium, analyzed by Hendrickson and Kocur (1981) and Newell (1987), can be quite complex. However the following special case, first analyzed by Vickrey (1969, 1973) and further elaborated by Fargier (1983), is tractable and leads to surprisingly elegant and insightful results.

Suppose, then, that $V_d(t)$ is constant at V_d during the interval $[t_p , t_{p'}]$ and zero outside that interval. Hence there are a total of $Q \equiv V_d \cdot q$ travelers when demand is inelastic, where $q = t_{p'} - t_p$ denotes how long the peak period would last if capacity were unrestricted. Assume $\beta < \alpha$, which is supported by the empirical evidence of Section 2.3.2 and which is necessary to achieve an equilibrium without massed departures at a single instant in time. Consider the case $V_d > V_K$, so that the desired exit rate cannot be achieved and thus queuing and/or schedule delay must occur. Our analysis follows the logic and much of the notation of Arnott, De Palma, and Lindsey (ADL, 1990b).[32]

---

[32] The ADL analysis treats only the special case in which the desired schedule t_d is identical for all commuters, i.e. $Q \equiv V_d \cdot q$ is fixed but q=0 and V_d is infinite. This version, which we call the "basic bottleneck model," is widely used. We achieve more realism at a modest cost in complexity by retaining Vickrey's original assumption of a uniform distribution of t_d with nonzero q and finite V_d. In doing so we also render the assumption of zero travel time before

For a commuter exiting the queue before the desired time t_d, equilibrium requires that the chosen queue-entry time minimizes the combined costs for early exits in (3.24): $\alpha \cdot T_D(t) + \beta \cdot [t_d - t - T_D(t)]$. This requires that $T_D(t)$ change at rate $\beta/(\alpha - \beta)$ so long as anyone entering the queue at time t is exiting early. Similarly, so long as anyone entering at t is exiting late, $\alpha \cdot T_D(t) + \gamma \cdot [t + T_D(t) - t_d]$ must be minimized, so $T_D(t)$ must change at rate $-\gamma/(\alpha+\gamma)$.[33] The first and last commuters exiting must face a zero queue length in equilibrium, because otherwise a discretely lower travel cost could be realized by departing just before t_q or after $t_{q'}$.

Comparing these equilibrium rates of change in T_D to that implied by equation (3.15), namely $[(V/V_K)-1]$, we see that vehicles must be entering the queue at rates

$$V_a^{early} = V_K \cdot \frac{\alpha}{\alpha - \beta}; \qquad V_a^{late} = V_K \cdot \frac{\alpha}{\alpha + \gamma} \tag{3.25}$$

during the early and late parts of the peak period, respectively. The resulting pattern is shown in Figure 3.10, in which $N(t)$ is the number of vehicles in the queue, $\tilde{t}$ is the entry time for the commuter incurring maximum queuing delay T_{Dm}, and this commuter's exit time is:

$$t^* \equiv \tilde{t} + T_D(\tilde{t}) \equiv \tilde{t} + T_{Dm}. \tag{3.26}$$

Commuters with $t_d < t^*$ enter the queue before (or possibly at) $\tilde{t}$, and those with $t_d > t^*$ enter after (or possibly at) $\tilde{t}$. (Due to linearity in the cost function, each commuter is in fact indifferent among departure times $[t_q, \tilde{t}]$ or else among departure times $[\tilde{t}, t_{q'}]$; however we can remove this indeterminacy by making the quite natural assumption that commuters exit the queue in the same order as their desired queue-exit times.)

FIGURE 3.10

---

and after the bottleneck relatively innocuous because the varying preference for queue-exit time could be interpreted as arising from individuals having different free-flow times T_f that are outside the model.

[33] These theoretical rates of change are compared to actual rates on congested roads in Paris by Fargier (1983, pp. 247-252) in order to estimate the behavioral parameters β/α and γ/α. He gets values much smaller than the direct behavioral estimates of Small (1982), possibly reflecting the limited realism of the model or a relatively wide dispersion in desired arrival times.

It remains to determine $\tilde{t}$. We accomplish this by equating the total numbers of travelers entering and exiting. The exit rate is constant at V_K during some interval $[t_q , t_{q'}]$ which, in order to accommodate the total of Q vehicles, must be of duration:

$$t_{q'} - t_q = Q / V_K = q \cdot V_d / V_K > q . \tag{3.27}$$

This peak travel period encompasses but exceeds the desired peak $[t_p , t_{p'}]$, whose duration is q; congestion begins prior to the earliest desired queue-exit and lasts beyond the latest desired queue-exit. Furthermore, the duration of this interval depends inversely on V_K, showing that expanding capacity narrows the peak period — as postulated for example by Downs (1962).

In order to solve for the entire equilibrium configuration, define

$$\sigma = \frac{t^* - t_p}{q} = \frac{t^* - t_p}{t_{p'} - t_p} \tag{3.28}$$

as the proportion of commuters who exit the queue before t^* (equivalently, the proportion who enter the queue before $\tilde{t}$). They enter at rate V_a^{early} , so their number must be:

$$\sigma \cdot Q = V_a^{early} \cdot \left(\tilde{t} - t_q\right) . \tag{3.29}$$

Similarly, the proportion $(1–\sigma)$ who enter after $\tilde{t}$ do so at rate V_a^{late} , so

$$\left(1 - \sigma\right) \cdot Q = V_a^{late} \cdot \left(t_{q'} - \tilde{t}\right) . \tag{3.30}$$

Equations (3.25) through (3.30) can be solved for:

$$\sigma = \gamma /(\beta + \gamma) \tag{3.31}$$

$$t_q = t_p - \sigma \cdot q \cdot \left[(V_d / V_K) - 1\right] \tag{3.32}$$

$$t_{q'} = t_{p'} + (1 - \sigma) \cdot q \cdot \left[(V_d / V_K) - 1\right] \tag{3.33}$$

$$t^* = t_p + \sigma \cdot q$$

$$\tilde{t} = t_p + \sigma \cdot q - T_{Dm}$$

$$T_{Dm} = \delta \cdot Q /(\alpha \cdot V_K) \tag{3.34}$$

where

$$\delta \equiv \beta\gamma /(\beta + \gamma) = \beta\sigma . \tag{3.35}$$

The maximum delay (3.34) corresponds to a maximum travel-delay cost $\delta \cdot Q/V_K$.

Figure 3.11 shows how costs vary over time. Travel-delay cost per traveler, $c_T(t)$, rises linearly from zero to T_{Dm} (reached at time $\tilde{t}$) and falls linearly back to zero. Schedule-delay cost per traveler, $c_S(t)$, falls linearly from a maximum of $\beta \cdot (t_p - t_q)$ for the earliest traveler to zero (at $\tilde{t}$), then rises linearly to a maximum of $\gamma \cdot (t_{q'} - t_{p'})$; computing these maxima from (3.31) - (3.33), we find they are both equal to $(\delta \cdot Q/V_K) \cdot (1 - V_K/V_d)$. Note that their sum $c_g(t)$ need not be constant in equilibrium because each consumer has a different desired schedule t_d. (This differs from the ADL model.)

FIGURE 3.11

From the piecewise-linear cost patterns just described, we see easily that the time-averaged cost components, c_T and c_S, are just half their maximum values. Thus time-averaged travel-delay cost per traveler is:

$$\bar{c}_T = \tfrac{1}{2} \cdot \frac{\delta \cdot Q}{V_K} \equiv \frac{\delta \cdot q}{2} \cdot \frac{V_d}{V_K} \ . \tag{3.36}$$

The middle expression is the same formula as that derived by Fargier (1983, p. 246) and ADL (1990b, p. 116) for the special case q=0. Surprisingly, it depends only on the total number of travelers Q, not on the distribution of their desired queue-exit times. Similarly, the time-averaged schedule-delay cost per traveler is:

$$\bar{c}_S = \tfrac{1}{2} \cdot \frac{\delta \cdot Q}{V_K} \cdot \left(1 - \frac{V_K}{V_d}\right) \equiv \frac{\delta \cdot q}{2} \cdot \left(\frac{V_d}{V_K} - 1\right). \tag{3.37}$$

This does depend on the distribution of desired exit times; for a given number of travelers $Q \equiv V_d \cdot q$, distributing the desired exit times over a shorter interval q raises V_d and thereby raises average schedule delay cost. In the extreme case when q=0 while V_d=∞ (with Q finite), $\bar{c}_S$ becomes equal to $\bar{c}_T$ as derived by ADL (1990b).

Adding $\bar{c}_T$ and $\bar{c}_S$ and including the possibility of $V_d \le V_K$ with its lack of queuing, we can write the time-averaged congestion cost as:

$$\bar{c}_g(V_d, q; V_K) = \begin{cases} 0 & \text{if } V_d \le V_K \\ \dfrac{\delta \cdot Q}{V_K} \cdot \left(1 - \dfrac{V_K}{2V_d}\right) \equiv \delta \cdot \mathrm{q} \cdot \left(\dfrac{\mathrm{V_d}}{\mathrm{V_K}} - \dfrac{1}{2}\right) & \text{otherwise.} \end{cases} \tag{3.38}$$

Equation (3.38) is the average congestion cost given the constraint that people are free to adjust their schedules according to their tradeoff between queuing delay and schedule delay. It should be viewed as part of a second-best aggregate cost function, in which the entry pattern $V_d(t)$ is determined sub-optimally. This is why it is discontinuous at $V_d=V_K$: as soon as there is any congestion, the average queuing delay (3.36) jumps from zero to $\frac{1}{2}\cdot\delta\cdot q$. As we shall see in Chapter 4, a different queue-entry pattern would eliminate queuing delay and thereby reduce $\bar{c}_g$ to $\bar{c}_S$, making it the congestion cost for a first-best cost function and also making it continuous in V_d.

These costs have the remarkable feature of not depending on value of travel time, α. So long as α remains greater than β so that the analysis applies, increasing the value of time causes no change in the duration or timing of the peak interval $[t_q\,,\, t_{q'}]$, nor in the proportion of travelers who exit early; instead, it causes the queuing delay to decrease just enough to hold queuing cost constant, while schedule delay remains unchanged. This point was first noted, for a closely related model, by De Palma and Arnott (1986).

Equally remarkable is that the entire pattern of queue entries and queue exits shown in Figure 3.10 is unaffected by how demand Q is factored into q and V_d, so long as t^* is unchanged and V_d is greater than capacity. This again results from the perverse private incentives that cause a substantial queue to form even if V_d exceeds capacity by only a tiny amount. Spreading Q over a wider interval does, however, reduce scheduling costs because the pattern in Figure 3.10 imposes fewer costs when some people already prefer to arrive at some time other than the most popular one.

In the special case where all users are identical and have the same desired arrival time t^*, there is always congestion for any nonzero Q and (3.38) simplifies to a linear average cost function (since $V_d=\infty$):

$$\bar{c}_g(Q; V_K) = \frac{\delta \cdot Q}{V_K}. \tag{3.39}$$

The lack of heterogeneity across commuters now causes equilibrium travel cost to be constant over time. This model has the advantage that demand is summarized by a single quantity, Q, rather than two quantities (q and V_d) as in (3.38). It is therefore very easy to interact it with an inverse demand function to derive the equilibrium. Indeed, this is one of the advantages noted by ADL (1993): once the model is solved in this form, as a function of Q, average cost looks exactly like that of the stationary-state model (3.22) with travel-delay cost proportional to Q. We will, in what follows, use the term 'basic bottleneck model' to refer to this widely used version of the model with identical desired arrival times, a linear schedule delay cost function, and pure bottleneck congestion.

The derivation of dynamic equilibrium for other dynamic congestion technologies involves roughly the same steps as for the bottleneck model. Because most dynamic congestion technologies are nonlinear, the analytics become more cumbersome and typically no closed-form analytical solutions can be obtained. Other demand structures can also be assumed: for example, Ben-Akiva, Cyna, and De Palma (1984) analyze a model incorporating a probabilistic demand similar to that of Section 2.3.2.

Summary

We have considered three types of models to study short-run variable cost: stationary-state, time-averaged, and dynamic. Each leads to a tractable formula for short-run variable cost under certain assumptions. Stationary-state and time-averaged models are both characterized by a rising average cost function (in one case with a vertical asymptote), so conceptual analyses are often similar for both models; when this is the case in later chapters, we will treat the two models jointly and refer to them as "static models." Dynamic models show how scheduling flexibility reduces or eliminates the time variation in costs that underlie time-averaged models. Dynamic models permit internally consistent analyses of staggered and flexible work schedules as well as the "shifting-peak phenomenon" discussed in the next chapter. Furthermore, as we have seen, a conventional static model can be derived from a dynamic model as a reduced-form relationship among time averages or cumulative quantities.

3.4.4 Network Equilibrium

Up to this point we have ignored the fact that traffic usually operates on a network. Accounting for this requires us to recognize that the cost of a trip depends on flows on one or more links, each of which may be serving several trip types. Furthermore, users will seek out the best routes for their trips, and the resulting cost will depend on the allocation of traffic to links that results from this process. Typically we assume that the search for routes settles down rather quickly to an equilibrium characterized by each user choosing the route that minimizes cost for that particular trip. Such a situation is called a *user equilibrium* (UE) because it results from individual optimization by each user, as opposed to any collaborative procedure. [34]

To analyze such problems, we define a network structure consisting of M origin-destination pairs or "markets" (denoted m=1,…,M), R routes (denoted r=1,…,R), and L directed links (denoted l=1,…,L). "Directed" means that a two-way roadway is represented by two links carrying traffic in opposite directions. A single origin-destination (OD) pair may be served by multiple routes; each route may comprise multiple links; and any link may be part of more than one route. As a result, traffic serving different origin-destination pairs is likely to interact on certain links, and of course this affects how congestion forms. We define a set of dummy indicators δ_{rm} to denote whether route r serves market m (in which case δ_{rm}=1), and another set δ_{lr} to denote whether link l is part of route r.

The simplest case is when all users are identical, alternative routes are perfect substitutes, and congestion on a link depends only on the flow on that link (as opposed to, say, an intersection). An appropriate concept for the user equilibrium is then *Wardrop's first principle* (Wardrop, 1952): for a given OD pair, all used routes (those with positive flows) should have equal average cost, and there should be no unused routes with lower costs. So long as users take aggregate traffic conditions as given, this principle is consistent with the standard game-theoretic concept of Nash equilibrium: no user can reduce cost by unilaterally changing route. When demand for trips between an OD pair is elastic, an additional equilibrium condition is that the equalized average cost for used routes be equal to the marginal willingness to pay for trips between that origin and destination.

---

[34] Somewhat confusingly, the UE is sometimes called a "user optimum," while the socially optimal flow pattern (to be discussed in Chapter 4) is called a "system optimum."

These conditions can be expressed mathematically in terms of route flows V_r as follows, where the first statement signifies that it applies only for every $\{m,r\}$ for which δ_{rm}=1:

$$\forall \delta_{rm} = 1 : \begin{cases} \sum_{l=1}^{L} \delta_{lr} \cdot c_l(V_l) - d_m(V_m) \geq 0 \\ V_r \geq 0 \\ V_r \cdot \left[\sum_{l=1}^{L} \delta_{lr} \cdot c_l(V_l) - d_m(V_m) \right] = 0 \end{cases} \tag{3.40}$$

where

$$V_l = \sum_{\rho=1}^{R} \delta_{l\rho} \cdot V_\rho \quad \text{and} \quad V_m = \sum_{\rho=1}^{R} \delta_{\rho m} \cdot V_\rho$$

are the link and market flows, respectively, and where ρ denotes a route. The inverse demand function d_m is defined as a function of OD flow V_m, rather than of flows on distinct routes, because of the assumed perfect substitutability: people do not care about any characteristics of routes except their costs.

Beckmann, McGuire and Winsten (1956) have shown that the equilibrium problem (3.40) can be formulated and solved as an equivalent convex optimization problem. This remains true even when direct link interactions are present, provided they are symmetric (Sheffi, 1985). Such a formulation facilitates analysis of the existence and uniqueness of equilibria as well as finding them numerically. (Equilibria are typically unique in terms of link flows and OD flows but not in terms of route flows.) The objective to be minimized in this equivalent optimization problem involves integrals of average cost functions $c_l(\cdot)$ between 0 and V_l, summed over all links, minus the integrals of marginal benefits $d_m(\cdot)$ between 0 and V_m, summed over all OD pairs. This objective has no meaningful economic interpretation, and is best viewed as an artificial mathematical construct that produces the equilibrium conditions (3.40) as the necessary first-order conditions. The classic algorithm to solve this minimization problem numerically is that of Frank and Wolfe (1956); improved algorithms are also available (Sheffi, 1985; Patriksson, 2004).

Network equilibrium may sometimes lead to surprising and counterintuitive implications for public policy. A famous example is the so-called *Braess paradox* (Braess, 1968): adding a new link to a congested network may cause equilibrium travel times to increase! Intuitively, this

can happen if using a newly available route results in a lower average time but a higher marginal contribution to congestion than using competing routes.[35] Formally, it is possible because the objective function whose minimization yields the UE conditions, described above, is different from the negative of social surplus (benefits minus costs); therefore users may use the new link even if, due to congestion, using it lowers social surplus. Another paradox, known as the Downs-Thomson Paradox, occurs in a simple two-link, two-mode network in which one mode (public transport) operates with scale economies. When the capacity of the other mode (a road) is increased, the average cost of both modes can go up!

Such paradoxes are extreme examples of "induced demand," which is simply a consequence of downward-sloping demand curves as discussed in Section 5.1.3. They occur because user prices are not set optimally. We show in Chapter 4 that optimal pricing for a network involves link-based tolls that bridge the gap between average and marginal cost. Including these in the user equilibrium conditions (3.40) would makes those conditions correspond to the necessary first-order conditions for maximizing social surplus; the Braess paradox could then no longer occur.

If route choice is stochastic, the UE concept extends to the *stochastic user equilibrium* (Daganzo and Sheffi, 1977), in which no traveler can reduce *expected* travel cost by unilaterally changing route.

---

[35] The following example is from Arnott and Small (1994). Suppose two bridges A and B, cross a river. When bridge i carries traffic volume V_i, congestion causes its travel time (in minutes) to be $V_i/100$. Two cities a few miles apart and on opposite sides of the river are connected by two routes. Route A uses bridge A and an uncongested road that takes 15 minutes to travel; route B uses bridge B and a different but identical road. (The roads follow the river bank on opposite sides.) Total traffic of 1000 reaches a user equilibrium when traffic divides equally across the two routes, resulting in flows $V_A=V_B=500$ and travel times $t_A=t_B = 15+(500/100) = 20$. An engineer notices that the roads along the river banks are circuitous, and proposes a straight causeway connecting the far ends of the two bridges; it takes only 7.5 minutes to traverse, but to do so requires crossing both bridges. It looks like a time saver because currently each bridge has only 5 minutes of congestion, so the new route C covering both bridges and the causeway takes only 17.5 minutes. However, after the causeway is built, congestion on the bridges rises: travel time on route i is now $t_i=15+(V_i+V_C)/100$, for i=A,B, while that on route C is $t_C= 7.5 + (V_A+V_C)/100 + (V_B+V_C)/100$. Equilibrium requires that all three travel times be equal; this occurs when $V_A=V_B=250$, $V_C=500$, and $t_A=t_B=t_C=22.5$. Because the causeway has enticed half the travelers onto a route with a higher marginal congestion cost than the other routes (due to its including both bridges), its availability raises travel costs for everyone compared to the situation where it had never been built.

More recent advances have shown how the existence and uniqueness of a solution to Wardrop's equilibrium conditions can be guaranteed even when asymmetric link interactions are present. The conditions can be interpreted as the solution to a mathematical optimization problem known as *variational inequality* (Dafermos, 1980). Doing so permits some far-reaching generalizations by means of a trick: a larger network is defined as multiple copies of the original one with certain links interacting with the corresponding links in a copy. Two examples illustrate the usefulness of this approach.

The first example is to model networks dynamically by creating "time-space networks," in which each copy of a physical link corresponds to a different time period. Link interactions then arise because the travel time on the physical link during a certain period depends on past and current flows. (These interactions are asymmetric because travel time does not depend on future flows.) An example of such a model is METROPOLIS (De Palma and Marchal, 2002).

A second example treats multiple user classes that differ with respect to value of time or other preferences (Dafermos, 1972; Boyce and Bar-Gera, 2004). Each user class occupies its own copy of the network; cost interactions between links in the extended network then capture congestion arising from more than one class using the same physical link of the original network. These two examples, and other aspects of network models, are reviewed in several recent works including Ran and Boyce (1996), Nagurney (1999), Chen (1999), and Patriksson (2004).

3.4.5 Parking Search

In many urban locations, parking is sufficiently scarce that drivers spend a substantial amount of time searching for an empty space, an activity known as *cruising*. Shoup (2005, ch. 11) lists studies estimating that in various cities, 8 to 74 percent of cars in downtown traffic at any moment are cruising, for an average time of 3.5 to 14 minutes per trip. If parking spaces are underpriced, then anyone occupying one of them imposes an externality on others by making it harder for them to locate a vacant space.

This is a kind of crowding externality that shares certain features with congestion. Both are examples in which common property resources are overused. Some of these resources are more valuable than others — for example the most popular routes in the case of congestion, the best-located spaces in the case of parking. Thus the problem has two features: there is too much

use overall, and the use is inefficiently distributed across locations, with the more desirable locations being more overused.

Some interesting models capture one or the other of these features. Arnott and Rowse (1999) focus on the first. All spaces are equally desirable because they are located evenly around a circular city, as are destinations. People can choose either a mode that uses parking spaces (auto) or one that doesn't (walking). All residents considering a potential destination of interest will walk if it is close, drive if it is somewhat more distant, and forego the trip if it is more distant still, these reservation distances all being determined endogenously within the model. If they drive, they also choose another reservation distance, that at which they will accept a vacant space if they see one. The model *may* generate multiple equilibria, including a desirable one in which most people drive and easily find a place, walk quickly to their destination, transact their business, and leave, keeping parking-space occupancy low. But it *always* generates at least one undesirable equilibrium, in which people who drive must park far from their destination, thus keeping their space occupied while they walk a long distance, which in turn maintains a high occupancy rate of parking spaces causing long search times.

Anderson and de Palma (2004) focus on the second feature mentioned earlier: the relative overuse of parking at the most desirable locations. In their model, total demand for parking is fixed, and spaces are ranked according to their distance x from a common destination (e.g. a CBD). Each traveler must choose x and commit to searching for a space there — for example by entering a side street or an off-street parking lot. The time required for that search depends on the proportion of spaces that are occupied. Just as with congestion, users ignore an external cost that is greater the more desirable the location. The result is that too many people park close to the CBD and have to spend too much time finding a parking space.

Parking is also linked to congestion because vehicles engaged in entering, exiting, or looking for a parking space slow down other vehicles.[36] Anderson and de Palma allow for this factor by postulating that vehicles parking at location x slow down those drivers wishing to park closer to the CBD than x (since they must pass location x to do so). This sets up another

[36] Shoup (2005, pp. 303-304) describes quantitative estimates of this effect in the UK and Calcutta.

externality, which tends to offset the first because the effect of cruisers is to discourage people from seeking the best located parking spaces.

A more elaborate model by Arnott and Inci (2005) considers travel to the destination and cruising for parking as two parts of a trip, the demand for which depends on the time spent in both parts. They obtain stable equilibria that they refer to as "hypercongested," in which cruising actually declines, while total time traveled increases, as demand shifts outward. They show further that the second-best optimal number of parking spaces provided, given inefficiently low parking prices, is greater than first-best optimal number, even though these spaces remove some capacity from the streets that could otherwise be used to improve traffic flow.

As we will see in the next subsection, urban parking is often supplied in such abundance that search costs are negligible. The Arnott-Inci result just cited might justify such a situation if one accepts that parking fees are impossible to enact. But other cities use parking scarcity to limit downtown traffic, a policy with considerable intuitive appeal but largely ignored within the types of model just described. Thus we anticipate a continuation of the tentative steps taken so far to model parking search and its contribution to street congestion. What is clear so far is that the results of such models are sensitive to fine details of the situation.

3.4.6 Empirical Evidence on Short-Run Variable Costs

In this section we compile estimates of short-run average variable costs of urban automobile travel. Our purpose is to illustrate how one can use existing information, including many far more exhaustive studies, to glean the most relevant information for use in policy analysis. Such information includes the overall size of such costs in typical urban areas, the relative sizes of their constituent categories, and the factors that determine which costs are external to the user. A subsidiary purpose is to immunize the reader against some of the more extreme claims that are sometimes made by advocates of particular approaches to urban transportation policy.

In the interest of simplicity and comparability, we focus on US urban commuters and present estimates in US dollars per vehicle-mile, at 2003 prices, for a medium-size car.[37] Due to the prevalence of ample parking at US workplaces, we do not include parking search costs. In converting between distance- and time-related costs, we assume that trip distance and time are those for the average US urban commute using private modes, namely 12.1 miles and 22.5 minutes (implying average speed of 32.3 mi/hr).[38]

We distinguish between private and social average cost. The former includes fuel taxes as well as the user's own congestion costs, while the latter excludes taxes but adds external costs imposed by highway users on non-users. In the case of social cost, we also distinguish between average and marginal cost. Average social cost is the total cost borne by society, including all users, divided by total vehicle-miles. Marginal social cost is the corresponding incremental cost borne due to one additional vehicle-mile of travel; it thus includes the external component of costs imposed by congestion.

For convenience, we divide variable costs into the two categories shown in Table 3.3: those borne directly by highway users as a group (so that private and social average costs are equal) and those borne at least partly by nonusers.

Variable Costs Borne Primarily by Users

(1) Operating and Maintenance. The costs of fuel, maintenance, and tires are usually assumed proportional to distance traveled, and are typically estimated from such data as fuel consumption, tire wear, maintenance experience, and prices. The American Automobile Association estimates these expenses in 2003 to be $0.131 per mile, of which just over half is for fuel and oil.[39] About $0.017 of this is for fuel taxes, which we subtract here but add later as a

[37] Interest in this topic is worldwide, resulting in a literature we cannot begin to summarize adequately. Europeans have been especially active in pursuing it, resulting in a number of reviews, models, and policy-related applications such as ECMT (1998), Van den Bossche *et al.* (2003), Quinet (2004), De Ceuster *et al.* (2005), Newbery (2005), and Nash and Matthews (2005).

[38] These are estimates from the 2001 National Household Travel Survey. See Hu and Reuscher (2004), Table 26.

[39] Davis and Diegel (2004), Table 10.11.

private payment toward fixed roadway costs, which is how they are normally viewed in the US.[40] Barnes and Langworthy (2003) show that maintenance cost rises dramatically with age, a fact that comes as no surprise to anyone who has owned an automobile more than a few years old.

(2) Vehicle Capital. Motor-vehicle ownership costs are sometimes treated in ways that to an economist seem astonishingly naïve, including confusion between fixed and variable cost and between economic and accounting cost. They can be analyzed using standard discounting techniques for capital assets (Nash, 1974). As a starting point, we can approximate the combined interest and depreciation costs, averaged over the life of the car, by applying the *capital recovery factor* to the price of a new car.[41] In the US, that price was on average $21,120 in 2003, cars were driven 12,242 miles per year, and the median lifetime was approximately 16.9 years.[42] Using the continuous-time capital recovery factor to annualize at an interest rate of 6 percent, the average ownership cost comes to $0.162 per mile, a potentially important cost component.

Vehicle capital cost varies considerably by age of the vehicle. This is determined by examining the shape of depreciation: i.e., how the loss in value each year varies over the life of the vehicle. Using international data, Storchmann (2004) finds that the market price of automobiles largely follows a declining exponential pattern by age, declining on average by 31 and 15 percent per year in OECD countries and developing countries, respectively. It can be shown that such a depreciation pattern, in which the absolute depreciation cost is greater for a

---

[40] Average US federal and state gasoline tax rates in 2003 were $.184 and $.191 per gallon: US FHWA (2004), Table MF-121T. Average fuel consumption for automobiles was one gallon per 22.3 miles: *ibid.*, Table VM-1.

[41] The capital recovery factor ρ is the annual expenditure in each year from 1 to T that has a present value of 1.0 (computed at interest rate r). For an asset whose initial cost is K, ρK can be interpreted as interest plus depreciation on the asset's current value, when that current value takes the unique time path keeping interest plus depreciation constant. The formula for the capital recovery factor, given by Meyer, Kain, and Wohl (1965, p. 177), can be written as $r/(1-\delta^T)$, where $\delta=1/(1+r)$. For a clear derivation, see DeNeufville and Stafford (1971), ch. 8. If interest is compounded continuously, the formula becomes $r/(1-e^{-rT})$. The discrete version is more commonly used; in this chapter, we use it unless T is not an integer.

[42] New car price: Davis and Diegel (2004), Table 10.10; the 2002 figure is multiplied by 0.985, the change in the consumer price index for new cars (Table 10.14). Annual miles per vehicle: *Highway Statistics, 2003*, Table VM-1. Median lifetime: Davis and Diegel (2004), Table 3.9 for a 1990 model car. This lifetime has grown dramatically, from 12.5 years for a 1980 model car.

newer than an older car, implies that value of the car to the user is also declining, presumably because of rising maintenance costs and technological obsolescence.

This average ownership cost does not tell us the marginal depreciation cost of operating a car conditional on owning it. Barnes and Langworthy (2003) analyze data from *Official Used Car Guide* of the National Association of Automobile Dealers in order to determine the effect of increased mileage on a given vehicle's market price. They conclude that marginal depreciation cost is $0.062/mile, or 38 percent of the average ownership cost just computed.

In asking about the cost of driving, do we want only this incremental depreciation, or should we include the much larger time-related depreciation as well? That is, should we consider most capital cost as fixed? It depends on how output (vehicle-miles per year) is expanded for the question being asked. Consider two policies that increase the aggregate number of vehicle-miles traveled on commuting trips (to and from work), one by affecting commute length and the other by affecting commute mode. The first does not affect the size of the vehicle fleet, so the applicable marginal ownership cost includes only distance-related depreciation. The other causes some workers to increase their auto ownership and still others to impose inconvenience on family members by tying up the family vehicle for part of each week. In this latter case, some or all of interest and time-related depreciation cost is variable as well. In the table, we show the case where it is all variable, i.e. the vehicle fleet expands proportionally to vehicle-miles.

(3)Travel Time. Our review in Section 2.6.5 suggests that a typical value of time for work trips is 50 percent of the wage, or approximately $9.14/hr for U.S. metropolitan areas in 2003.[43] This amounts to $0.286 /mi for a single-occupant automobile moving at 32 mi/hr, the average speed for US urban commuting trips by private vehicle in 2001.[44] We use this figure for both average private and average social cost, since it is all borne by users as a group.

Our analysis of congestion shows that there is in addition an external social cost of driving in urban areas due to the contribution of a given vehicle to travel delays for others. This

[43] Mean hourly wages for metropolitan areas in 2003 were $18.29, from US BLS (2004), Table 1-1.

[44] Hu and Reuscher (2004), Table 26.

external cost varies greatly by time and location. Parry and Small (2005) review a number of studies and suggest that a nationwide average for 2000 might be $0.035/mi in the US and twice that in the UK. Most of the costs are in urban areas, which account for 65 percent of vehicle travel (US FHWA, 2004, Table VM-1), so we may assume the figures are about 40 percent higher in US urban areas, or $0.054/mi after updating to 2003.[45] Because commuting is more heavily concentrated in peak periods, the external congestion cost of a commute trip is probably higher; we add 50 percent to the above figure to account for this, making it $0.081. This external congestion cost is therefore added to the average social cost of travel time to obtain marginal social cost of $0.367. As a point of reference, the ratio of marginal to average travel time just derived (1.28) would for the BPR function of (3.9) occur at a volume-capacity ratio V/V_K of 0.84 (0.82 for the updated BPR for freeways).

Traffic congestion also imposes time costs on pedestrians, but we are aware of no estimates so do not include one here. We suspect that aside from accidents, which we treat below, the external cost of traffic to pedestrians is mostly aesthetic and depends strongly on the specifics of urban design.

We also omit the external time cost imposed on transit agencies and their riders due to buses and streetcars sharing streets with automobiles. To take an extreme example, the 15 percent reduction in automobile traffic within Central London during the daytime hours due to congestion charging is estimated by Transport for London to have speeded up bus travel in the area by six percent;[46] scaling the results of Small (2004) accordingly, this may have provided benefits of time savings to drivers and users together valued at nearly one-fourth of initial agency operating costs. [47] We also omit the additional costs to freight that may occur due to paid

[45] We update using hourly earnings in private industries, which grew by 9.64 percent between 2000 and 2003 (US CEA, 2005, Table B-47).

[46] Transport for London (2004), pp. 2-3.

[47] Small (2004), using earlier TfL estimates of a 9 percent increase in bus speed, estimates time savings equal to 35 percent of initial agency operating costs.

drivers' higher values of time and the inventory costs of delays to expensive vehicles and their loads.[48]

(4) Schedule Delay and Unreliability. Just as the tradeoff between travel time and money implies a value of time, the tradeoff between non-ideal travel schedules and money defines the cost of putting up with those schedules. These are called *schedule-delay costs*. As in Section 3.4.3, we let β and γ be the marginal cost of making an early schedule earlier or making a late schedule later, respectively, assumed to be constant as in the model by Small (1982) described in Section 2.3.2.

Empirical estimates show schedule-delay costs, like travel-time costs, to be substantial. For example, the average commuter in the sample used by Small (1982) incurred an amount of schedule delay equivalent, given the estimated coefficients, to 7.0 minutes of in-vehicle travel time, or \$1.07 at the value of time just given.[49] With the average urban commuting distance stated above, this cost is \$0.088/mi. We assume this includes the cost of unreliability due to congestion.

Now consider the external component to be added to this for social marginal cost. The estimates above suggest that in the current situation, the ratio of schedule delay to time cost of congestion is 0.93.[50] Note that this is close to the ratio of 1 that applies in the unpriced equilibrium of the basic bottleneck model with identical desired arrival times, while this ratio decreases when introducing a dispersion in desired arrival times; see equation (3.37). We assume

---

[48] See, for example, Golob and Regan (2001). The economic effects of congestion are widespread (Weisbrod , Vary, and Treyz, 2001) but since most of these are manifestations of the delays and uncertainties that we attempt to measure directly, including them would involve a great deal of double-counting. This issue is discussed at greater length in Chapter 5.

[49]This calculation is based on the full frequency distribution for schedule delay, which is summarized but not fully reported in Small (1982, Table 1). Of the 527 commuters, 318 arrive an average of 17.0 minutes early and 22 arrive an average of 7.27 minutes late. I assume no schedule-delay cost for the 187 who arrive on time, thus ignoring the penalty indicated by the coefficient of variable *DL*. As noted in Section 2.3, the equation implies that each minute of early or late schedule delay is worth 0.61 min. or 2.40 min. travel time, respectively. Therefore, the average commuter's schedule delay is worth [(318)(17.0)(0.61) + (22)(7.27)(2.40)]/527 = 7.0 min. travel time.

[50] We assume the average time cost of congestion is one-third of the average time cost of the entire trip, or \$0.095. Equivalently, we assume that congestion accounted for one-third of the average commute trip time of 22.5 minutes, and that schedule delay imposed a cost equivalent to 7.0 minutes; thus the ratio is 7.0/(22.5/3) = 0.93.

that this same ratio of 0.93 characterizes the marginal external costs (*mec's*) of travel time and of schedule delay. This implies an *mec* for schedule delay of $0.075/mi, which is added to the social average cost to get social marginal cost.

Variable Costs Borne Substantially by Non-users

(5) Traffic Accidents. Accident costs affect such diverse policy issues as fuel taxes, drunk-driving laws, and fuel-efficiency standards. However, estimating them requires care and sophistication, and estimating their external components even more so because the responsibility for accident costs is shared in complex ways among victims, their relatives and friends, other parties to accidents, insurance companies, and government agencies.

A good starting point is the study by Blincoe *et al.* (2002), who estimate total tangible economic costs due to US motor vehicle accidents in 2000. Their result is $249.6 billion, or $0.091 per vehicle-mile, at 2003 prices.[51] The largest categories are productivity loss due to injury and death, property damage, travel delay, and insurance administration.

Productivity loss, however, is a poor measure of the value to an individual of avoiding a casualty or injury. A more theoretically justified measure is the individuals' willingness to pay for reducing the probability of such an event. Traffic hazards raise all drivers' risk of being hurt or killed; their willingness to pay for a reduction in that risk is the relevant measure of the cost those hazards impose. To take an example: suppose people are willing to pay $5,000 each to reduce the risk of fatality from 0.001 to zero. The willingness to pay per unit of risk reduction is then $5,000/0.001 = $5 million. Equivalently, 1,000 such people would in aggregate pay $5 million and would reduce expected fatalities by one. As a short-hand, such willingness to pay is summarized as a *value of a statistical life* (VSL) of $5 million. Similarly, the magnitude of

[51] Their cost estimates, in 2000 dollars, are given in their Table 1. To state them in 2003 dollars, we increase them by 8.25%, the average between the growth of hourly earnings and the growth of the Consume Price Index for all urban consumers and all items (CPI-U); see CEA (2005), Tables B-47, B-60. Vehicle-miles traveled in 2000 were 2,747 billion, from US FHWA (2002), Table VM-1.

people's willingness to pay to reduce the risk of specific kinds of injuries may be expressed as the *value of a statistical injury*.[52]

There is considerable empirical evidence on the magnitude of these values. For fatalities, most comes from observing wage premiums required by workers in competitive labor markets, a measurement with two advantages: risk levels tend to be stable, and people are likely to have some knowledge of them. Several reviews and meta-analyses are available to assess the large number of studies on this topic, especially prevalent for the US. The most significant unresolved issue is whether to account for variation in unmeasured working conditions across industries (Mrozek and Taylor, 2002, p. 269). Because industry differentials constitute one of the main sources of variation in risk, using industry-specific dummy variables greatly reduces the remaining risk variation and so lowers statistical precision. But if high-risk industries also offer less attractive working conditions that require a compensating wage differential, then the wage premium attributable to that risk is overestimated unless such dummy variables are used.

Mrozek and Taylor (2002), advocating inclusion of industry dummy variables, find VSL from a meta-analysis to average $1.7–2.9 million from US studies (after updating to 2003 prices). Viscusi and Aldy (2003), arguing against such inclusion, find the predicted mean for US studies to be $6.0–8.2 million. Day (1999) obtains a "best" estimate of $6.9 million from a meta-analysis including studies both with and without industry dummies, but mostly without.[53] From this evidence, it seems reasonable to adopt a range for value of life of $2–8 million, with the most likely value $5 million.

For nonfatal injuries, Viscusi and Aldy review 39 studies from around the world, finding most estimates for developed nations in the range of $20-70 thousand per serious injury

---

[52] Blincoe *et al.* (2003, Appendix A) do consider a related measure: willingness to pay for "quality adjusted life years," which they apply to both fatalities and injuries. However, they present this only as side information on alternative approaches.

[53] Mrozek and Taylor report summary figures of $1.5-2.5 million (1998 dollars, p. 253), while Viscusi and Aldy obtain mean predicted values of $5.5-7.6 million for US studies (2000 dollars, last row of Table 8). Viscusi and Aldy also report the median value for all their studies at about $7 million (2000 dollars, p. 18). Day's "best" estimate is $5.63 million in 1996 dollars (p. 24). We update to 2003 as explained in the next footnote.

(sometimes defined as an injury resulting in at least one lost workday). Miller (1993) provides estimates disaggregated by type of injury.

It is well established that the average value of a statistical life or injury rises with income, as would be expected if safety is a normal good (i.e., demand for safety exhibits positive income elasticity). However, the income elasticity appears to be considerably less than one, probably between 0.5 and 0.6 (Viscusi and Aldy, pp. 36-42). This is relevant for transferring results from the US or Europe, where most studies have been undertaken, to developing nations. It is also relevant for updating figures measured in one year to price levels of later years.[54]

Just replacing the productivity costs of fatalities used by Blincoe *et al.* (2002) by a VSL of $5 million (in 2003 dollars) raises accident costs by 71 percent. A more detailed estimate is provided by Parry (2004), who analyzes 1998-2000 US accident data.[55] Parry applies a VSL of $3 million and, for other categories including non-fatal injuries, updates valuations by Miller (1993). Parry provides figures by type of cost and type of accident, the latter described by the worst injury sustained. In Table 3.4, we show his results at 2003 prices, recalculated using a VSL of $5 million. Average social cost is $0.130/veh-mi, a figure we adopt for our overall compilation of automobile costs. The table shows that costs are dominated by accidents involving death or injury and by the "intangible" willingness to pay to avoid such outcomes.

How much of these costs are external to the individual user? To fully address this question would require models of driver behavior, insurance, tort and criminal law, and the effects of congestion on accident rates. The literature is only beginning to produce such models, despite early discussions by Vickrey (1968). To give an idea of the difficulties, consider the

---

[54] In our own procedure for updating accident costs to 2003 dollars, we approximate this finding of Viscusi and Aldy by increasing all accident costs, regardless of type, by the average between growth in the overall consumer price index and growth in nominal earnings. This is approximately equivalent to assuming that VSL, stated in real (inflation-adjusted) terms, grows at half the rate of real wages, as suggested by studies: for example, if i is the growth rate in the Consumer Price Index and g is the growth rate in hourly earnings, both for 1996-2003, then $\text{VSL}(2003\$)/\text{VSL}(1996\$) = (1+i)\cdot\{1+0.5g\} = (1+i)\cdot\{1+0.5\cdot[(1+g)/(1+i)-1]\} \approx 1 + i + 0.5\cdot(g-i) = 1+0.5\cdot(g+i)$. For i (for urban consumers, expressed as a fraction) we use 17.3% for 1996-2003, 12.9% for 1998-2003, 6.9% 2000-2003; for g we use 27.6%, 18.1% and 9.6%, respectively (US CEA, 2005, Tables B-60, B-47).

[55] Fatalities and nonfatal injury data are from the Fatality Analysis Reporting System (FARS) and the General Estimates System (GES), respectively. The latter is less accurate because it derives from field reports of police officers, who do not necessarily observe the victims, much less obtain a medical diagnosis.

simple question of whether adding more vehicles to the road raises accident costs for existing drivers – if so, this would be an inter-user externality just like congestion, and could be analyzed similarly. Empirical evidence suggests that more congestion causes a higher rate of accidents but that they are less severe, probably due to lower speeds.[56] The resulting effect on average accident costs is ambiguous, and for this reason it is sometimes assumed to be zero.[57]

Parry (2004) allocates various costs of single- and multiple-vehicle accidents in plausible but largely heuristic ways to determine how much is external. For example, 85 percent of medical and emergency-service costs are assumed external in his "medium" scenario, as is 50 percent of property damage.[58] More important, half the cost of injuries or fatalities to occupants of a given vehicle involved in a two-vehicle crash is assumed to be caused by the other vehicle, hence external. All these external costs cause social marginal cost to exceed private cost; those imposed on non-users also cause social average cost to exceed private cost. We allocate the external costs computed by Parry assuming that those based on willingness to pay for risk reduction are inter-user externalities while all others are imposed on non-users. The resulting estimates are shown in Table 3.3; the accident externality (social *mc* minus private *ac*) is 53 percent of private *ac*—a serious impediment to efficient resource allocation. This percentage can

---

[56] See Traynor (1994) or Fridstrøm *et al.* (1995). Lindberg (2001) notes that detailed Swedish studies indicate that in urban areas, the net effect of traffic on accident rates (not costs) is negative for crashes between cars and "unprotected" users (pedestrians, bicycles, mopeds), with elasticity ~-0.5; zero for car crashes between intersections; and positive for multi-car crashes at intersections, with elasticity ~0.20–0.45. An interesting implication is that the marginal external cost of a bicycle or moped is negative: by adding to the traffic stream, it lowers the probability of accidents involving vehicles other than itself, perhaps by causing other drivers to be more alert.

[57] The results of Lindberg (2001, Table 6) for urban cars, given his value of θ=0.5 for the share of two-car collision costs borne by each vehicle, imply that E, the risk elasticity, is usually substantially negative. On the other hand, Edlin and Karaca-Mandic (2003) find that traffic density (at the level of a US state) increases insurance premiums, insurers' payouts, and possibly fatalities (with borderline statistical significance) in high-traffic states. Parry's (2004) "medium" scenario that we cite below implies that E is mildly positive. This follows mainly from his assumption that in two-car collisions, each car imposes an external cost equal to half the other driver's injury damages (p. 356); given that each car's occupants bear their own injury damages, this assumption is equivalent to a risk-elasticity E=0.5 for such collisions.

[58] The 50% proportion, mistakenly stated as 25% in Parry's text (p. 356), is confirmed by personal correspondence, 15 April 2005.

be compared to the 74 percent figure implied by Lindberg (2001, Table 6) for urban car traffic in Sweden, using a very different procedure.[59]

The externality derived here assumes that all drivers and all vehicles are identical. Thus, it measures the over-incentive to drive if the external cost is not offset through other incentives. When one considers more specific decisions, such as driver behavior or vehicle choice, the problem of external costs may be even greater. Two prominent issues involving such decisions are alcohol consumption and vehicle size. Levitt and Porter (2001) estimate that drivers who have been drinking impose more than seven times the external accident risk of other drivers.[60] This kind of finding is one justification for the considerable attention devoted to policies toward drunk driving.

As for vehicle size, White (2004) finds that that the probability of an automobile driver or passenger being killed in a two-vehicle crash is 61 percent higher if the other vehicle is a "light truck" (van, pickup truck, or sport utility vehicle) than if it is another car. For a pedestrian or a motorcyclist, the risk is 82 percent and 125 percent higher, respectively, if hit by a light truck. She also calculates that replacing a million light trucks by automobiles in the US would eliminate 30 to 81 fatalities annually. The incentive problem is highlighted by another finding: the larger vehicle is safer for its own occupants in a given two-car crash.[61] So long as the vehicle stock is diverse, so that many accidents involve vehicles of different sizes, White's findings suggest that the greater use of light trucks as passenger vehicles in the United States may be imposing very large accident costs. White also notes several reasons why such external costs are not likely to be eliminated by insurance or tort law.

Current laws in most nations address accident externalities in various ways: for example through tort law, criminal law, and insurance regulation. Boyer and Dionne (1987) and White

---

[59] We compute this as $28/(76\cdot\theta)$, with θ=0.5 being the fraction of costs incurred by a typical car in a multi-car crash, from Lindberg's Table 4. Private average cost in Lindberg's notation is $\theta\cdot r\cdot(a+b)$.

[60] An even larger ratio is estimated by Miller , Spicer, and Levy (1999), based on older data.

[61] White finds that this perceived safety is more than offset if the driver changes behavior to that typical of drivers of light trucks as a whole. Even if drivers engage in this kind of compensating risk-taking behavior, so that their observed accident risk is unaffected by the size of their vehicle, the incentive to drive a bigger vehicle remains once one accounts for the effort required for vigilance against accidents (see Steimetz, 2004, ch. 2).

(2004) provide helpful analyses. Whether such arrangements provide efficient incentives to avoid causing accidents depends to a large extent on how they affect the marginal decisions of those potentially causing an accident, prior to its realization. (This is an important reason why per-mile insurance premiums, also known as "Pay-as-you-drive" schemes, have gained interest as a substitute for fixed, yearly premiums.) Infrastructure investments and mandated safety features on vehicles are often justified on grounds of reducing accident costs, and indeed they may do so. However, consumers are likely to partly offset such benefits by choosing more dangerous driving habits and by simply driving more, since the risk associated with more and/or more aggressive driving have been reduced.[62]

(6) Government Services. Governments provide any services to highway users, from pavement maintenance to police patrols. We estimate this as the sum of three components of disbursements identified by the US Federal Highway Administration (2003, Table HF-10): (a) maintenance and traffic service, (b) administration and research (after subtracting a portion that we prorate to capital outlays), and (c) highway law enforcement and safety. These total $53.2 billion or $0.018/vehicle-mile, of which the majority is for highway maintenance. Most is covered by state and local governments, in about equal proportions. We somewhat arbitrarily take the private portion of these costs to be represented by state-imposed fees for vehicle licenses, vehicle title certificates, and drivers' licenses, which we estimate at $0.005/mi.[63] States also use fuel taxes for these costs, but we allocate fuel taxes later as payments toward road capital costs. The local portion of government services is largely covered by general tax revenues, so may be considered a subsidy to motor-vehicle operations. This subsidy is quite large in aggregate, but a very small portion of average costs of travel.

---

[62] Such offsetting behavior was postulated by Peltzman (1975) and has been tested empirically in many contexts, mostly confirming. For a review and recent example, see Winston, Maheshri, and Mannering (2006).

[63] State registration fee receipts in 2003 were $7.478 billion, divided by passenger-car vehicle-miles traveled of 1,661 billion (US FHWA, 2003, Tables MV-2, VM-1). Fees for drivers' license and certificates of titles were $2.288 billion, divided by total vehicle-miles of 2,891 billion (same sources).

(7) Environmental Externalities. Extensive studies have been carried out of the health costs of major air pollutants in the lower atmosphere — particulate matter, nitrogen oxides, volatile organic compounds, sulfur oxides, carbon monoxide, and ozone (a product of atmospheric reactions involving other primary pollutants). To form such estimates, one must know emission rates, how emissions determine ambient air concentrations, how ambient concentrations damage people's health, and the costs of that damage including people's willingness to pay to avoid it. There are uncertainties in each of these steps, leading to a range of estimates; yet a reasonable consensus has emerged on the order of magnitude of the costs.

Even more striking is the agreement on the main components of these costs. Numerically, health costs of air pollution are overwhelmingly dominated by mortality, which in turn is dominated by the effects of particulate matter. Some of the particulate matter is emitted directly, but a substantial portion is formed in the atmosphere from nitrogen oxides, sulfur oxides, and hydrocarbons. Ozone also has important health effects but has generally not been linked to long-term mortality.[64]

With mortality dominating air-pollution costs, VSL is even more important for them than for accident costs. Furthermore, fatalities from air pollution usually occur many years after the time of exposure and among elderly people, which raises two additional analytical issues. First, because VSL is measured from the relationship between *current* willingness to pay and *current* fatality risk (as in labor-market studies), the willingness to pay for changes in fatality risk that take place far in the future should be discounted to the present time just like any other expenditure. This is widely accepted among analysts, although some people mistakenly think of it as "discounting lives" and therefore objectionable. Second, an individual's VSL may depend on that person's remaining expected life span. Here the evidence is equivocal. Alberini *et al.* (2004) "find weak support for the notion that [VSL] declines with age, and then, only for the

[64] Bell *et al.* (2004) identify a correlation between mortality and ozone concentrations over very short time periods, using daily data. As with any such study using daily time series, some or all of the correlation may be due to "harvesting," whereby short-term changes in air quality determine the exact timing of a death that was going to occur soon for other reasons (McCubbin and Delucchi, 1999). The only reliable way to discern how much of the correlation is due to harvesting is to also measure cross-sectional correlations over longer time spans, for example current annual mortality as a function of exposure over several decades. Such cross-sectional studies have clearly demonstrated mortality effects of particulates but not of ozone.

oldest respondents (aged 70 or above)" (p. 769). By contrast, Viscusi and Aldy (pp. 50-53) conclude that VSL does decline with age.[65] Considerations of discounting and possible age dependence together make it reasonable to assume a lower VSL in evaluating environmental mortality than accident mortality. In what follows, we accept for air pollution the VSL used by the US Department of Transportation, which updated to 2003 is $3.93 million, or 79 percent of the VSL we use for traffic accidents.[66]

Using the principles just outlined, US FHWA (2000) estimates the average pollution cost of an automobile in the US in 2000 at $0.016/mile (in 2003 prices).[67] We note that 99.8 percent of this cost is due to particulate matter (both directly emitted and produced through atmospheric reactions), and 77 percent is due to fatalities. We adjust their estimate upward to reflect urban emissions, but downward to reflect reduced emissions rates between 2000 and 2003, for a result of $0.014/mile.[68] An identical number can be derived from a somewhat older study by McCubbin and Delucchi (1999).[69]

---

[65] However, they find that it does not vary in strict proportion to remaining life span, an assumption embedded in a broader technique in which risk of injuries and fatalities at different ages and levels of health status are all valued through a single constant measuring willingness to pay for a "quality-adjusted life year" (QALY). For further discussion of the QALY concept, see Krupnick (2004).

[66] The US Department of Transportation uses $2.7 million in 1990 dollars. Following the same procedure as with accident costs, we update by the average 1990-2003 growth in nominal wages (50.6 percent) and consumer price index (40.8 percent).

[67] US FHWA, 2000, Table 10, increased by 45.7 percent to convert from 1990 to 2003 prices.

[68] We take health damage per vehicle-mile to be 16.7 percent higher in urban areas than the US as a whole, based on US FHWA (2000, Fig. 6). To account for reduced emissions per car from 2000 to 2003, we extrapolate the 50 percent reduction in weighted per-mile emissions from a California gasoline car between 1992 and 2000 that was projected by Small and Kazimi (1995, Table 8); we do this by assuming emissions to be exponentially declining at a rate 8.66 percent per year, for a three-year reduction of 23 percent. Although California has tighter emissions standards than the US as a whole, allowable emissions have been declining in tandem so this should be a reasonable estimate of the rate of change in US average emissions rates.

[69] The "mid-range" estimate of air pollution costs given by FHWA and used here is roughly the geometric mean of two other estimates, "high" and "low," which they provide (their Table 10). McCubbin and Delucchi give only a high and low estimate, differing by approximately the same factor of ten as is the case for FHWA. We therefore take as a mid-range estimate for McCubbin and Delucchi the geometric mean of their high and low estimates for light-duty gasoline vehicles (top right entries, their Table 4); we update to 2003 by the average of hourly wages and consumer prices (45.7 percent); we adjust the US figure upward to reflect urban areas using their urban-to-US ratio of cost per kilogram for PM10, from their Table 5 (40 percent); and we assume the same extrapolated rate of decline in emissions per vehicle-mile as we did for the FHWA estimate (8.66 percent per year, or 67.6 percent total decline 1990-2003).

Conventional lower-atmospheric pollution has many well documented effects besides those on human health, including soiling of materials, reduced visibility, and damage to crops, materials, and ecosystems. However, attempts to measure the costs of these effects are virtually unanimous in yielding far smaller estimates than for human health effects (see e.g. Delucchi, 2000). The same is true for water pollution and noise, at least from automobiles.[70] We therefore omit such costs.

Motor vehicles are also a major contributor to global warming through the "greenhouse effect," due to their emissions of carbon dioxide. Precise prediction of effects of carbon dioxide on climate is impossible. Furthermore, most of the effects occur many decades after the emission, making it highly speculative to forecast what economic impacts those changes will produce — especially in light of the ability of individuals and societies to take countermeasures, perhaps using technologies that are currently unknown. Nevertheless, a number of studies have estimated the present value of projected costs of current emissions. Tol *et al.* (2000, p. 99) conclude that the marginal damage cost is very likely less than $50/tC" (metric ton carbon), a bound that in fact substantially exceeds most of the estimates. Indeed, more recent studies that incorporate standard discounting techniques (essential to any intelligible interpretation of future costs) and account for adaptation obtain results well below this bound. Based on this evidence, we follow Parry and Small (2005) in adopting a value of $25/tC, which for our US commuter converts to $0.061/gal or $0.003/mile.[71] Even this is probably on the high side, and the real cost could be much smaller. Nevertheless, this estimate is less than one-fourth our estimate of cost from conventional air pollutants; the two combined are $0.017/mile, the figure entered in Table 3.3.

---

[70] See US FHWA (2000) and Delucchi (2000). For a good review of literature assessing willingness to pay to reduce noise, see Navrud (2003). For a recent study using stated-preference data, see Arsenio, Bristow, and Wardman (2006).

[71] The conversion rate of 413 gal/tC is based on US National Research Council (2001), p. 5-5. The average fuel economy of US passenger cars in 2003 was 22.3 mi/gal, from US FHWA (2003, table VM-1). Global warming cost in other nations would be the same per ton carbon and therefore lower per mile due to the higher fuel efficiency of automobiles outside the US.

The upshot is that the environmental costs of motor vehicles are large in aggregate, justifying substantial expenditures on control measures, but far smaller than other costs of driving. Consequently, internalizing them on a per-mile basis would make little difference to people's travel decisions. It follows that optimal policy toward environmental externalities from automobiles would focus on specific measures to reduce the externalities rather than general measures to reduce automobile travel.

Summary

The upper panel of Table 3.3 summarizes the figures just presented for variable costs of automobile travel. These costs are very large and give some idea of the importance of policy decisions affecting use of motor vehicles. Their relative size, and especially the size of the gaps between private and social marginal cost, highlight the importance of certain categories — especially travel time, travel scheduling, and motor vehicle accidents — in such policy decisions.

3.5 Highway Travel: Long-Run Cost Functions

In order to complete the cost analysis for highway travel, we need to include the capital cost of building roads, converted to an annual or daily flow. Defining this cost as a function of capacity and adding it to a short-run cost function enables us to derive a long-run cost function by choosing, for each output, the size of highway that minimizes the two costs combined. Such a function provides a comprehensive summary of what it costs society to undertake different amounts of motor-vehicle travel in a corridor. This is important, for example, in evaluating policies designed to influence the total amount of highway travel.

The long-run cost function may be derived under the assumption that capacity is continuously variable, or that it can be built only in discrete units such as lanes. In the latter case, the resulting function is not smooth but rather is the lower envelope of several distinct short-run curves. The actual possibilities for capacity, however, are probably continuous, despite the fact that *changes* in road capacity are usually made in discrete units. The design capacity of a lane can vary widely depending on lane width, shoulders, curves, median, exits and entrances, intersections, and traffic signals. Hence it is possible to design a highway with virtually any capacity, and the question really becomes whether the cost function exhibits small or large

bumps. Choice among continuous highway capacities can be formulated analytically in an illuminating manner, as we illustrate in this section.

We first derive analytic long-run cost functions for some common situations. We then consider the impact of capacity-augmenting information technologies. Finally, we provide some empirical evidence on capital costs.

3.5.1 Analytic Long-Run Cost Functions

As shown in Section 3.1, an analytic long-run cost function can be derived by combining a short-run cost function with information about the cost of capacity. To simplify, we begin with several assumptions. First, we assume a single uniform output, measured as vehicle trips or flow on a road of unit length.[72] Second, we assume that capital investment serves solely to expand road capacity; we therefore ignore parking requirements as well as any auxiliary benefits of road investment such as higher free-flow speeds, lower operating costs, and greater safety.[73] Third, capital cost is linear in capacity:

$$K(V_K) = K_0 + K_1 \cdot V_K \tag{3.41}$$

with $K_1>0$. Fourth, interest and depreciation on capital can be written as $\rho \cdot K(V_K)$ per day, where ρ is the annual capital recovery factor[74] divided by the number of days per year during which the travel conditions under consideration apply.

We specify the average short-run variable cost (3.21) to be a function of volume-capacity ratio V/V_K during that period, as it is in every example considered in Section 3.4:

$$c(V) = c_0 + c_g\left(V / V_K\right) \tag{3.42}$$

where $c_g(\cdot)$ describes congestion-related average user cost and $c_g(0)=0$. Depending on the model,

---

[72] Additional output distinctions may prove useful in certain circumstances. One may separate high-occupancy from low-occupancy vehicles, both analytically and physically (Mohring, 1979; Small, 1983a). One may separate automobiles from trucks, and consider the additional capital dimension of pavement thickness (Small, Winston, and Evans, 1989). One could consider the joint costs of using a right of way for highway and rail transit. All these extensions lead to considerations of economies of scope and multiproduct economies of scale, an example of which is treated by Small, Winston, and Evans (1989, ch. 6).

[73] Larsen (1993) finds that these factors can substantially modify the optimal choice of capacity in practice.

[74] See Section 3.4.6, part (2).

V may represent static flow, time-averaged flow, or desired arrival rate, each measured over a given time period of duration q; thus short-run total variable cost over that period is $q \cdot V \cdot c(V)$. Several such periods h may be considered, in which case a function like (3.42) applies in each one; if so, we assume for simplicity that each has the same value of c_0.

Multiplying (3.41) by ρ and adding it to (3.42), short-run total cost may be written as:

$$C(\boldsymbol{V}, \boldsymbol{q}; V_K) = \rho K_0 + c_0 Q + C_g(\boldsymbol{V}, \boldsymbol{q}; V_K)$$

where $\boldsymbol{V}$ is the vector of flows V_h, $\boldsymbol{q}$ that of durations q_h, $Q=\boldsymbol{V}\cdot\boldsymbol{q}$ is total vehicle-trips, and C_g includes all the parts of (3.41) and (3.42) involving V_K. The first two terms of this equation are unaffected by capacity, so we can ignore them in deriving investment criteria. We thus focus on the tradeoff between the costs of capacity and congestion in the third, congestion-related, term. We consider two alternate congestion models: a static (stationary-state) model and the dynamic bottleneck model with endogenous scheduling.

Static Congestion Model

Suppose the typical weekday is divided into distinct periods h=1,...,H, each with constant flow V_h for a duration q_h. With short-run variable cost given by (3.42) in each period, the resulting congestion-related part of the long-run total cost function (cost per day) is:

$$\widetilde{C}_g(\mathbf{V},\mathbf{q}) = \min_{V_K} C_g(\mathbf{V},\mathbf{q};V_K) = \min_{V_K}\left\{\sum_h q_h \cdot V_h \cdot c_g(V_h/V_K) + \rho K_1 V_K\right\} \tag{3.43}$$

The first-order condition for minimization in (3.43) leads to the following investment rule:

$$\rho K_1 = -\sum_h q_h \cdot V_h \cdot \frac{\partial c_g(\cdot)}{\partial V_K} = \sum_h q_h \cdot (V_h/V_K)^2 \cdot c_g'(\cdot), \tag{3.44}$$

where c'_g denotes the derivative of c_g with respect to the ratio V_h/V_K. The marginal capital cost of expanding the highway is equated to marginal travel-cost savings. Solving this for V_K as a function of the vector $\boldsymbol{V}$ and substituting into the minimand in (3.43) gives the long-run cost function, individual terms of which give the corresponding congestion cost in each time period.

Kraus, Mohring, and Pinfold (KMP, 1976) and Keeler and Small (KS, 1977) estimate such a cost function, using expressway data from the Minneapolis-St. Paul and San Francisco regions, respectively. KMP use two time periods and KS use five. Both assume that the ratios of

volumes in different periods remain constant as traffic expands, thereby reducing output to just one dimension. Both assume the function $K(V_K)$ to be continuous and linear, as here, with a positive intercept K_0 in KMP and a zero intercept in KS. KMP find that optimal capacity produces a peak-period speed between 32 and 56 miles per hour, depending on parameters.[75] KS find optimal peak speed between 47 and 56 miles per hour, depending on capital cost, interest rate, and value of time.[76] Starrs and Starkie (1986) apply the Keeler-Small model, with a locally estimated speed-flow curve, to urban arterials in Adelaide, South Australia, finding optimal peak speeds of about 24 miles per hour. These results illustrate the point that peak-period congestion would not be eliminated by socially optimal investment in capacity.

It is illuminating to write the long-run cost function analytically for the special case of just one time period. Dropping the time subscripts, the total number of vehicle-trips served is $Q \equiv V \cdot q$; but as discussed earlier, V and q are really distinct outputs. This is because the cost of providing for Q differs depending on whether it results from a very high volume for a short duration, or from a lower volume for a long duration, the latter being cheaper to accommodate. The failure of most literature in transportation economics to distinguish between these two outputs has limited its ability to analyze policies that affect the duration of the peak period. We illustrate here for the case where $c_g(\cdot)$ is the power function $\alpha \cdot T_f \cdot a \cdot (\mathrm{V}/V_K)^b$ as in (3.23). Applying (3.44) and solving for V_K, we obtain:

$$\rho K_1 = q \cdot \left(\frac{V}{V_K}\right)^2 \cdot \alpha T_f ab \cdot \left(\frac{V}{V_K}\right)^{b-1} \Rightarrow V_K^* = V \cdot \left(\frac{q \alpha T_f ab}{\rho K_1}\right)^{1/(b+1)} .$$

The capacity chosen is proportional to traffic volume V, with proportionality constant depending on duration of the congested period, value of time, parameters of the congestion function, and capacity cost. Substituting V_K^* into the total cost function, we obtain after some calculations:

$$\widetilde{C}_g(V,q) = c_{bpr} \cdot V \cdot q^{1/(b+1)} \quad \text{with: } c_{bpr} = \left(\rho K_1\right)^{b/(b+1)} \cdot \left(\alpha T_f ab\right)^{1/(b+1)} \cdot \left(1 + b^{-1}\right).$$

---

[75]This applies their assumed relationship, $60/S = 1 + (V/V_K)^{2.5}$ (p. 544), to the range of peak volume-capacity ratios (V/V_K) in their Table 1 (p. 537).

[76]Keeler and Small (1977), Table 5, p. 18, second column. Their assumption of constant returns makes their results independent of demand.

We see that the congestion-related part of the long-run cost function exhibits no scale economies or diseconomies with respect to V; but it exhibits scale economies with respect to q because the same investment in capacity can accommodate more people at a given level of service if the time period is longer. This, of course, is the basis for attempts to spread traffic peaks, for example by staggering work hours. These scale economies in q are greater the more sharply curved is the congestion function, i.e., the greater is the exponent b. In the special case b=1, congestion-related costs are proportional to $V \cdot q^{1/2}$.

Dynamic Congestion Model with Endogenous Scheduling

We turn to dynamic bottleneck congestion. We have already seen that the average variable congestion-related cost is $\bar{c}_g$ as given by (3.38), in which demand is represented by volume V_d with duration q (of desired queue-exits). This equation again conforms to the restriction in (3.42) that congestion depend on volume only through the volume-capacity ratio V_d/V_K. Thus daily congestion-related total cost is in the form (3.43) with just one time period, but now with V_d replacing V_h, q replacing q_h, and $\bar{c}_g$ replacing c_g where

$$\bar{c}_g(V_d/V_K) = \begin{cases} 0 & \text{if } V_d \le V_K \\ \delta \cdot q \cdot \left(\dfrac{V_d}{V_K} - \dfrac{1}{2} \right) & \text{otherwise.} \end{cases}$$

There are two possible solution regimes. If ρK_1 is small, it will be cheaper to provide enough capacity so that no queuing occurs; *i.e.*, $V_K^* = V_d$. Capacity is then proportional to V_d and is independent of q. If ρK_1 is larger, it will be cheaper to allow some queuing, in which case (3.44) applies with $c'_g(\cdot) = \delta q$; this yields the investment rule $\rho K_1 = q \cdot (V_d/V_K)^2 \cdot \delta q$, whose solution is:

$$V_K^* = \left(\frac{\delta}{\rho K_1} \right)^{1/2} \cdot V_d \cdot q \,.$$

In this regime, optimal capacity is proportional to $Q \equiv V_d \cdot q$; the proportionality constant is greater if the composite scheduling-cost parameter δ is large or if the capacity-expansion cost ρK_1 is small.

The total congestion-related cost in the first regime is simply $\tilde{C}_g = \rho K_1 V_d$, and in the second it can be written as:

$$\widetilde{C}_g(V_d, q) = V_d \cdot q \cdot \left(c_{bot} - \frac{\delta \cdot q}{2} \right) \quad \text{with: } c_{bot} = 2\left(\rho K_1 \delta\right)^{1/2}.$$

In both regimes, total congestion-related costs show scale economies with respect to duration q: average long-run congestion-related cost $\widetilde{C}_g/(V_d \cdot q)$ is equal to $\rho K_1/q$ in the first regime and $(c_{bot} - \delta q/2)$ in the second, in both cases declining with q. This is again because if demand is spread out more, it takes less capacity to keep congestion to a reasonable level.[77]

Thus if there are no scale economies or diseconomies in capital cost (i.e. K_0=0), total long-run cost is proportional to the peak volume of desired trip completions, but less than proportional to the duration of this demand – just as we found for the static model with congestion given by a power function. In both models, then, it is important to distinguish flow from duration in considering the properties of long-run costs.

As with the short-run function on which it is based, this long-run cost function is second-best because it is constrained by the requirement that users time their trips according to their own individual interests, which does not yield the lowest possible total cost.

3.5.2 The Role of Information Technology

There has been a growing interest in the role of various information and communication technologies in the functioning of congested roads and networks. These technologies may offer other ways besides physical road expansion to increase road capacity. Two main types of information technology are discussed here: automated highway systems and advanced traveler information systems.

Automated Highway Systems

Automated highway systems (AHS) use information and control technologies that allow "hands-off and feet-off" driving. With vehicles' speeds controlled electronically, eliminating fluctuations due to human factors, safety may increase substantially: crashes could be reduced by

[77] In the limiting special case where $q \to 0$ while keeping $V \cdot q \equiv Q$ finite, only the second regime applies, and its cost becomes $C_g(Q) = c_{bot} \cdot Q$, showing no scale economies or diseconomies in Q.

26 to 85 percent on urban highways according to Transportation Research Board (1998). Because smaller vehicle spacings can be allowed, highway capacity may also rise considerably, with lane capacities potentially ranging from one to five times those for driver-controlled traffic. There is a trade-off between capacity and safety, just as with driver-controlled traffic. Due to frequent on and off ramps, the potential of AHS may be smaller in urbanized areas — precisely where capacity augmentation is most important due to high construction costs; but capacity may still nearly be double that of a conventional highway (Hall and Caliskan, 1999).

Despite the potential of AHS, various considerations warn against too much optimism. First, improved highway capacity is of limited use when bottlenecks remain. For example, if AHS technology is applied on highways leading into a city where the urban street network has limited capacity, it may merely shift congestion from highways to city streets. Second, mixed use of automated and manually operated vehicles may pose particularly high demands on the technical performance of the AHS, while at the same time yielding limited improvements in flow. Third, a single AHS highway in a network of conventional highways would induce route shifts, which could be counterproductive if Braess-type effects occur. Fourth, legal considerations and driver resistance to relinquishing control may be barriers to implementation. Other factors include costs, the need to standardize equipment across locations, and vulnerability to sabotage. It therefore remains to be seen whether or not AHS can play a major role.

Advanced Traveler Information Systems

Advanced traveler information systems (ATIS) use information and telecommunication technologies targeted to road users. Emmerink and Nijkamp (1999) provide an overview. A major purpose is to reduce the effects of unexpected incidents on travel times, thereby improving both expected travel time and reliability. Schrank and Lomax (2002), for example, suggest that roughly half of total delay across major US regions is attributable to nonrecurring incidents.

There is little doubt that information could lead to a more efficient use of an otherwise optimized road network – specifically, if the externalities in the user equilibrium of Section 3.4.4 were all eliminated (De Palma and Lindsey, 1998). However, a network with a user equilibrium need not necessarily be improved by information – yet another paradox analogous to the Braess paradox mentioned earlier. Ben-Akiva, De Palma and Kaysi (1991) provide examples where information provision is welfare-reducing.

Figure 3.12, based on Verhoef *et al.* (1996), shows why this can happen. Suppose capacity is stochastic, with a high-capacity state (subscript 0) and low-capacity state (subscript 1) occurring with equal probability. Therefore either of two average cost curves, shown as c_s, s=0,1, may apply. Since each is rising, there is an associated (social) marginal cost curve mc_s above it. Expected average cost $E(c)$ and inverse demand d are also shown, all as functions of flow V. If drivers have no information about the actual state and they are risk-neutral, equilibrium road use occurs at the intersection of d and $E(c)$, at point V_N. With perfect information, the equilibrium depends on which state occurs: for each state s it is at V_s, the intersection of d and c_s. (The optima in both states, not indicated in the diagrams, are at the intersections of d and mc_s; see Chapter 4.)

FIGURE 3.12

Let us define expected social surplus as expected total benefits minus expected total cost. Its change due to perfect information is measured by determining in each state the change in area under the demand function minus the change in total cost (which is the area under the marginal cost function). Averaging the results for the two states gives the change in expected social surplus. With net benefits shaded lightly and net costs shaded darkly, the left panel shows that expected social surplus increases. It can be shown that this result always holds for linear demand and cost functions, under relatively mild conditions.[78] It often holds in more general settings as well, for example when two parallel roads are available (Emmerink, 1998). The basic reason is that although information produces a net loss in state 0 (because it increases use of the road, which is already more than optimal), it produces a bigger net gain in state 1 (because the extreme losses due to congestion are ameliorated by a reduction in demand).

However, the right panel shows that the result is not generally true if the inverse demand function is convex, here represented in an extreme way as a kinked demand function. In this case, the net benefits in state 1 disappear because demand is unaffected, and only the net loss in

[78] A sufficient condition is that the intercept of the average cost function be no lower in state 1 than in state 0.

state 0 remains. Thus providing information causes the congested road to be even more overused in the good state, with no compensating reduction of use in the bad state, resulting in a net loss of expected social surplus.

Arnott, De Palma, and Lindsey (1991a) similarly examine information provision in the bottleneck model with stochastic capacity. They find that *perfect* information increases social surplus, but information subject to some uncertainty may reduce social surplus. So not only is there a paradox in which information can be harmful, but the results are not even monotonic with respect to how accurate the information is.

These insights suggest that the value of information may be enhanced when pricing is also in place to control congestion externalities, a question we will return to in Chapter 4.

3.5.3 Empirical Evidence on Capital Costs

In this section we address the two most important capital costs that are fixed in what we have defined as the short run, but are variable in the long run. These are the costs of building roads and parking spaces.

Roads

Capital costs vary greatly with terrain — e.g. flat, rolling, or mountainous — and with degree of urbanization. Both affect such factors as the number and types of structures required (e.g., bridges, overpasses, intersections, drainage facilities, retaining walls, sound walls), ease of access to construction sites, difficulty of grading, extent of demolition, and of course land prices.

Scale economies with respect to capacity, s_K, may be defined analogously to equation (3.3) as the ratio of average to marginal cost of capacity. Scale economies might arise from fixed costs of administration, better equipment utilization, fixed land requirements such as shoulders and medians, and efficiencies of multilane traffic flows (Mohring, 1976, pp. 140-145). Scale diseconomies could result from the increased cost of building more or bigger intersections, especially when they require complex signals or overpasses (Kraus, 1981);[79] or from a rising

[79] As Kraus notes, the envelope theorem guarantees that scale economies for a road network are identical whether capacity is expanded by widening existing roads or by adding new ones.

supply price of urban land, especially in large cities where urban land is scarce and roads use a substantial fraction of it (Small, 1999a).

Several empirical studies have examined scale economies and the magnitude of highway capital costs. Meyer, Kain, and Wohl (MKW, 1965) estimate a cost function like (3.41) based on engineering standards, assuming scale economies due to fixed costs of administration and fixed land requirements. For a six-lane expressway in a typical suburban area, their results imply scale economies, with s=1.74 (MKW, p. 207). However, there is reason to doubt their assumption that the right of way needed for median and shoulders is independent of the number of traffic lanes; physical separation of traffic and provision for stopped vehicles are often used to maintain safety in the face of high total traffic levels.

Keeler and Small (1977) estimate construction and land costs statistically from a sample of 57 highway segments in the San Francisco Bay area, based on construction data for 1947-72. Urbanization is represented by three categories: central city (Oakland or San Francisco), urban (other incorporated cities), and rural (unincorporated areas). Highway type is expressway or other arterial. Construction cost for any of these categories is assumed to be proportional to the number of lanes raised to the power $1/s$. They estimate scale economies s=1.03 (standard error 0.39), which may be taken as weak evidence for neutral scale economies.[80] Land costs are estimated as fractions of construction costs, the fractions ranging from 26.7 to 36.7 percent. Starrs and Starkie (1986) similarly estimate a power function, using data from twenty-seven projects involving urban arterials in South Australia; omitting land, they estimate scale economies s=1.28 (standard error 0.22).[81]

Kraus (1981) estimates the degree of scale economies on urban road networks while explicitly accounting for the costs of intersections, which he finds to be quite large. Using British data on costs and design standards, he estimates overall scale economies at s=1.19 for a circular

---

[80] Keeler, and Small, 1977, Table 1, using their loglinear specification. As their discussion on p. 7 makes clear, the estimates of a_6 reported in their paper should have minus signs.

[81] This uses their second estimate on p. 4, with w the number of driving lanes and including a dummy variable for curbside parking; returns to scale are $1/b$. Their first estimate, with w the width of the entire roadway in meters and no control for parking, yields scale economies of only 1.05.

urban area of 10-mile radius containing a specified highway network.[82] This value reflects substantial scale *economies* in constructing individual road segments, significantly offset by scale *diseconomies* of intersections (the latter due to the fact that their size tends to be proportional to the square of the width of the highways being connected). This offset, seen only at the network level, is ignored by the other studies just mentioned. Those studies (and Kraus as well) may furthermore overestimate scale economies by taking land prices as fixed.

Altogether, the evidence supports the likelihood of mild scale economies for the overall highway network in major cities. Scale economies are probably substantial in smaller cities in which one or two major expressways are important, and may disappear altogether in very large cities where expanding expressways is extraordinarily expensive due to high urban density.

What can we say about the average capital cost per vehicle-mile? The US Department of Commerce (1998, Table 11) estimates the depreciated value of the entire US highway capital stock, excluding land, at $1,359 billion in 1997. Annualizing at a 7% real interest rate and a 20-year average remaining life, and updating to 2003 prices, this implies an annual cost of $151 billion.[83] If we allocate this cost to vehicle classes, as in US FHWA (1997, Table V-21), we find that 51 percent is attributable to automobiles, which comes to $0.051 per automobile vehicle-mile for the entire US.[84] Given the evidence that scale economies, if any, are small, this figure is a reasonable estimate of the average cost of physical capital for urban areas as well. We add 30% for the cost of urban land, based on Keeler and Small (1977, p. 9), bringing the average capital cost to $.067/veh-mi. Because the size of the capital stock is not necessarily optimized, this figure is listed as a "short-run fixed cost" at the bottom of Table 3.3.

---

[82]The degree of scale economies, as defined here, is the inverse of the cost elasticity, estimated by Kraus at 0.84 (p. 20 and n. 4).

[83] US OMB (1992) recommends a 7% real interest rate for project evaluation. The corresponding capital recovery factor with 20-year life is 0.0944.

[84] We update capital costs from 1997 to 2002 using the ENR (formerly Engineering News-Record) construction cost index, which rose by 12.3%; and from 2002 to 2003 using the U.S. Census Bureau index for houses under construction (excluding land), which rose 4.9%. These figures are from US Census Bureau (2001, Table 928, and 2004, Table 9\21). This value for annualized capital cost of construction is more than twice as large as current capital outlays in 1997, which we estimate to be $63.8 billion (including a portion of administration and research) in 2003 dollars, from US Federal Highway Administration (1998), Table HF-10. Automobile vehicle-miles traveled in 1997 were 1,503 billion (US FHWA, 1998, Table VM-1).

Passenger vehicles contribute toward paying roadway costs in the US mainly through fuel taxes, which we estimated above at $0.017/mi and list in the table as a private average cost for roadways.

Parking

Providing parking spaces is costly wherever land is expensive. As emphasized by Shoup (2005), there are three to four parking spaces for every registered vehicle in cities. If we consider just commuting trips, and allow for a 20 percent vacancy rate, we could assume that adding a trip by car requires 1/0.8=1.25 parking spaces (at the workplace) if the vehicle fleet is not expanded, and 2.25 (including one at the residence) if it is.

Willson (1998, p. 39) and Shoup (2005, ch. 6) present evidence on costs from southern California. In suburban office parks, Willson finds the average cost per space to be $7,870 for surface lots and $15,420 for structures, both restated at 2003 prices.[85] In urban westside Los Angeles, specifically the University of California at Los Angeles (UCLA) campus, where surface lots are uneconomic, Shoup measures the incremental cost per space in parking structures, relative to surface lots, obtaining $29,160. The latter figure is calculated by dividing the construction cost of the structure (without land) by the number of additional spaces it holds beyond what a surface lot would hold; it thus has the advantage of being independent of land cost except insofar as optimal structure height depends on land costs.[86] These figures are easily within the national ranges suggested by Cambridge Systematics, Cervero, and Aschauer (1998, p. 9-18).

Annualizing with a 40-year lifetime and 7 percent real interest rate and adding an estimate of annual operating cost,[87] these figures imply an annual fixed cost per parking space at

---

[85] We update Willson's figures from 1995 to 2002 using the ENR (formerly Engineering News-Record) construction cost index, which rose by 19.5%; and from 2002 to 2003 using the U.S. Census Bureau index for houses under construction (excluding land), which rose 4.9% (US Census Bureau, 2004, Table 921).

[86] The data are for the 9 new structures or additions built at UCLA between 1977 and 2002. Based on Shoup's Table 6-1, these include four underground and five above-ground structures, the latter apparently 3 to 8 stories high.

[87] Shoup assumes a 40-year life, which at 7% real interest implies a capital recovery factor of 0.0750. For annual operating costs per space, we use the midpoint of the ranges of annual costs implied by Cambridge Systematics *et*

a workplace of $794, $1,645, and $2,676 for suburban surface lot, suburban structure, and urban structure, respectively. Assuming 250 round trips per year and a 20 percent vacancy rate,[88] the corresponding average capital cost for parking at the workplace adds $3.97, $8.23, or $13.38 per day to the average cost of a commute trip.[89] For the short-run fixed cost of parking in Table 3.3, we use the average of the two suburban figures, divided by round-trip distance of 24.2 miles, yielding $0.252/mi. It may seem anomalous that parking costs are more than three times roadway costs; but this reflects our focus on an urban commuting trip, whose parking space typically has a high opportunity cost and is not shared by any other trips, whereas the roads used for such trips are used at other times of day and so have their costs averaged over more users.

For private cost, we hazard the guess that US urban commuters pay for at most 2.5 percent of workplace parking cost on average, or $0.006/mi.[90]

Urban parking costs are clearly a significant portion of the cost of automobile travel, and are all the more remarkable because they are fully absorbed by the vast majority of US employers rather than being charged to the commuter.

3.5.4 Is Highway Travel Subsidized?

Calculations such as those shown in Table 3.3 are often used to debate whether automobile travel, or highway travel more broadly, is subsidized. Such debates can be confusing because "subsidy" has several different meanings, and for each there are conceptual issues in how to measure costs and user payments that cannot necessarily be resolved in a scientific manner.

---

al. (1998, Table 9.3, note 5), updated to 2003 prices using the Consumer Price Index; these updated annual figures are $204 for surface lots and $489 for structures.

[88] This assumes tight planning by the employer; the average vacancy rate for all US parking spaces is said to be 50 percent (Cambridge Systematics *et al.*, 1998, p. 9-17, note 15).

[89] These figures are slightly higher than the value of 1.9 ECU per space per trip used for Brussels in 1996 by Calthrop, Proost, and Van Dender (2000, p. 68). Converted to dollars ($1.27/ECU), updated to 2003 in the same way as US figures (22.0 percent), and adding a 20% vacancy rate, this amounts to $3.53/trip.

[90] Shoup (2005, p. 267) estimates that only five percent of automobile commuters pay to park, and it is clear that they usually pay only a fraction of average cost, which we take for illustration to be one-half.

We can discern at least four meanings for "subsidy," not mutually exclusive. The first is *fiscal*: is a particular set of government accounts in balance? We might seek such balance as a way of facilitating public scrutiny of financial decisions in order to encourage honest and competent management; we might also care about budgetary imbalances because raising public funds to cover deficits generally has some economic cost (the so-called "excess burden"). These concerns are reflected in the frequent use of ear-marking, or hypothecation, of highway-based revenues to be spent only on highway-related purposes.

A second meaning is *distributional*: does the system of highway finance benefit certain groups at the expense of others? Here the motivation might be understanding the political economy of decision-making, or simply the desire to promote a broad trust in the fairness of the system. These concerns, as well as fiscal ones, are prominent in the highway cost allocation studies that have been done periodically at both the federal and state levels in the US.

A third meaning involves *long-run allocation*: does the gap, if any, between social costs and revenues from highway transportation indicate likely misallocations of investment? Some discussions, especially in the literature on privatization, appear to take the position that allocation of investment across sectors of the economy are best made as in unregulated private markets, by allowing investment funds to flow out of sectors that make losses into sectors that make profits. This could be justified, for example, if the industry exhibits neutral scale economies and minimal external costs, so that average total cost approximates long-run marginal cost.

A fourth meaning is *efficiency*. Is there a discrepancy between social and private marginal cost that will cause market failures? This question is the basis for most economic analysis of optimal pricing and investment, which we describe in subsequent chapters. It has been prominent in many research projects on cost measurement sponsored by the European Union, as suggested by such project titles as "UNIfication of Accounts and Marginal Costs for Transport Efficiency" (UNITE) (Nash *et al.*, 2003). Such questions also influenced, although incompletely, the most recent US federal highway cost allocation study (US FHWA, 1997, 2000). In contrast to the first three meanings of "subsidy," this one has more of a short-run focus because of the prominence of short-run marginal cost in economic pricing theory; it is also more focused on the marginal decision-making of a single user, as opposed to a policy-maker who can influence many users simultaneously.

Conceptual difficulties abound. In assessing fiscal balance, which taxes should be considered to be user taxes? For example, what about the portion of a normal sales or value-added tax that falls on fuel? Or what about a specific exemption of fuel from such taxes? Similarly, which expenditures are undertaken primarily for the benefit of highway users, especially when they are part of broader efforts to promote public safety (as with police, emergency response, and alcohol-abuse prevention)? The same problems afflict attempts to assess distributional impacts, hence all the more so the formulation of a concept of fairness. The third and fourth meanings are somewhat more amenable to precise definitions because they can be used to address precise questions, such as what would happen if a particular suite of policies were introduced to restrain downtown road traffic; but then the most direct approach is to model the policies explicitly and forego the step of measuring the long-run average or marginal cost of expanding road traffic.

Even in addressing efficiency, the most valuable lessons from computing social and private costs most likely come from the individual components. For example, the bottom row of Table 3.3 could be used to argue that in the short run, private decision-makers are "subsidized" at the margin by $0.220 per vehicle-mile if they choose to travel by car. Given the capital decisions that have been made with respect to provision of roads and parking spaces, this measures the extent of the distortion in the average incentive facing car users. But more striking is that most of this discrepancy arises from the congestion externality, and most of the rest arises from inter-user externalities connected with accidents. These externalities are well understood to vary greatly by circumstance. So from an efficiency point of view, the table is most useful by pointing to congestion and accidents as two places to look for big savings from more efficient policies. Similarly, looking at long-run policies, we see that parking is supplied at an enormous subsidy — to such an extent, in fact, that we deemed the short-run costs of searching for parking spaces too small to bother to quantify for the US. So very likely there are big savings to be reaped from policies that reduce provision of parking spaces, especially if some of the fixed costs can be recovered by converting parking lots to other uses or by arranging for existing parking structures to be shared with nearby new users.

We therefore see little value in arguing over the extent of aggregate subsidies to road users, but much value in examining specific cost components and how decisions affecting them are made.

3.6 Intermodal Cost Comparisons

One way to use cost functions such as those developed in this chapter is to compare them for different modes. This is not as easy as it might seem, because of the many required assumptions regarding demand, geography, land use, and other factors. Inevitably, some of the cost categories discussed earlier are omitted for simplicity. Furthermore, because a cost comparison does not incorporate an explicit demand model, it cannot take into account preferences of users for service characteristics other than those quantified in the study, nor can it predict the mix of modal choices that would be efficient. Nevertheless, it is a conceptually transparent way to summarize the advantages and disadvantages of alternate modes for producing a carefully specified type of service.

The pioneering study is by Meyer, Kain, and Wohl (MKW, 1965). Their costs exclude the value of user-supplied inputs like time, but they compensate for this by constraining the various modes to provide comparable levels of service. Several later studies, including Boyd, Asher, and Wetzler (1973, 1978), Keeler *et al*. (1975), Dewees (1976), and Allport (1981), incorporate the value of user-supplied time explicitly. Others discuss differences in service quality but do not incorporate them formally (Smith, 1973; Skinner and Bhatt, 1978).

Figure 3.13, adapted from MKW, shows a typical result for commuting trips along a corridor connecting residential areas to a high-density business district. All results are for a ten-mile limited-access line-haul facility (auto-only expressway, exclusive busway, or rapid-rail line), combined with a two-mile downtown distribution route. Bus service is integrated, meaning that collection, line haul, and distribution is all done by a single vehicle. Downtown distribution for rail and for one of the bus systems is accomplished using exclusive underground right of way; for the other bus system and for auto, it is accomplished using city streets.

FIGURE 3.13

Costs for all transit modes decline as a function of hourly passenger volume along the corridor, reflecting scale economies in vehicle size. At the lowest volumes, automobile travel is cheapest. At somewhat higher volumes (above approximately 5,000 passengers per hour in this

example), the bus becomes more economical. At still higher volumes, rail transit may become the cheapest, although not for this particular set of parameters.

These results help delineate the natural markets for each of these modes. One of the problems with transit subsidies is that they have encouraged expansion of transit services beyond where they are most suitable. Rail systems are built in small cities which are more economically served by bus; bus systems are extended to low-density suburbs where auto is cheapest. Meyer and Gómez-Ibáñez (1981, pp. 51-55) identify the primary markets for bus transit in the U.S. as high-income radial commuters and low-income central-city residents. Research suggests that service has been expanded beyond these markets at least in part because of federal subsidies (Anderson 1983, Pucher and Markstedt 1983).

MKW do find that rail transit can be cheaper than bus above 30,000 passengers per hour if residential densities are higher than those applicable to Figure 3.13. The findings of some other studies are even less optimistic for rail. Neither Keeler *et al*. (1975) nor Boyd *et al*. (1973, 1978) find any situation where rail is cheaper than bus. The study by Keeler *et al*. estimates rail costs using the San Francisco area's Bay Area Rapid Transit (BART) system, which may be an atypically high-cost system (Viton, 1980b); while Boyd *et al*. use several North-American rail systems built since 1945. In contrast, Allport, with the cheaper Rotterdam rail system as his model, finds that elevated rail transit is cheaper than bus for corridor volumes above 10,000 per hour. Allport also analyzes a light-rail system, which dominates both bus and rapid rail for peak-direction passenger volumes in the range of 8,400 to 13,100 per hour[91]—in contrast to the results of Wunsch, described earlier, for a broad cross-section of European cities.

The biggest factor accounting for these differences is the capital cost of infrastructure. Kain (1999) focuses on this item, comparing more recent evidence for four types of express transit in North American cities: rapid rail, light rail, bus on exclusive busway, and bus on shared carpool lanes. Some of Kain's results are shown in Table 3.5. It is evident that at the ridership levels achieved by these systems, both heavy and light rail are many times more expensive than

[91] These figures are computed from Allport's discussion on p. 638. We have multiplied his two-way weekday 24-hour passenger demands by 0.075, the assumed ratio of peak-direction peak-hour volume to 24-hour volume (Allport, p. 636).

express bus, even before accounting for their higher operating costs as documented in Section 3.2.

Comparisons such as these have led to widespread skepticism among economists toward new rail systems. The evidence is strong that in all but very dense cities, equivalent transportation can be provided far more cheaply by a good bus system, using exclusive right of way where necessary to bypass congestion. Recent attention has been focused on designing bus systems that match more closely the quality of service offered by rail. This concept, known as "rapid bus transit," has been successfully implemented in a number of cities – most famously Curitiba, Brazil.[92]

3.7 Conclusions

Cost models may be simple or excruciatingly complex, depending on their purpose. But to be useful they must at least be rigorous. That means carefully distinguishing what is an output, what is a parameter, and what is being held constant. By doing this, the researcher can summarize a host of information in useful ways, such as average costs, marginal costs, and economies of scale and scope. Carefully distinguishing between long-run and short-run variations, and rigorously defining the associated cost functions, help to clarify what costs need to be considered in particular policy decisions.

Such summary measures and their underlying components can, in turn, be used to understand significant features affecting how transportation services are, or could be, supplied. As just one example, recognizing that the time supplied by users is a necessary input into the production of trips leads to a recognition of scale economies for producing scheduled services such public transit, which in turn implies that marginal-cost pricing, often recommended on efficiency grounds, will produce revenue shortfalls.

Turning to congested highways, it is possible to use cost functions to understand the outcome of allowing many users to choose trip frequencies, routes, and possibly travel schedules

[92] See, for example, US General Accounting Office (1991), International Energy Agency (2002), Levinson *et al.* (2003), and Hess, Taylor, and Yoh (2005).

endogenously, even as their choices in aggregate determine the pattern of travel times they each face. The relatively simple static model of congestion provides a valuable insight about the inefficiencies of such situations and (as we shall see in the next chapter) about the possibilities of pricing to alleviate them. When applied to an artificial network of two or three competing routes, the static model shows that unexpectedly perverse results can occasionally occur from widening a highway or from adding a new highway link; when applied to a realistic network, it enables planners to estimate the effects of changes in transportation demand or infrastructure, so long as these changes are modest enough so that the model's assumptions continue to be valid approximations. Dynamic models provide further insights: for example, that the total congestion cost may be dominated by how strongly people care about their preferred schedules rather than how much they value the time lost in delays. We have argued that further progress in refining our understanding of policies on the public agenda today will require researchers to combine economic analysis with increasingly detailed engineering models of dynamic congestion formation and evolution.

The most important components of transportation costs have been studied sufficiently that reasonable empirical estimates of them can be compiled and compared. We have done this for urban commuting by automobile in the US. The results suggest that the costs are dominated by congestion, accidents, and parking — all areas where charging users for these costs is not straightforward. There are considerable differences between marginal social costs and average private costs, suggesting the likelihood of overuse. We will have more to say about this when we consider the implications of different ways to price automobile users in the next chapter.

Figure 3.1 The fundamental diagram of traffic congestion in three forms.

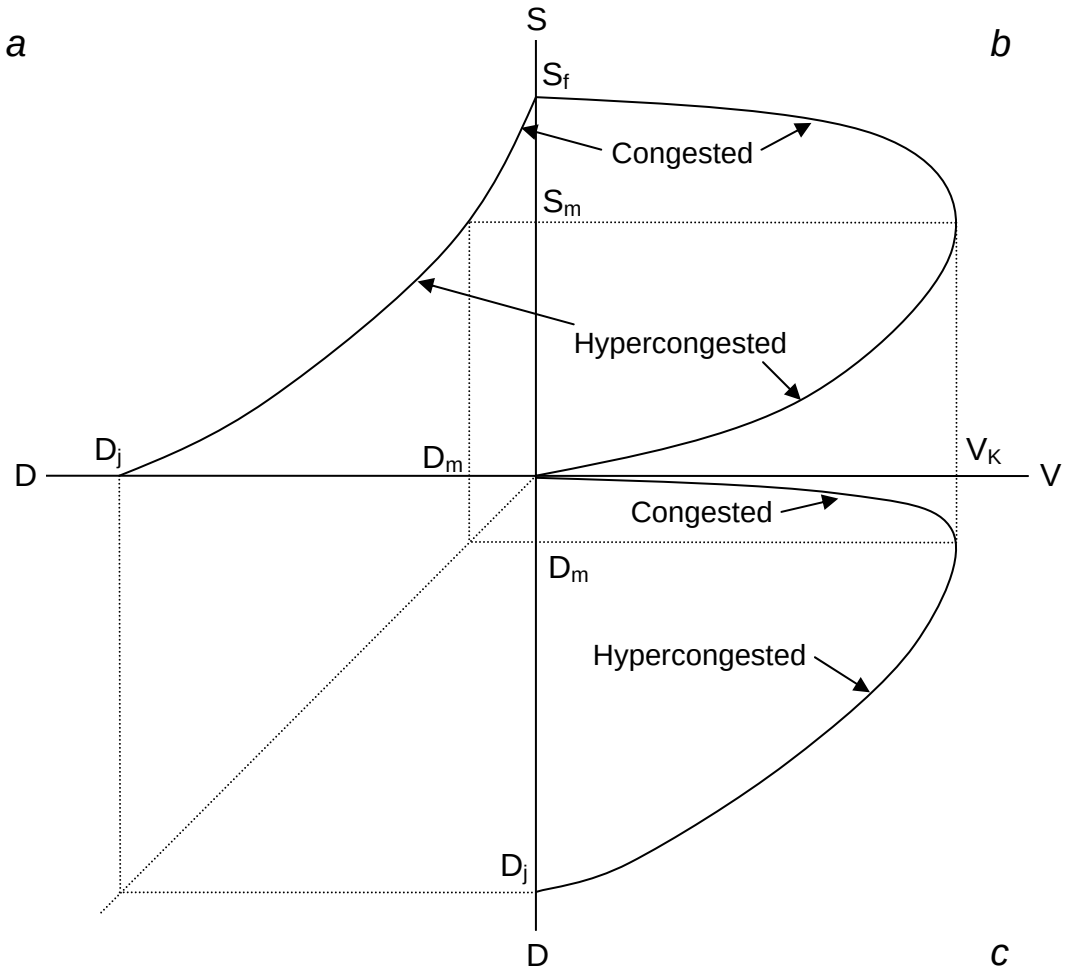
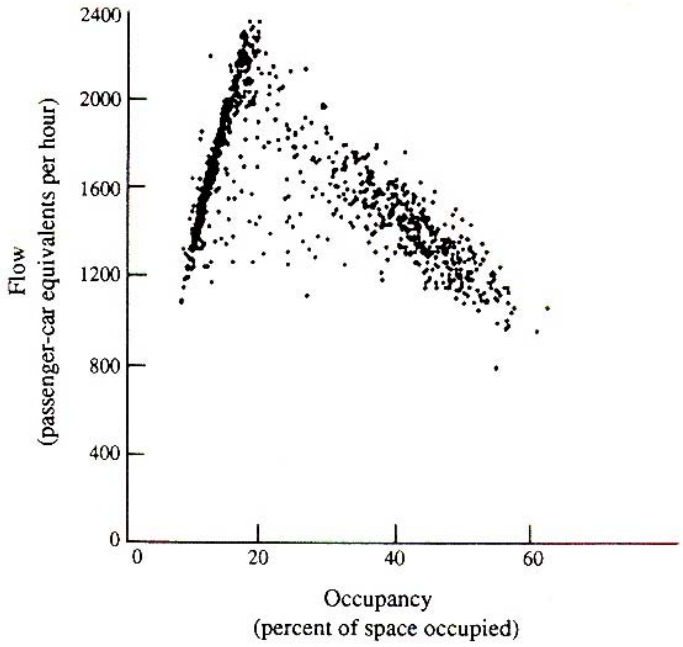


Figure 3.2 Flow-Density and Speed-Density scatter plots.

(a) Queen Elizabeth Way (Toronto). Adapted with permission from Hall *et al.* (1986, p. 204), copyright 1986, Permagon Press plc.



(b) Santa Monica Freeway (Los Angeles). Adapted from Payne (1984, p. 145), with permission from Transportation Research Board, National Research Council, Washington, D.C.

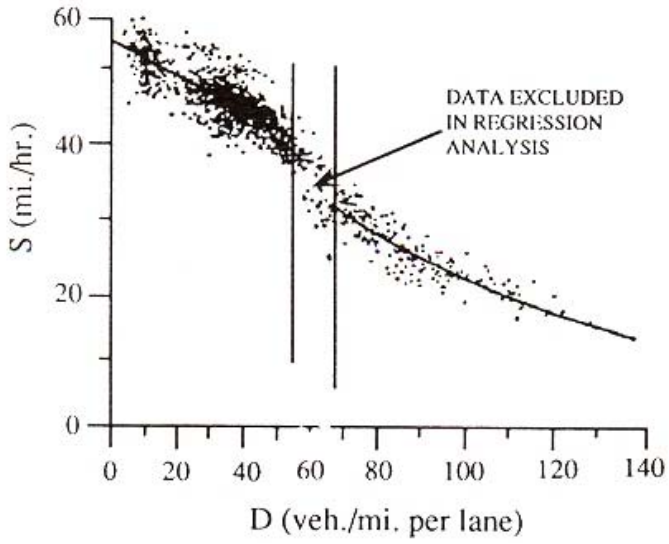


Figure 3.3 Washington, D.C. Beltway (Adapted with permission from Boardman and Lave (1977, p. 346), copyright 1977, Academic Press, Inc.

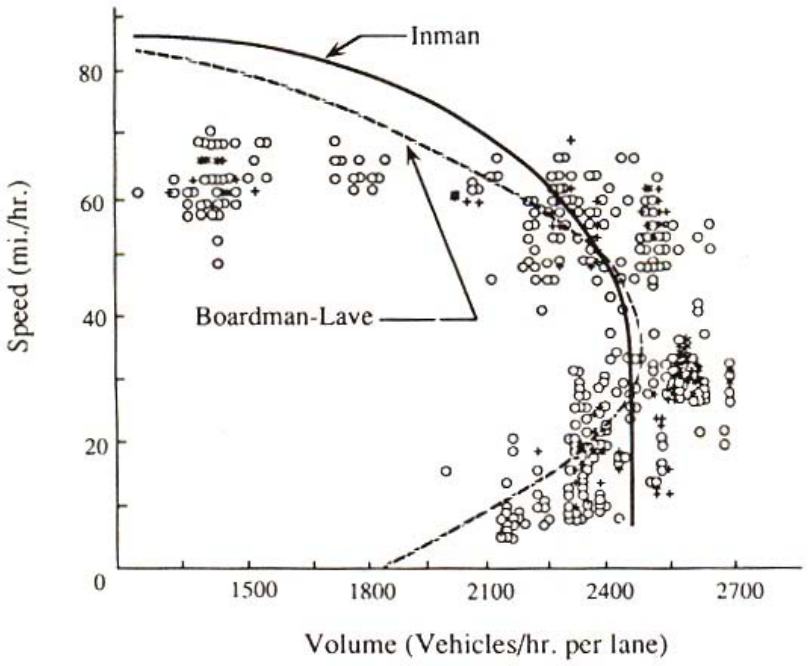
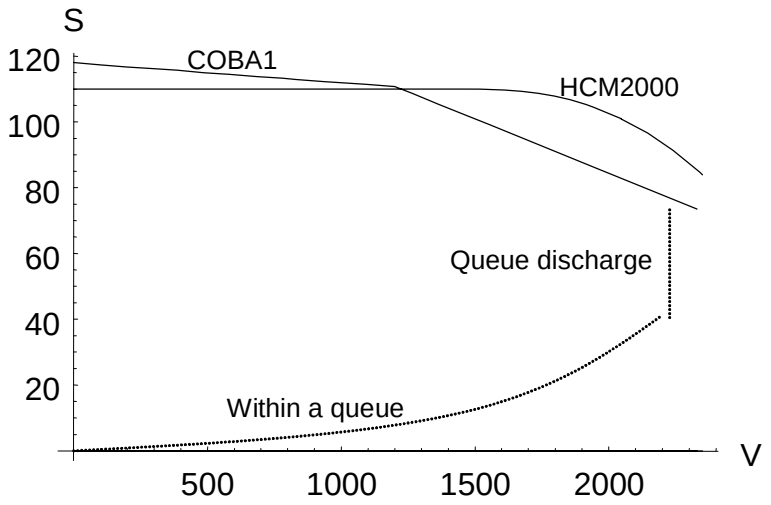


Figure 3.4 Sample speed-flow curves for US and UK government analyses.



Notes: S in km/h, V in veh/l/h. HCM2000 and COBA11 curves assume an expressway with no hills, bends, or heavy vehicles. Capacities under these idealized conditions are V_K=2330 (COBA) and V_K=2350 (HCM), and the break points V_B are 1200 (COBA) and 1450 (HCM). The two COBA segments are given by $S = 118 - 0.006 \cdot V$ for $V \le V_B$ and $S = 110.8 - (33/1000) \cdot (V - V_B)$ for $V_B < V \le V_K$. The two HCM segments are given by $S = 110$ for $V \le V_B$ and $S = 110 - [(730/28) \cdot ((V - 1450)/900)^{2.6}]$ for $V_B < V \le V_K$. The two additional (dotted) segments proposed by Hall *et al.* (1992) are hand-drawn.

Figure 3.5 Inflow rates and travel times in time-averaged models

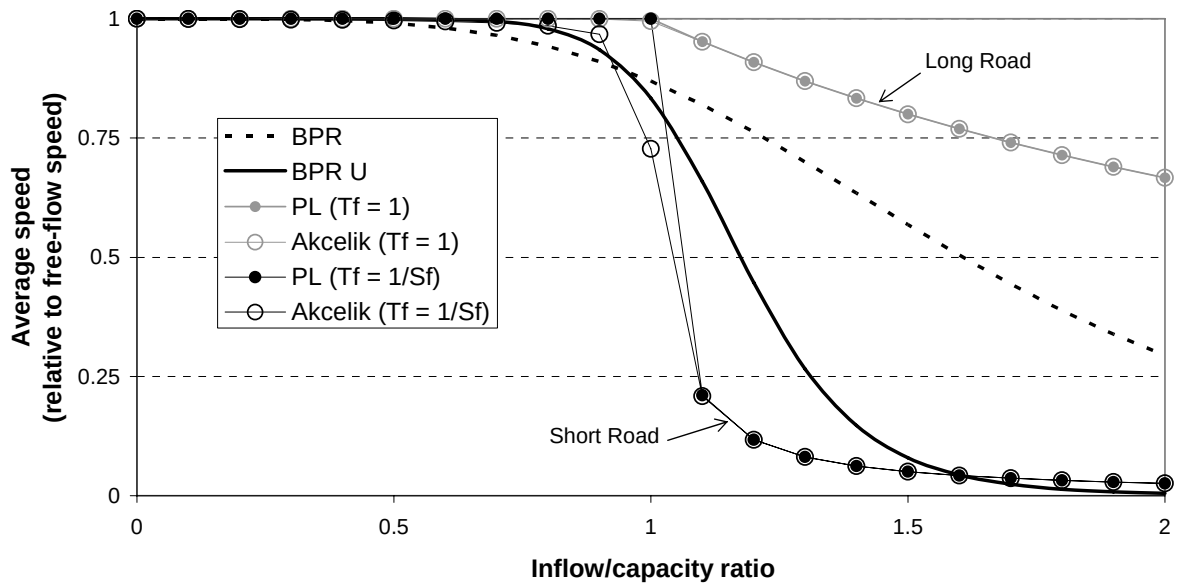


Figure 3.6 Deterministic Queueing

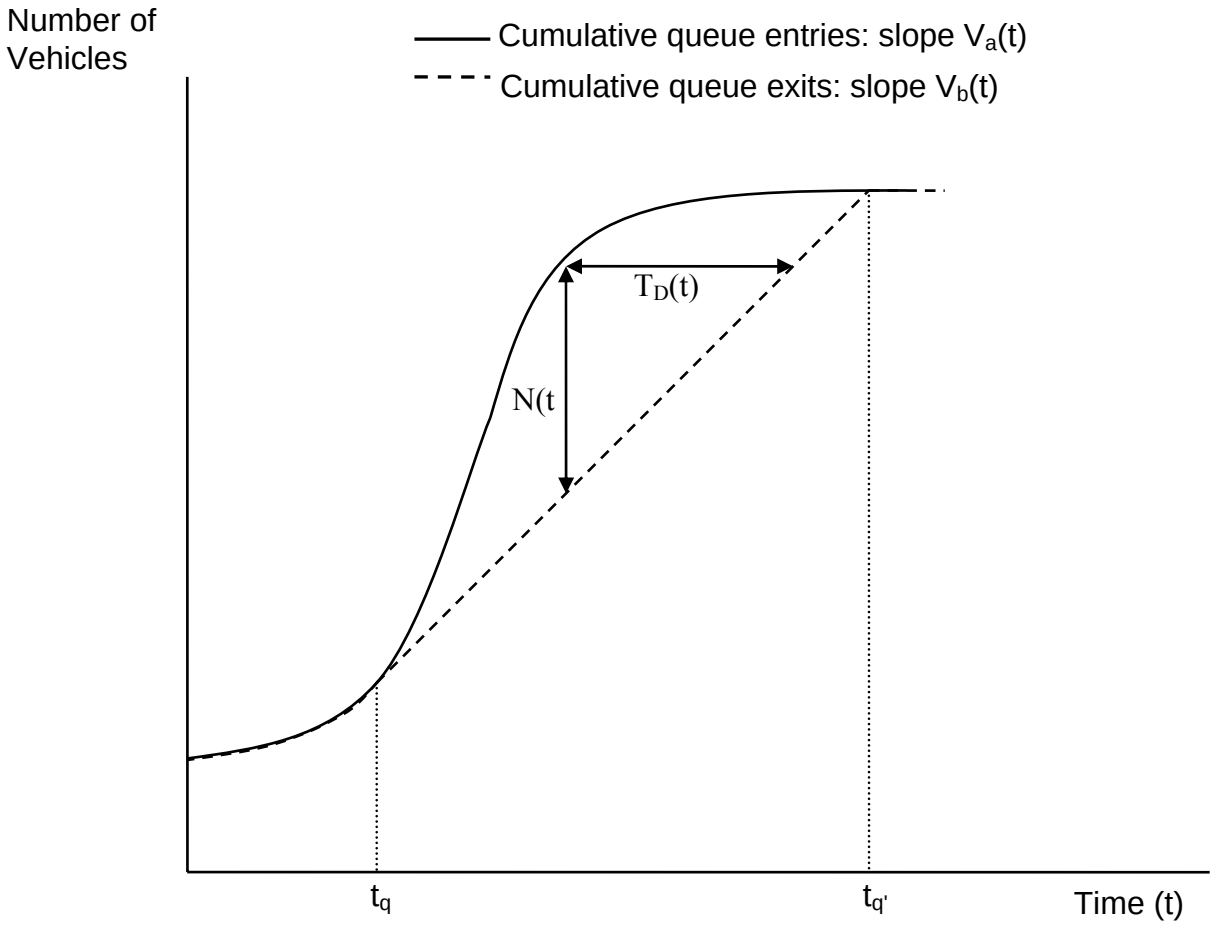


Figure 3.7 The conventional stationary-state average cost curve

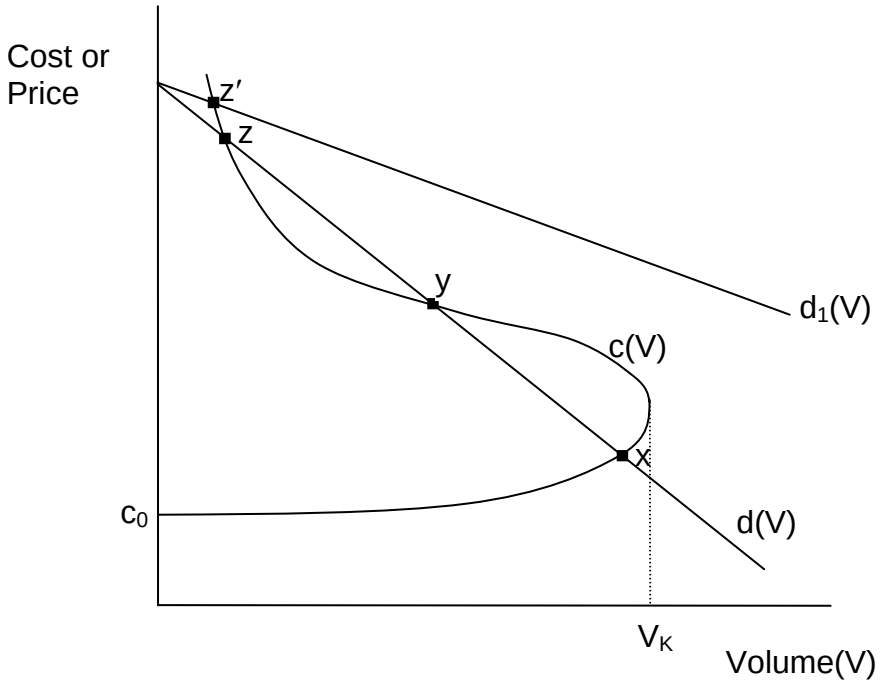


Figure 3.8 Stability of stationary-state equilibria and the stationary-state average cost function

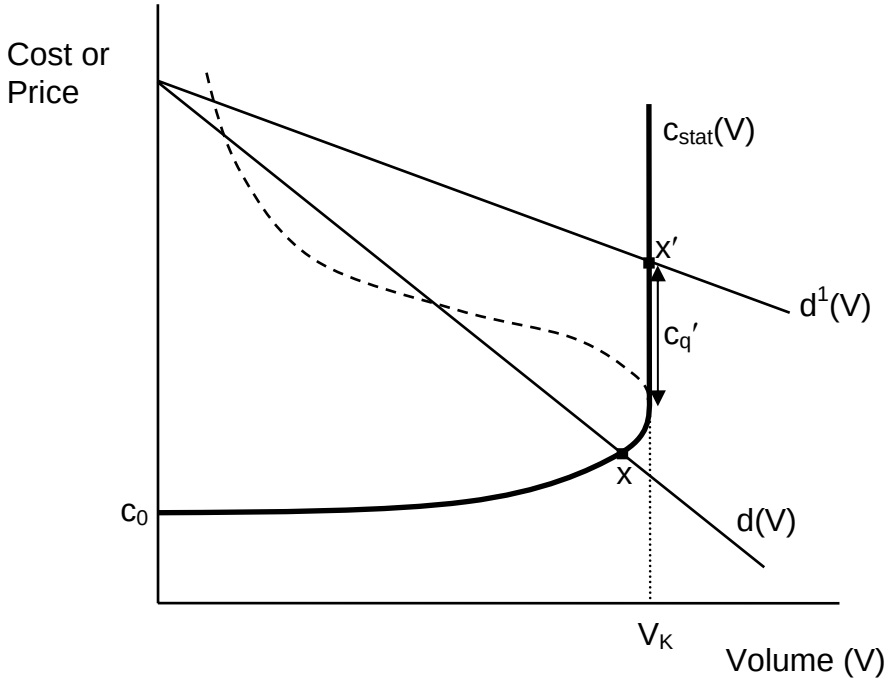


Figure 3.9 Short-run variable cost in stationary-state model (c_{stat}) and two time-averaged models: piecewise linear (c_{PL}) and Akçelik (c_{AK})

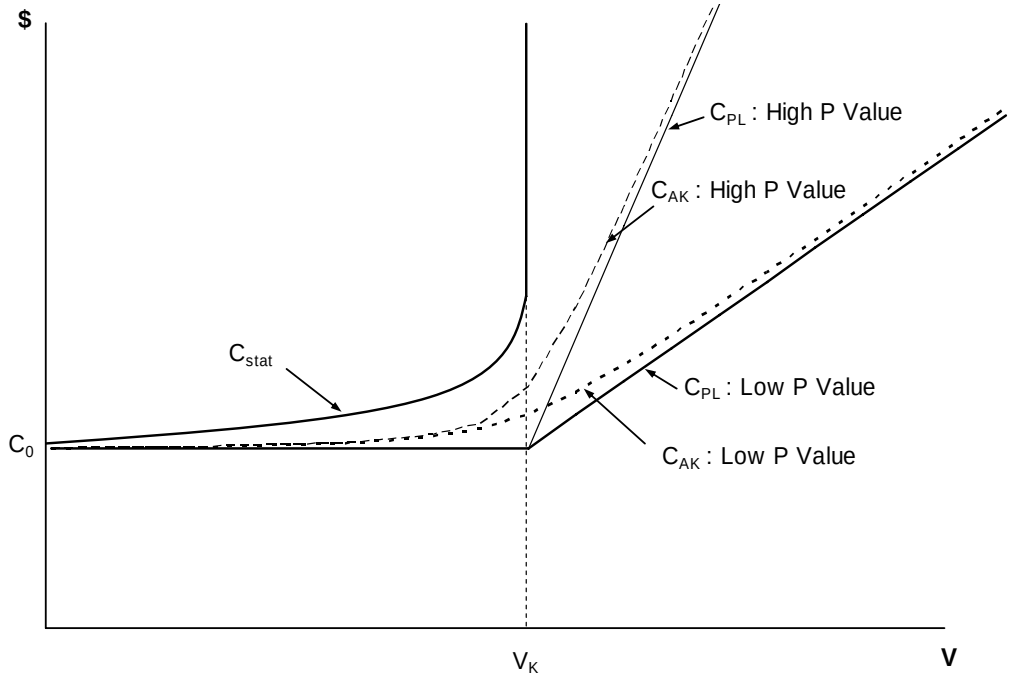
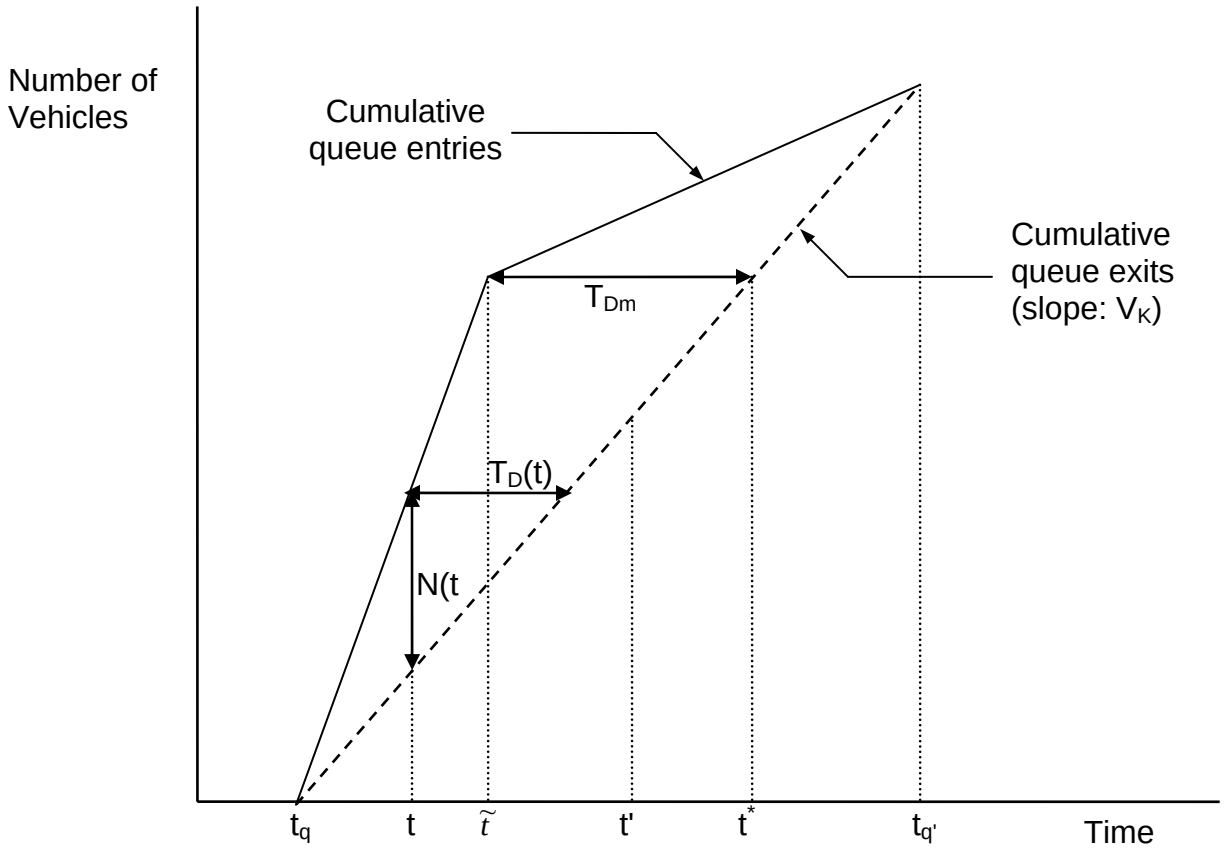


Figure 3.10 Dynamic queuing equilibrium



Note: Adapted with permission from Arnott *et al.* (1990b, p. 117), copyright 1990, Academic Press, Inc.

Figure 3.11 Equilibrium average costs of time delay, schedule delay, and their sum, under linear schedule delay cost functions and with a dispersion of desired queue-exit times.

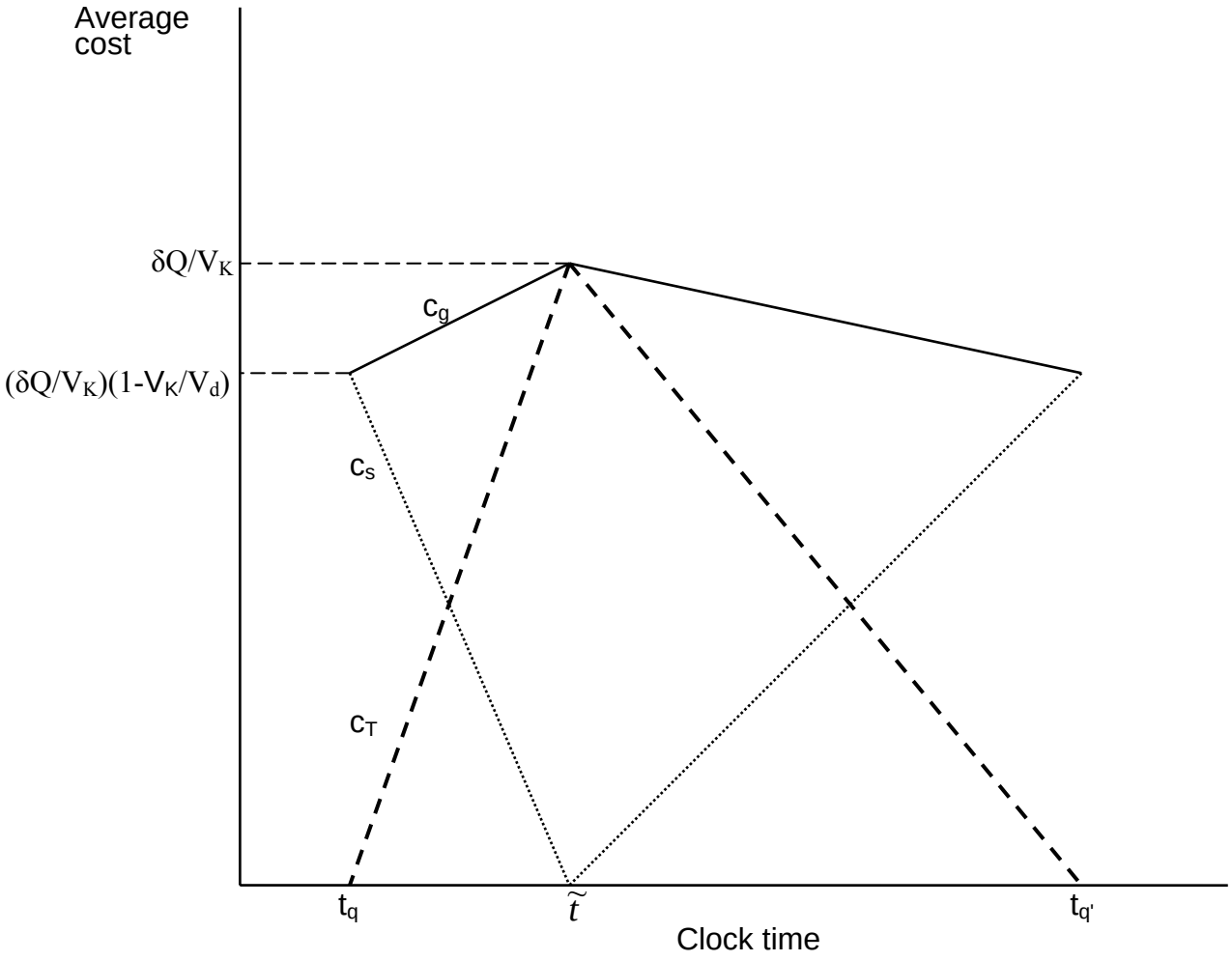


Figure 3.12 Stochastic road capacity and information provision: welfare gains (light shading) and welfare losses (dark shading)

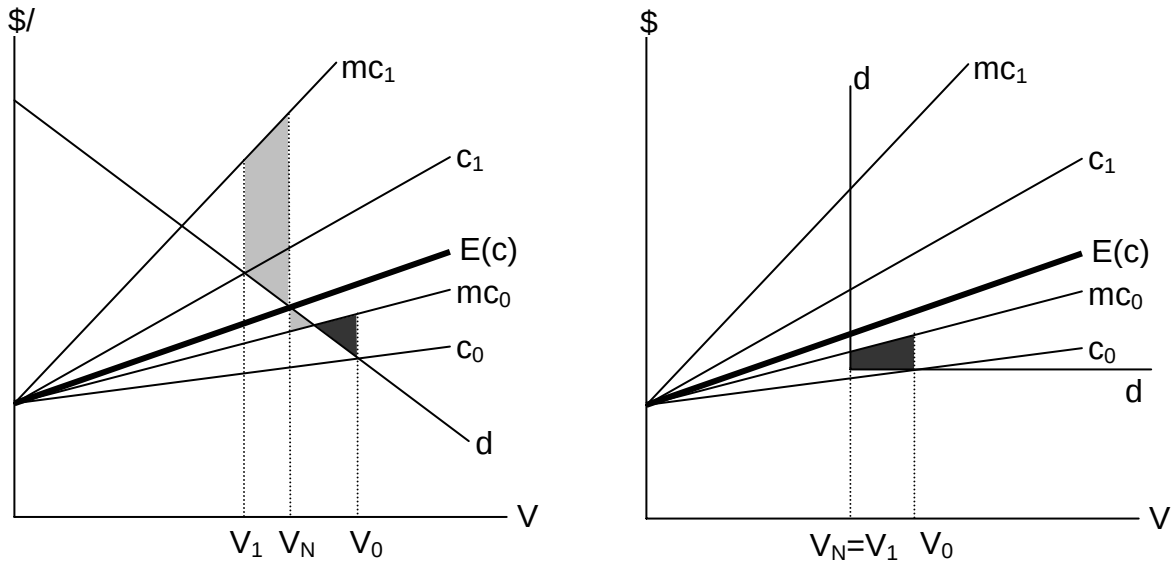


Figure 3.13 Results of Intermodal Cost Comparisons. Adapted with permission from Meyer, Kain, and Wohl (1965, p. 300), copyright RAND Corporation 1965.

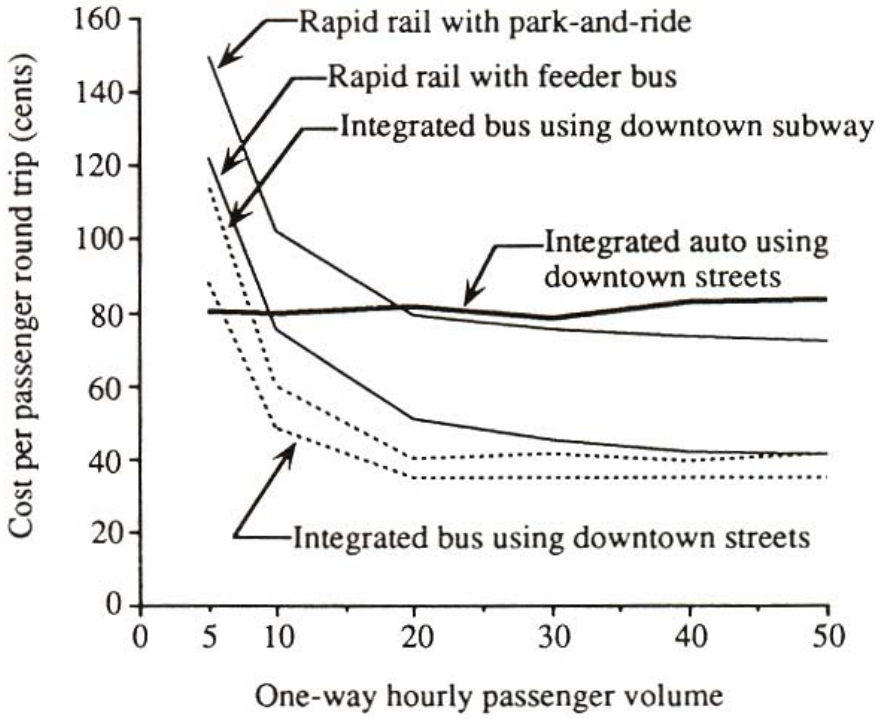


Table 3.1
Accounting Cost Functions for Public Transit: Incremental Costs

	Rapid rail		*Light rail*	*Bus*	
	Boyd[a]	Allport[b]	Allport	Allport	Boyd
Capital Cost:[c]					
Per Route-Mile ($M/yr)	5.95	3.48	0.70	NA	0.90[d]
Per Peak Vehicle ($K/yr)	94.3	57.3	76.0	14.5	25.7
Operating Cost:					
Per Route-Mile ($M/year)	0	0.834	0.209	0.008	0
Per Peak Vehicle ($K/yr)	0	57.9	41.0	26.5	0
Per Convoy-Hour[e] ($)	7.18	37.69	49.44	47.44	27.25
Per Vehicle-Mile[e] ($)	7.39	2.67	2.83	1.26	1.20

Notes:
All figures are in 2003 US dollars, updated using the transportation component of the consumer price index for all urban consumers.

NA means the costs in this category was excluded by the author(s); in contrast to 0, which indicates the costs were included but allocated to other outputs.

[a]Boyd, Asher, and Wetzler (1978, pp. 5-6), and (1973, pp. 29, A-47, and E-1). Figures given by the authors are projected 1980 costs in 1972 U.S. prices.

[b]Allport (1981). Figures given by the author are estimates from accounts of the Rotterdam system in 1978 but adjusted to British conditions in 1980; we have converted to U.S. dollars at the average 1978-80 exchange rate of £1=$2.12, then updated as noted above. Distances are converted using 1 km = 0.6214 mile. Costs per station or stop are converted to costs per route-mile using the average spacings given for Rotterdam (Allport, p. 633).

[c]Annualized capital cost of way, structures, and rolling stock. Allport computes them using an interest rate of 5% per year, and appropriate lifetimes; for rapid rail we average his figures for underground and elevated systems and for different vehicle specifications, while for light rail we use his "high demand" figures, which apply to peak-direction peak-hour passenger demand volumes "considerably higher" than Rotterdam's 450-1,010 (p. 633). Boyd et al. give capital outlay; we annualize it using their assumptions of: 5% interest; 30-year lifetime for way, structures, and rail cars (hence capital recovery factor is 0.0651); 12-year lifetime for bus (capital recovery factor 0.1128). Also, we add a 20% "spare ratio" to their vehicle costs to account for vehicles not in service (this ratio was used by the U.S. Urban Mass Transportation Administration for its funding formulae: Metro Magazine, July/August 1990, p. 16).

[d]Cost of a two-lane exclusive busway, using land cost in 1972 prices of $0.68M/route-mile (Boyd et al., 1978, p. 6) plus construction cost of $1.40M/lane-mile (Boyd et al., 1973, p. 29), both annualized as in the previous note.

[e]A "vehicle" is one rail car or bus. A "convoy" consists of one train (rapid rail), one light-rail vehicle, or one bus. The trains considered by Allport are 2-6 vehicles (p. 633), while those considered by Boyd et al. are 2-10 vehicles (1973, pp. E-2 and E-3); we assume a 4-car train.

Table 3.2
Accounting Cost Model for Cairo, Egypt

	Minibus	Regular Bus	Tram
Unit Costs:[a]			
per peak convoy[b] per year (*PV*)	196	557	1782
per convoy-hour (*VH*)	3.4	13.8	44.6
per convoy-mi (*VM*)	0.35	0.48	3.57
Unit costs divided by capacity *n* (for seated plus standing people):			
per person per year ($PV \cdot n$)	7.84	16.88	23.45
per person-hour ($VH \cdot n$)	0.14	0.42	0.59
per person-mi ($VM \cdot n$)	0.0142	0.0146	0.0470
Percentage of category in total cost:			
$c_2 \cdot PV$	66.0	65.2	61.2
$c_3 \cdot VH$	16.5	24.8	23.4
$c_4 \cdot VM$	17.5	10.0	15.4
Employees per convoy	7.5	16.7	51.9

Source: Abbas and Abd-Allah (1999), Tables 1, 4
Notes:
[a] Monetary units are Egyptian pounds (EGP) for fiscal year1996-97. The exchange rate was 3.4 EGP = US$1.

[b] A convoy means one or more vehicles traveling together. It consists of one minibus, one regular bus, or two tram cars.

Table 3.3
Some Typical Short-Run Costs of Automobile Travel:
US Urban Commuters

Type of cost	Private (Average)[a]	Social Average	Social Marginal
Variable costs			
Costs borne mainly by highway users in aggregate			
(1) Operating & maintenance	0.114	0.114	0.114
(2) Vehicle capital	0.162	0.162	0.162
(3) Travel time	0.286	0.286	0.367
(4) Schedule delay & unreliability	0.088	0.088	0.163
Costs borne substantially by non-users			
(5) Accidents	0.108	0.130	0.165
(6) Government services	0.005	0.018	0.018
(7) Environmental externalities	0	0.017	0.017
Short-run variable costs	0.763	0.815	1.006
Fixed costs			
(8) Roadway	0.017	0.067	
(9) Parking	0.006	0.252	
Short-run fixed costs	0.023	0.319	
Total costs	0.786	1.134	1.006

Notes: All costs in US$ per vehicle-mile at 2003 prices.
[a] If increased vehicle travel requires a proportionate expansion of the car fleet, then private marginal cost is approximately the same as private average cost. In the opposite extreme, where increased travel occurs solely in the form of more miles per vehicle, then the following items should be considered as fixed cost (hence not part of private marginal cost): 62 percent of (2), and all of (6) and (9). We arbitrarily allocate user fees among private cost categories as follows: vehicle and license fees count toward government services, and fuel taxes toward roadway capital; hence item (8) is actually part of variable cost for purposes of private cost.

Table 3.4
Components of Social Average Cost of Accidents

By Type of Cost		By Type of Accident	
Type	Cost ($/veh-mi)	Type	Cost ($/veh-mi)
WTP of death, injury	0.095	Fatality	0.070
Productivity	0.013	Disabling injury	0.023
Medical expenses	0.008	Other injury	0.032
Property damage	0.007	Prop. Damage only	0.004
Legal, police, fire	0.003	Unknown	0.002
Insurance admin	0.003		
Traffic delay	0.002		
TOTAL	0.130	TOTAL	0.130

Source: computed from Parry (2004), Tables 1, 2.
Notes: WTP = willingness to pay (for avoidance). All costs are for US, 1998-2000, stated in 2003 prices. Price levels are updated by multiplying the 1998-2000 costs by 1.119, the average between the growth factors of hourly earnings and of the Consume Price Index for all urban consumers, all items (Council of Economic Advisors, *Annual Report*, 2005, Tables B-47, B-60).

Table 3.5
Construction Costs of North American of Transit Systems

	Construction Cost (2003 US dollars)[a]	
	per route mile ($ millions)	per daily trip[b] ($ thousands)
Heavy rail (average of 4 systems)	177.4	30.51
Light rail (average of 4 systems)	55.1	29.57
Exclusive busway:		
Ottawa	46.8	3.00
Pittsburgh	24.7	5.66
Shared carpool lane:		
El Monte Busway, Los Angeles	14.9	3.79
Shirley Hwy, northern Virginia	15.8	2.98
I-66, northern Virginia	27.8	8.55
Houston transitways (average)[c]	6.6	5.71

Source: Kain (1999), Table 11-3; original sources are Kain *et al.* (1992) and Pickrell (1989).

[a] Updated from 1989 to 2002 prices using the ENR (formerly Engineering News-Record) construction cost index (rose 41.7%); and from 2002 to 2003 using the U.S. Census Bureau index for houses under construction (excluding land) (rose 4.9%). The figures are from US Census Bureau (1992, Table 1203) and (2004, Table 921).

[b] Includes trips in carpools for shared carpool lane. Cost per daily trip by transit is two to nine times greater.

[c] Average of four new "transitways" operational in 1989: Katy, Gulf, North, and Northwest.

4. PRICING

Having described demand and cost structures, we can now ask what happens when the two operate simultaneously: that is, when economic actors constrained by these two structures interact in markets. Our analysis of short-run equilibria in the previous chapter provided some first insights. But to obtain a more complete picture, we must also specify how suppliers take into account demand and cost structures in formulating pricing, investment, or other policies.

There are many ways to do this, each leading to an equilibrium analysis that answers a different question. For example, if we specify that firms maximize profits, we determine equilibrium in an unregulated environment. If we specify that highway capacity is fixed and there is no charge for using it, as in Chapter 3, we determine road use in short-run unpriced equilibrium. If we specify that prices are set to maximize some social objective function, we determine a short-run optimum. These and other possibilities are considered in this and the following two chapters.

Our discussion is organized around policy questions. This chapter describes first-best and second-best optimal short-run pricing, aimed at maximizing social welfare. Chapter 5 discusses investments and how their financing relates to revenues from pricing, including a brief discussion of cost-benefit analysis. Chapter 6 considers institutional arrangements and how they facilitate or hinder the achievement of efficient pricing and investment. These institutions include public or private ownership of highways, public or private provision of public transit, regulation, and freedom of entry by firms into established markets.

The present chapter exemplifies the ramifications of a standard argument in economics: that efficiency is achieved in many situations by marginal-cost pricing. This means that each economic agent, in deciding whether to undertake an activity, faces a perceived price for doing so equal to that activity's social marginal cost. Landmark works applying this principle to transportation include Dupuit (1844, 1849), Hotelling (1938), and Walters (1968). Doing so in practice, however, is often impossible and so we must also consider various constraints on what prices can be charged. We then enter the world of "second-best" as exposited for general situations by Lipsey and Lancaster (1956), Baumol and Bradford (1970), and others. We will see that second-best pricing is the rule rather than the exception in applied transport pricing.

Section 4.1 sets the scene by discussing first-best pricing in the context of congested highway traffic. Section 4.2 then moves on to consider second-best road pricing with such examples as the inability to price all links on a network, the inability to distinguish between classes of users, the inability to vary tolls continuously over time, imperfect information by users, and distortions outside the transport market. Section 4.3 discusses some practical applications of congestion pricing, all of which involve second-best pricing. Section 4.4 considers a very common applied pricing problem: parking. Section 4.5 discusses the pricing of public transport.

4.1 First-Best Congestion Pricing of Highways

Economists have long recognized that the principle of marginal-cost pricing applies to peak-load problems in general (Bye, 1926; Boiteux, 1949; Steiner, 1957) and to roads in particular (Pigou, 1920; Knight, 1924). These authors' basic concepts have been elaborated and extended by many including Walters (1961), Strotz (1965), Mohring and Harwitz (1962), and Vickrey (1963, 1968b, 1969). The resulting models have been applied empirically to cities around the world.[1]

This section presents the basic economic motivation for congestion pricing. To that end, we abstract from various complications and consider a world in which traffic congestion is the sole distortion in the economy and fully flexible road pricing is possible. Thus, for example, we assume for now that any other externalities associated with travel are priced, e.g. by a fuel tax; and we exclude so-called "external benefits" that may occur if the marginal social benefit of a trip exceeds the benefit to the traveler.

The basic idea is easily stated within the framework of highway costs developed in Chapter 3. Recall that congestion technology and the costs of travel have been placed directly into the average cost function, $c(\cdot)$. All congestion functions considered in Chapter 3 have an important feature in common: short-run average cost increases with the level of road use, be it

[1] A sampling of studies includes Mohring (1965), Smeed (1968), Keeler and Small (1977), Dewees (1979), Gómez-Ibáñez and Fauth (1980), Anderson and Mohring (1997), Nguyen (1999), May and Milne (2000), De Borger and Proost (2001), Li (2002), Niskanen and Nash (2004), Santos (2004), De Palma, Lindsey, and Proost (2006), and Eliasson and Mattsson (2006). Implementation issues have also been studied extensively: for example in Britain (U.K. Ministry of Transport, 1964); Singapore (Watson and Holland, 1978); Hong Kong (Dawson and Brown, 1985); the United States (National Research Council, 1994); and the European Union (Niskanen and Nash, 2004).

expressed in traffic flow (V) as in static models, or in the total number of travelers over the peak (Q) as in dynamic models. This implies that short-run marginal cost exceeds short-run average cost. Intuitively, this is because short-run marginal social cost *mc* includes not only the cost incurred by the traveler herself but also the additional cost she imposes on all other travelers by adding to the congestion they encounter.[2] This additional cost is known as the *marginal external congestion cost*, here denoted *mecc*.

An efficient level of road use is obtained when each trip that is made provides benefits as least as great as its social cost, *mc*, and when no trip meeting this condition is suppressed. To obtain this situation through pricing, each traveler should face the marginal social cost of her trip. This requires a charge equal to the difference between the marginal cost and the cost already borne by the traveler, which is short-run average variable cost, c. This charge, known as the optimal *congestion fee* or *congestion toll*, is therefore $\tau = mc - c = mecc$.

These arguments can be formalized by means of a planning problem that determines a *Pareto optimal* distribution of traffic, defined as one that maximizes any one person's utility while holding all others' utilities constant and meeting any aggregate resource and technological constraints. Equivalently, we find the allocation of road space to users that maximizes *net welfare*, defined as the difference between aggregate consumer benefit and total cost. For first-best pricing, total cost embodies all the resource and technological constraints; when we consider second-best pricing, we will add various feasibility constraints when maximizing net welfare. Here we consider first-best pricing with static congestion (Section 4.1.1), then with dynamic congestion (Section 4.1.2).

4.1.1 Static Congestion

Chapter 3 distinguished between two types of static models of traffic congestion: models of stationary-state congestion and models of time-averaged congestion. The analytical derivation of optimal road prices does not differ fundamentally between these models, provided the inflow period P is kept fixed in time-averaged models. We will therefore treat them together, and refer to them as "static" models.

---

[2] Formally, total cost is $V \cdot c(V)$, so marginal cost is $mc(V) = c(V) + V \cdot c'(V)$ (by the product rule of differentiation). With rising average cost, i.e. $c'(V)>0$, we see that $mc(V)$ exceeds $c(V)$ for any positive V.

Single Road, Single Time Period

Consider first the case of a single road and a single time period. As before, let $d(V)$ denote the inverse demand function and $c(V)$ the average cost function. In equilibrium, individuals equate their marginal willingness-to-pay $d(V)$ to generalized price, p, which is defined as average cost $c(V)$ plus toll τ.

$$d(V) = p \equiv c(V) + \tau \,. \tag{4.1}$$

The aggregate benefit from road use, B, is the value of travel to users, as measured by the area under the inverse demand curve up to the equilibrium travel flow:

$$B = \int_0^V d(v)\,\mathrm{d}v \,. \tag{4.2}$$

Total cost, holding capacity fixed in the short run, is:

$$C = V \cdot c(V) + \rho K \,, \tag{4.3}$$

where, as before, ρK is the annualized cost of capital expenditures K.

An appropriate measure of aggregate welfare in this setting is social surplus W, defined as total benefit minus total cost. Maximizing $W = B - C$ with respect to V yields the necessary first-order condition:

$$d(V) - c(V) - V \cdot \frac{\partial c}{\partial V} = 0 \quad \Rightarrow \quad d(V) = mc(V) \,. \tag{4.4}$$

This implies, using (4.), that the optimal price is:

$$p = mc(V) = c(V) + V \cdot \frac{\partial c}{\partial V} \,, \tag{4.5}$$

thus yielding the usual marginal-cost pricing rule. Equivalently, the optimal toll is:

$$\tau = mc(V) - c(V) = V \cdot \frac{\partial c}{\partial V} = mecc \,. \tag{4.6}$$

The optimal congestion toll is thus the difference between short-run marginal and average cost. It is often referred to as a "Pigouvian toll", named after Arthur Pigou (1920).

Figure 4.1 provides a graphical illustration. The left panel shows the conventional diagram of optimal congestion pricing. The unpriced equilibrium occurs at the intersection of $d(V)$ and $c(V)$; it involves traffic flow V^0 and cost c^0. The optimal flow V^1 occurs at the intersection of $d(V)$ and $mc(V)$, according to (4.4); it can be achieved by imposing the optimal fee τ shown in the diagram. The generalized cost c falls from c^0 to c^1, but the generalized price rises from $p^0 = c^0$ to $p^1 = c^1 + \tau$. The quantity $(c^0 - c_0) \cdot V^0$ can be interpreted as the total cost of congestion in the unpriced equilibrium; but the total cost of *inefficient* congestion, $(c^0 - c^1) \cdot V^0$, is smaller because some congestion is optimal in this example.

Figure 4.1

In many cases the Pigovian toll takes an intuitive form mathematically. Using the BPR congestion function of equation (3.9), average cost is given by (3.23), which we can write as:

$$c = c_0 + \alpha T_f a \cdot \left(\frac{V}{V_k} \right)^b , \tag{4.7}$$

where again α is the value of time, $c_0 \equiv c_{00} + \alpha T_f$ includes the value of free-flow travel time T_f, and a and b are parameters of the speed-flow curve. Then:

$$mc \equiv \frac{\partial (V \cdot c)}{\partial V} = c_0 + \alpha T_f a \cdot (b+1) \cdot \left(\frac{V}{V_k} \right)^b , \tag{4.8}$$

and:

$$\tau = mecc \equiv mc - c = \alpha T_f ab \cdot \left(\frac{V}{V_k} \right)^b . \tag{4.9}$$

In this case, the optimal congestion fee is just b times the average congestion cost in the optimum, a result derived by Vickrey (1965). To determine τ numerically, one must still solve (4.6) simultaneously with the demand function (4.), which can be written as $\tau = d(V) - c$, yielding

the condition $d(V)=mc$, as also can be seen in the figure. Depending on the form of $d(V)$, this may or may not have an analytical solution.

Using instead a piecewise-linear cost function based on congestion function (3.10), a corner solution is possible as illustrated in the right panel of Figure 4.1. The optimal volume is where the demand curve crosses the *mc* curve, namely V_K. The optimal congestion fee is the fee that keeps demand at that level; in this case, it completely eliminates congestion. Because the cost function is kinked there, marginal cost *mc* is formally undefined; but conceptually it can be interpreted as the value of the marginal trip displaced by the last traveler using the available capacity. For other demand curves, a corner solution need not obtain. If the demand curve were lower, so as to intersect the average cost function on its flat segment, then the optimal toll would be zero and the optimum would coincide with the unpriced equilibrium. If instead the demand curve were higher, so as to intersect $mc(V)$ on its rising segment, the optimal volume would entail some congestion, just as in the left panel. In this case, the relevant part of the cost function is:

$$c = c_0 + \tfrac{1}{2}\alpha P \cdot \left(\frac{V}{V_k} - 1 \right), \tag{4.10}$$

where *P* denotes the exogenous inflow period; the optimal toll is then:

$$\tau = mecc \equiv mc - c = \frac{\partial (V \cdot c)}{\partial V} - c = \tfrac{1}{2}\alpha P \cdot \left(\frac{V}{V_k} \right). \tag{4.11}$$

Again, to determine a numerical value for the toll, this equation must be solved simultaneously with condition (4.) for user equilibrium.

A cost function that becomes vertical, like c_{stat} in Figure 3.8, also implies a section where *mc* is vertical. If the marginal-cost toll occurs on that section, it has an interpretation like that just described for the piecewise-linear congestion function

Single Road, Multiple Time Periods

The above model is easily extended to multiple periods of stationary-state congestion. Consider, as before, a highway of capacity V_K serving *H* distinct daily time periods, each of exogenous duration q_h with endogenous flow V_h under stationary traffic conditions. Then the $q_h V_h$ users of

the highway during period h incur short-run total variable cost $q_h V_h c_h$; the short-run marginal cost of adding another user is therefore:

$$mc_h \equiv \frac{\partial(q_h V_h c_h)}{\partial(q_h V_h)} = \frac{\partial(V_h \cdot c_h)}{\partial V_h} = c_h + V_h \cdot \frac{\partial c_h}{\partial V_h}. \tag{4.12}$$

With independent inverse demand functions $d_h(V_h)$ applying for each period, social surplus W is:

$$W = \sum_h q_h \int_0^{V_h} d_h(v)\mathrm{d}v - \sum_h q_h V_h \cdot c_h(V_h) - \rho K . \tag{4.13}$$

The first-order conditions to maximize W produce optimal tolls just like (4.6):

$$\tau_h = mc_h(V_h) - c_h(V_h) = V_h \cdot \frac{\partial c_h}{\partial V_h}. \tag{4.14}$$

This toll is typically higher in those periods when equilibrium demand V_h is higher, and is guaranteed to be so if $c_h''(V_h) \geq 0$ as is usual; hence it is an example of peak-load pricing.

Network Optimum, Single Time Period

A second extension of interest concerns optimal congestion pricing on a network. Equation (3.40) in Section 3.4.4 presented the *user equilibrium* conditions for an unpriced network, also referred to as *Wardrop's first principle* (Wardrop, 1952). Recall that subscripts l denote links, r and ρ routes, and m markets or OD-pairs. When link-based tolls τ_l are in place, these equilibrium conditions continue to hold with $c_l(V_l)$ replaced by $c_l(V_l)+\tau_l$:

$$\forall \delta_{rm} = 1: \begin{cases} \sum_{l=1}^{L} \delta_{lr} \cdot \left(c_l(V_l) + \tau_l\right) - d_m(V_m) \geq 0 \\ V_r \geq 0 \\ V_r \cdot \left[\sum_{l=1}^{L} \delta_{lr} \cdot \left(c_l(V_l) + \tau_l\right) - d_m(V_m)\right] = 0 \end{cases} \tag{4.15}$$

where again $\delta_{rm}=1$ for any route r that serves market m, $\delta_{lr}=1$ for any link l that is part of route r, and

$$V_l = \sum_{\rho=1}^{R} \delta_{l\rho} \cdot V_\rho, \quad V_m = \sum_{\rho=1}^{R} \delta_{\rho m} \cdot V_\rho \ .$$

The interpretation is that in a user equilibrium, all used routes for an OD pair should have equal generalized prices, all equal to marginal benefits $d_m(V_m)$ for that OD-pair; and that there are no unused routes with lower generalized prices.

The optimal flow pattern, often referred to as the *system optimum* in the engineering literature, can be found by maximizing social surplus (W), now defined as the benefits in (4.2) summed over OD-pairs minus the costs in (4.3) summed over links. Including non-negativity constraints for route flows, the optimization problem becomes:

$$\begin{aligned} &\underset{V_{r=1\ldots R}}{\text{Max}}\, W = \sum_m \int_0^{V_m} d_m(v)\mathrm{d}v - \sum_l V_l \cdot c_l(V_l; V_{K,l}) - \sum_l \rho \cdot K_l(V_{K,l}) \\ &\text{with}: \quad V_l = \sum_{r=1}^{R} \delta_{lr} \cdot V_r \quad \text{and} \quad V_m = \sum_{r=1}^{R} \delta_{rm} \cdot V_r \qquad . \\ &\text{s.t.}: V_r \geq 0 \quad \forall r \end{aligned} \tag{4.16}$$

Taking derivatives with respect to all possible route flows V_r, we obtain Kuhn-Tucker first-order conditions that are identical to (4.15) except that τ_l is replaced by $V_l \cdot (\partial c_l / \partial V_l)$. These conditions express *Wardrop's second principle*: in a system optimum, all used routes for an OD pair have identical marginal costs equal to the marginal benefits for that OD-pair, and there are no unused routes with marginal costs lower than this. Comparing to the equilibrium conditions (4.15), it's easy to see that the following pricing rule guarantees satisfaction of the first-order conditions:

$$\tau_l = V_l \cdot \frac{\partial c_l}{\partial V_l} \quad \forall l \, . \tag{4.17}$$

The single-road toll in (4.6) is therefore just a special case of the optimal link tolls for a full network, shown in (4.17). Link-based marginal cost pricing throughout the network ensures that on all "active" routes (*i.e.*, with $V_r>0$), users will enter up to the point where marginal benefits are equal to marginal cost, while a route remains "passive" (*i.e.*, $V_r=0$) when marginal benefits on the OD-pair it serves are below the marginal cost of using the route.

The tolls in (4.17) need not be the unique set of tolls achieving the optimum. The comparison between the equilibrium conditions (4.15) and the optimality conditions only tells us that the sum of link-tolls over each route should be equal to the sum of marginal external costs

on all the route's links. This may give some freedom in setting link-based tolls. The simplest example is when two serial links are always used together if used at all: what affects the flow pattern is not the individual link tolls but their sum, so all combinations of tolls with a given sum are equivalent. A further degree of freedom arises when demand is perfectly inelastic for all OD-pairs: adding a constant to all the link-based tolls of (4.17) would then leave route choices unaltered, allowing one to choose them to meet a secondary objective (Hearn and Ramana, 1998).

The practical challenge of setting optimal link-based tolls are daunting given that neither the demand functions nor the link-specific speed-flow curves can be known precisely. This has raised interest in trial-and-error approaches, in which tolls are set and then adapted based on observed results. Li (2002) and Yang, Meng, and Lee (2004) develop strategies for doing this.

Heterogeneity of Users

When users are heterogeneous, should tolls be differentiated across users? If so, the practical problems of toll collection are exacerbated; if not, we say the tolls can be *anonymous*.

The principle that toll equal marginal external congestion cost (*mecc*) does not change. The value of *mecc* for different users may differ, however, for two possible reasons. First, their vehicles may contribute differently to congestion, for example truck and cars; in that case, optimal tolls are not anonymous. Second, they may self-sort into different parts of the network with different values of *mecc* in equilibrium. In that case, the toll can still be anonymous, but must be differentiated across links as in (4.17). For example, Verhoef and Small (2004) consider differentiated first-best tolls for a simple parallel-route network with dispersion in values of time. The route with the higher toll is faster and therefore attracts drivers with higher values of time. The higher optimal toll on this route reflects a higher *mecc*, despite the fact that congestion is lower than on the other route; but the *mecc* imposed by a given user is not high because of that user's own value of time, but rather because of the value of time of her co-travelers, on whom she imposes a congestion externality.

Distributional Impacts and Acceptability of Road Pricing

Although imposition of the optimal toll generates a net welfare gain, the social and political acceptability of road pricing has proven to be very limited. The left panel in Figure 4.1 shows an

important reason why. Users between 0 and V^1 on the horizontal axis experience an increase in generalized price from p^0 to p^1 and therefore are worse off unless they receive benefits from the use of toll revenues. The users between V^1 and V^0 have shifted from the untolled road to some less-preferred alternative (such as public transport or not traveling at all), and are therefore also worse off.

It is interesting to compare these two groups in terms of how much worse off they are. The shifters' losses range from zero for the driver at V^0 to $p^1–p^0$ for the driver at V^1, with the others somewhere between. Thus their average loss in surplus is smaller than that for drivers who remain using the road (for example, with a linear demand curve, it is just half). The intuition is that, by changing behavior, drivers can avoid incurring the full loss $p^1–p^0$, and will only choose to do so when changing is more attractive than staying on the road.

Ignoring revenue allocation, then, all the initial travelers therefore lose from the policy (in the absence of heterogeneity). The only gain is to the public sector, in the form of toll revenues. Whether this gain is recognized by citizens in the political process depends on how revenues are used and on whether the authorities are able to effectively and credibly communicate these uses to the public. One consequence is that researchers have proposed various allocation schemes have been proposed in attempts to leave major user groups better off (Goodwin, 1989; Small, 1992b).

Another consequence is that revenue allocation has been identified as a key determinant of the political acceptability of congestion pricing. In a Dutch survey by Verhoef, Nijkamp, and Rietveld (1997), road users expressed the following preference (in decreasing order) for the use of toll revenues: investment in new roads, reduction in vehicle ownership taxes, reduction in fuel taxes, investment in public transport, subsidies for public transport, investment in carpool facilities, general tax cuts, and expansion of other public expenditures. Evidence suggests that the British put greater preference on public transport, while residents of the US put more on road construction or tax reduction. Interestingly, the most far-reaching of all congestion pricing schemes implemented in practice, namely in Central London starting in 2003, included a very explicit and well-publicized component of investing revenues in London's transit system. In North America and Norway, by contrast, nearly all pricing schemes that exist or are close to implementation have a direct connection to infrastructure finance.

When there is heterogeneity of travelers, some initial road users may be better off after imposition of optimal pricing, even before accounting for the use of revenues. This can happen because for an individual whose value of travel time is sufficiently high, the value of travel-time gains may exceed the toll. Aggregate demand by such users may then increase due to the implementation of congestion pricing. Aggregate demand added over *all* users groups would of course still decrease, for otherwise the travel-time gains would not happen.

Since high-value-of-time travelers benefit more, or lose less, from road pricing, it is likely to be regressive (not counting the use of toll revenues) because the value of time is positively correlated with income (Layard, 1977; Niskanen, 1987). This tendency is mitigated by the tendency of overall travel to increase with income (Foster, 1974). Also with toll cordons, higher income groups face more tolls than others if they live more predominantly outside the cordon. Of course, the ultimate net distributional effect of congestion pricing depends also on how the revenues are used and upon how the burdens are shifted through markets for land, labor, and commercial space.

Finally, public attitudes towards road pricing may differ before and after implementation. Tretvik (2003) reports how the local support for Norwegian toll rings, defined as the percentage of respondents who judge the scheme positively, has increased dramatically after implementation, from 19% to 58% in Bergen, form 30% to 41% for Oslo, and from 9% to 47% in Trondheim. Transport for London (2004) reports an increase from 39% support (average over 3 waves) to 54% (average over 4 waves). Such increases may reflect initial disbelief in the scheme's effectiveness.

4.1.2 Dynamic Congestion

We now discuss congestion pricing from a dynamic perspective, focusing on dynamic equilibrium models that endogenize departure-time decisions. We begin with the "basic bottleneck model" of Section 3.4.3, which allows for a particularly transparent analytical treatment and is currently the most widely used conceptual dynamic model of congestion pricing. We then turn to two of the extensions of it considered before: heterogeneous uses and networks. Finally, we describe pricing results from other types of dynamic congestion models.

First-best Pricing in the Basic Bottleneck Model

Recall that the "basic bottleneck model" considers a single "pure bottleneck," i.e. one for which there are no delays if inflow is below capacity V_K, and for which the rate of queue exits is equal to capacity (when a queue exists). The basic bottleneck model further simplifies by setting free-flow travel time equal to zero and by assuming that cost parameters are identical for all users: namely, the value of travel time α, the shadow prices of early and late arrivals β and γ, and the desired arrival time t^*. Total demand for passages Q is inelastic.

For ease of reference, we summarize the equilibrium conditions for this model, as derived in Section 3.4.3 in the limit where duration q of the desired queue-exit-time interval approaches zero while the total number of travelers, qV_d, remains finite at value Q. We add superscripts 0 to denote the unpriced equilibrium. The peak starts and ends at times:

$$t_q^0 = t^* - \frac{\gamma}{\beta+\gamma}\cdot\frac{Q}{V_K} \tag{4.18}$$

$$t_{q'}^0 = t^* + \frac{\beta}{\beta+\gamma}\cdot\frac{Q}{V_K}. \tag{4.19}$$

The queue-entry rates for early and late arrivals are given by (3.25), which ensures that average congestion cost $\bar{c}_g$ remain constant over all queue-entry times. Its equilibrium value $\bar{c}_g^0$ and its two components, average travel delay cost $\bar{c}_T^0$ and average schedule delay cost $\bar{c}_S^0$, are:

$$\bar{c}_g^0 \equiv p^0 = \delta\cdot\frac{Q}{V_K}\ ; \qquad \bar{c}_T^0 = \bar{c}_S^0 = \tfrac{1}{2}\cdot\bar{c}_g^0. \tag{4.20}$$

where $\delta=\beta\gamma/(\beta+\gamma)$. Total equilibrium costs are consequently:

$$\bar{C}_g^0 = \delta\cdot\frac{Q^2}{V_K}, \tag{4.21}$$

implying marginal cost equal to:

$$\overline{mc}_g^0 = 2\delta\cdot\frac{Q}{V_K} = 2\bar{c}_g^0. \tag{4.22}$$

Figure 4.2 depicts this unpriced equilibrium graphically, by showing schedule-delay cost $c_S(t')$ and travel-delay cost $c_T^0(t')$ as functions of arrival time t'. Because the queue-exit rate V_b is

constant over time, the time-averaged value of schedule-delay cost is given by the two triangular areas under $c_S(t')$ (between t_q and $t_{q'}$), divided by $(t_q - t_{q'})$; while for travel-delay cost it is the inverted triangle between $c_S(t')$ and the horizontal line at p^0, again divided by $(t_q - t_{q'})$. The equality between these two averages, stated in (4.20), is therefore visible geometrically from the diagram.

Figure 4.2

Now let's consider what an optimal travel pattern must look like. The optimum can be identified by intuitive reasoning. First, as long as exits occur, the exit rate should not be below V_K, as otherwise the period of exits could be shortened and hence total schedule-delay cost could be reduced without increasing travel delay. Second, no queue should exist in the optimum, as otherwise total travel-delay cost could be reduced without increasing schedule delay. These two observations together mean that the entry and exit rates should both be equal to V_K throughout the peak. Third, the timing of the period of exits should be such that the schedule-delay cost of the first and last drivers are equal, as otherwise total schedule-delay cost could be reduced by shifting the entire pattern of exits over time. This means that exits should occur between the same instants t_q and $t_{q'}$ as in the unpriced equilibrium.

The optimum just described thus involves the same pattern of *exits* from the bottleneck as the unpriced equilibrium, but it has a different pattern of *entries.* It can be decentralized by a triangular toll schedule $\tau(t')$, with two linear segments, that replicates the pattern of travel delay costs in the unpriced equilibrium, $c_T^0(t')$:[3]

$$\tau(t') = \begin{cases} 0 & \text{if } t' < t_q \\ \beta \cdot (t' - t_q) = \delta \cdot \dfrac{Q}{V_K} - \beta \cdot (t^* - t') & \text{if } t_q \le t' < t^* \\ \gamma \cdot (t_{q'} - t') = \delta \cdot \dfrac{Q}{V_K} - \gamma \cdot (t' - t^*) & \text{if } t^* \le t' \le t_{q'} \\ 0 & \text{if } t' > t_{q'} \end{cases} . \tag{4.23}$$

[3] For exit times outside the period $[t_q , t_{q'}]$, the zero tolls shown in (4.23) are more than sufficiently high to support the optimal pattern. Negative tolls would even be allowed so long as $\tau(t') > p^0 - c_S(t')$.

This toll schedule is shown in the diagram as $\tau(t')$. It results in the same constant generalized price $p^1 = p^0$ (where superscript 1 denotes the first-best tolled equilibrium) and the same pattern of schedule-delay cost as in the user equilibrium, but it produces zero travel delay cost. The resulting tolled-equilibrium departure pattern therefore satisfies:

$$V_a^1 = V_K \,. \tag{4.24}$$

From the figure, we find the following price and average costs levels:

$$\begin{aligned} p^1 &= \delta \cdot \frac{Q}{V_K} \\ \bar{c}_g^1 &= \tfrac{1}{2} \cdot \delta \cdot \frac{Q}{V_K}; \quad \bar{c}_T^1 = 0; \quad \bar{c}_S^1 = \bar{c}_g^1 \,. \end{aligned} \tag{4.25}$$

Total cost is therefore half as large as in the unpriced equilibrium:

$$\bar{C}_g^1 = \tfrac{1}{2} \cdot \bar{C}_g^0 = \tfrac{1}{2} \cdot \delta \cdot \frac{Q^2}{V_K} . \tag{4.26}$$

The net welfare gain from optimal pricing is equal to the value of travel time savings, and therefore also to the total toll revenues generated. Neither equality is generally true in the static model of Figure 4.1.

Several points deserve emphasis. First, in contrast to most static models, no travel delays exist in the optimum. Second, the generalized price remains unchanged after imposition of the optimal toll schedule, because tolls exactly replace travel time delays; this suggests that social acceptability should be less of a problem than what is predicted by the static model. Third, exit times and hence arrival times at the destination need not change between the unpriced equilibrium and the optimum, and no alternative to (solo) car use is required for the elimination of queues. In the context of the morning commute, only departure times from home have to be adjusted for queues to disappear. Moreover, if commuters retain their order of departure and their arrival times, everybody (except the very first and last driver) should depart later than in the unpriced equilibrium. This gives room for optimism on the possibility of achieving a significant reduction in queues in reality under optimal pricing. Note that these results are due to the kinked performance function for a pure bottleneck, and will at best only approximately apply for models with different performance functions.

We have assumed thus far that total demand Q over the peak is fixed, which leaves an ambiguity in the optimal toll: any constant could be added to the toll schedule of (4.23) and still support the optimal pattern. What if instead total demand has non-zero elasticity with respect to generalized price? We have seen that toll schedule (4.23) produces the same constant generalized price as the user equilibrium, $p^1=p^0$; therefore it also produces the same quantity demanded Q even if demand has some elasticity. This implies surprisingly that total demand Q^0 in the user equilibrium, which is inefficiently high in that situation, is just right when an optimal toll is in place.[4] One way to think about this puzzling result is that tolling reduces the adverse consequences of adding to total traffic so much that there is no longer a reason to curtail traffic, as there is in the unpriced equilibrium.

More formally, observe from (4.26) that in the optimal pattern, the marginal cost of increasing Q is:

$$\overline{mc}_g^1 = \delta \cdot \frac{Q}{V_K}, \tag{4.27}$$

which is the same as the generalized price p^1. Thus marginal cost equals marginal benefit with the toll schedule (4.23) in place, and no additional constant needs to be added to it even though it has failed to reduce Q from its no-toll value. The toll schedule (4.23) can also be interpreted as the marginal external cost of a traveler arriving at t', defined as the difference between the (time-independent) marginal social cost (4.27) and the (time-dependent) private cost $c_g(t')$.

Yet another way to look at these results is that imposing optimal time-varying tolls causes the average and marginal cost curves to rotate downwards by a factor ½, as seen by comparing (4.21) to (4.26) and (4.22) to (4.27). The cost curves in the static models of Section 4.1.1, in contrast, do not change when pricing is imposed.

Widespread adoption of time-varying fees might alter enough people's behavior to change the institutional and social environment determining certain behavioral parameters such as desired schedules and costs of schedule delay. A fuller description of long-run equilibrium would therefore incorporate the endogeneity of such parameters. There has been little if any research on this possibility.

---

[4] It also implies that (4.26) still holds, even when demand is elastic is zero, with no additional constant.

Heterogeneous Users

The logic of optimal pricing in the basic bottleneck model carries over to more complex settings. Newell (1987) provides a general analysis of equilibrium at a pure bottleneck with heterogeneous commuters, each of whom minimizes a deterministic cost function belonging to a specified parametric family. He provides diagrammatic and algorithmic solutions for determining the equilibrium queuing-delay function $T_D(t)$ (and, from that, the queue-entry times and resulting schedule delays for all commuters), given two cumulative distribution functions: that of desired queue-exit times, $F(t)$, and that of the cost-function parameters. (The desired queue-exit rate $V_d(t)$ in the previous chapter is just $F'(t)$ in this notation.)

Figure 4.3

A few very general features can be derived from minimal assumptions on these functions. Figure 4.3 shows cumulative counts $A(t)$ of queue-entry times, $B(t)$ of queue-exit times, and $F(t)$ of desired queue-exit times. At any given queue-entry time t and desired queue-exit time t_d, queuing delay $T_D(t)$ and schedule delay $S_D(t) \equiv t' - t_d \equiv t + T_D(t) - t_d$ are the horizontal distances shown in the Figure. (S_D is negative at the time t shown, indicating a queue-exit earlier than desired.) Denote the cost of queuing delay by $c_T(T_D)$, the cost of schedule delay by $c_S(S_D)$, and the derivatives of these functions with primes.[5] Each commuter chooses t to minimize $c_T(t)$+ $c_S(t)$, leading to the following first-order condition, which must make the sum constant over the time interval when users with the corresponding particular set of cost parameters depart:

$$c_T' \cdot T_D'(t) + c_S' \cdot \left(1 + T_D'(t)\right) = 0\,, \tag{4.28}$$

or:

$$\frac{c_S'}{c_T'} = -\frac{T_D'(t)}{1 + T_D'(t)}\,, \tag{4.29}$$

[5] This is a slight change of notation from Chapter 3, where c_T and c_S are functions of t. Here c_T' refers to (dc_T/dT_D), so that by the chain rule, the time-derivative of $c_T(T_D(t))$ is $c_T' \cdot T_D'$. Similarly for c_S.

assuming the derivatives exist.[6] Note that the assumed first-in, first-out queue discipline implies that $T_D'(t) > -1$.

Several qualitative results follow. First, the queue grows while early travelers (those who will exit before their desired arrival times) are arriving, and it shrinks while late travelers are arriving.[7] Second, a person exiting the queue exactly at his desired time must incur the maximum travel time incurred by users with the same characteristics.[8] Third, under certain circumstances, Newell obtains for this more general model a key pricing result of the basic bottleneck model: if user have identical values of time, a time-varying toll can be defined which has no allocative effects on number of travelers or their time of passage through the bottleneck. Such a toll collects revenues equal to the entire cost of queuing delay in the unpriced equilibrium, and it leaves each commuter exactly as well off (before redistribution of toll revenues) as in the no-toll equilibrium. This last-mentioned result illustrates a point that emerges from simulation exercises by ADL (1993): the reallocation of departure times may be a greater source of benefit from time-of-day pricing than the reduction in total trips.

Now suppose heterogeneity is more limited: the desired queue-exit time t_d varies across travelers, but the per-unit costs of travel and schedule delay do not. (One examples is our treatment of the bottleneck model in Chapter 3.) Hendrickson and Kocur (1981) provide an elegant and general analysis. Within limits, the heterogeneity in desired queue-exit times affects only schedule-delay costs, not travel-delay costs or the time pattern of queuing in the unpriced equilibrium. This is true so long as the cumulative desired exits, $F(t)$ in Figure 4.3, intersects the cumulative exits $B(t)$ only once. The optimal time-varying toll is then the same as for the basic

---

[6] Because $T_D' = (V - V_K)/V_K$ when there is a queue, equation (4.29) is consistent with the departure rates of (3.25) in the basic bottleneck model, where $c_T' = \alpha$, $c_S' = -\beta$ for early arrivals, and $c_S' = \gamma$ for late arrivals.

[7] Proof: We know that c_T' is positive and that c_S' has the same sign as S_D (the latter because c_S is by definition minimized at $S_D = 0$). For early exits, c_S' is therefore negative; so equilibrium condition (4.29) requires that $T_D'(t) > 0$ (growing queue). It also requires that $c_T' > |c_S'|$ — a consistence condition identical to the previously noted requirement $\beta < \alpha$ in the basic bottleneck model, where $c_T' = \alpha$ and $c_S' = -\beta$. For late exits, (4.29) requires that $T_D'(t) < 0$ (shrinking queue).

[8] Proof: If c_T, c_S, and F are everywhere differentiable, then $c_S' = 0$ implies $T_D'(t) = 0$; *i.e.*, $T_D(t)$ must be a maximum where c_S is a minimum.

bottleneck model, and will again eliminate all travel delay costs and leave schedule delay costs unaffected. Optimal tolling now reduces generalized cost by more than half (instead of exactly half as in the basic bottleneck model) because in this case total schedule delay cost is smaller than total travel delay cost in the unpriced equilibrium — as can be seen for example in Figure 3.11 and equations (3.36-3.37).

If there are multiple intersections of $F(t)$ and $B(t)$ in Figure 4.3, the queue will wax and wane more than once over the peak in the unpriced equilibrium (but not necessarily disappear completely in between). Then both schedule-delay costs and travel-delay costs are lower than in the basic bottleneck model. ADL (1988) show that in such equilibria, the queue-entry rates of (3.25) remain valid. Likewise, the optimal toll schedule has slopes β and $-\gamma$ just like that for the basic bottleneck model, and it again eliminates all travel delay costs and leaves schedule delay costs unaffected.

ADL (1988) also consider the contrasting case, in which distinct groups of travelers have identical desired schedules $t_d=t^*$ but different relative costs of schedule delay versus travel delay, as measured by the ratios β/α and γ/α. This model is interesting because it shows how marginal external costs may differ between seemingly identical vehicles, and also because it illustrates that in the unpriced equilibrium, the ordering of travelers as well as their departure rates may be non-optimal.

Figure 4.4

Figure 4.4 illustrates this case for a simple symmetric example with two groups of users, denoted A and B. For graphical convenience we assume that the groups are equal in size. The parameters α, β and γ are all lower for group A; the ratios β/γ are the same for both groups; but the ratios β/α and γ/α are higher for group A. Group A could be blue-collar workers, for whom all shadow prices are lower due to a lower income, but for whom scheduling is relatively more important than delays because they work in shifts.

The solid line in the upper panel shows travel delay $T_D(t')$ in the unpriced equilibrium, which is a piecewise linear function with slopes equal to β_G/α_G and $-\gamma_G/\alpha_G$ during the intervals when people from group G arrive early or late (G=A,B). Equilibrium entails temporal separation, with group A arriving closest to t^*. Both types of driver would lose individually if either were to

reschedule to an arrival time occupied by the other group. The dashed lines extrapolate a group's experienced travel delay function into the other group's arrival intervals, thus showing the required travel delays to make an arrival with these other group's drivers equally attractive as in the own interval(s). These extrapolations are always below equilibrium travel delays, confirming that there is no incentive to mix with drivers from the other group.

Drivers in Group A benefit from heterogeneity, in the sense that they would be worse off if a type B driver changed and became type A. This is because type B drivers spread out their departure times more and therefore cause less travel delay. The marginal external cost is therefore higher for type A drivers than for type B drivers: the identity of an extra driver is immaterial to group B, but group A prefers an additional driver to be of type B (see Arnott and Kraus, 1998a). As we will see later, the second-best "flat" (time-independent) tolls would therefore differ between the two groups, requiring making these tolls to be non-anonymous.

The optimal time-varying toll schedule, shown in the lower panel of Figure 4.4, eliminates all travel delay and therefore requires the slope of the toll schedule $\tau(t')$ to be equal to β_G and $-\gamma_G$ during group G's arrivals. It induces a voluntary reversal in arrival times: group B now arrives closest to t^*. Again, the dashed lines show for each group the hypothetical toll schedule that would make its members willing to switch to the other group's arrival interval; it is below the actual toll schedule so they do not switch. The toll eliminates all travel delays, as in the basic bottleneck model, and also produces another efficiency gain: it reduces aggregate schedule-delay cost because group B, with the higher β and γ, now exit the queue closer to the desired time t^*.

Another interesting reversal occurs: it is now group B that benefits from heterogeneity, while group A does not. This is consistent with another property: it can be shown that before redistribution of toll revenues, group A suffers from the imposition of tolling while group B benefits. The finding that high-value-of-time users can benefit from imposition of optimal pricing, while low-value-of-time users lose, also occurs with the static model.

Networks of Bottlenecks

It is not necessarily straightforward to generalize from the above single-bottleneck models to networks of bottlenecks. ADL (1998) provide an insightful discussion. Interactions between different bottlenecks may be either simpler or more complicated than between links in a conventional static network.

For example, a particularly simple result appears in the case of two bottlenecks in series, with different capacities and with no active origin or destination in between. The unpriced equilibrium, the optimum, and the optimal time-varying toll are all the same as in absence of the higher-capacity bottleneck – regardless of which of the two bottlenecks is upstream of the other. The higher-capacity bottleneck can therefore be entirely ignored in the analysis.[9]

A more complex case is two bottlenecks in parallel. ADL (1990a) consider this situation with homogeneous users. The free-flow travel times T_f can then no longer be set arbitrarily to zero unless they are equal between the two roads. Again, all travel-delay costs are eliminated in the optimum. The time-varying tolls are analogous to those for a single bottleneck, and the timing of exits and the route split are identical between the unpriced equilibrium and the optimum.

One might be tempted to conclude from these examples that queuing can never be socially optimal on a network of pure bottlenecks. However, De Palma and Jehiel (1995) show that some queuing may be optimal when drivers are heterogeneous. It seems hard to identify precisely the characteristics of a network that would make this happen. But regardless, setting tolls naïvely to eliminate all queues will typically produce substantial efficiency gains even if it is not optimal (De Palma, Kilani, and Lindsey, 2004).

For all but very simple networks, one is forced to numerical methods to find equilibria. Ben-Akiva, De Palma, and Kanaroglou (1986a) provide a computer simulation model that includes both mode and route choice. Rather than model equilibrium directly, they consider traveler behavior that responds to conditions on previous days, and simulate the approach to a stable pattern. The METROPOLIS model mentioned in Chapter 3 also uses this technique while simulating individual traveler decisions about trip schedule, route, and mode. Congestion METROPOLIS consists of queuing on each link, and scheduling decisions incorporate a preferred time-of-arrival window, schedule-delay costs from arrival outside this window, and logit choice between automobile and public transit.

---

[9] To eliminate travel delays, the optimal entry rate must be equal to the lower capacity, and the optimal time-varying toll is set accordingly. The higher-capacity bottleneck therefore remains inactive in the first-best optimum and is irrelevant. In the unpriced equilibrium, the higher-capacity bottleneck may become active when it is located upstream of the other bottleneck; but even so total queuing time is independent of whether or not the higher-capacity bottleneck exists.

To determine the optimal time-varying fee in such models, one must know all the parameters and be able to solve the entire model. In practice, one needs a way to use observed data to adjust the fee in response to changed conditions. De Palma and Arnott (1986) show that, in at least one very stylized case, setting a fee that varies instantaneously with queue length will approximately accomplish this.

Alternative Dynamic Congestion Technologies

The difficulties of combining endogenous scheduling with more complex traffic-flow models has deterred most researchers. We examine two such studies, focusing on how well the insight from the basic bottleneck model hold up when extended to more realistic settings.[10]

Chu (1995) investigates optimal pricing for a road that is characterized by no-propagation flow congestion as presented in Section 3.3.3, where the travel time depends solely on the flow at the road's exit at the instant that the trip is completed, $V_o(t')$. The demand side of his model is the same as for the basic bottleneck model discussed above. The optimal time-varying toll turns out to be a straightforward dynamic generalization of the standard toll for static congestion in (4.6):

$$\tau(t') = V_o(t')\alpha \frac{dT(V_o(t'))}{dV_o(t')}. \tag{4.30}$$

This expression follows from the optimality condition that the marginal social cost of an arrival at t' should be constant throughout the period during which arrivals occur.[11]

Chu provides an interesting comparison between his model and the basic bottleneck model. Several qualitative differences are worth emphasizing. First, total schedule delay cost is smaller than total travel delay cost in the unpriced equilibrium. This is due to the fact that in order to keep the generalized price p^0 constant over time, the exit rate needs to be higher the closer to t^* one completes the trip. Relatively many drivers therefore have relatively high travel delay costs and relatively low schedule delay cost. The simple geometry of Figure 4.2, which

[10] Two early examples are Chang, Mahmassani, and Herman (1985) and De Palma, Lefèvre, and Ben-Akiva (1987). See Mahmassani (2000) for a brief review.

[11] Mun (1999) combines Chu's approach with a downstream pure bottleneck, and finds that the optimal toll is described by (4.30) during the shoulders of the peak where the exit rate is below the bottleneck's capacity, and takes on the basic bottleneck form with slopes β and $-\gamma$ otherwise.

was due to the constancy of the queue-exit rate, breaks down. Second, the optimal toll does not eliminate all travel delays in Chu's model, and raises total schedule delay cost. Third, applying the optimal toll lengthens the duration of the peak because it reduces the rate at which trips near the desired arrival time are completed. Fourth, applying the optimal toll raises the generalized price of traveling, because of the increased peak duration.[12] Finally, the total variable cost in Chu's model depends on the value of travel delay, α.

An interesting question is whether these results are due to the difference between flow congestion and bottleneck congestion or to the absence of hypercongestion in the unpriced equilibrium in Chu's model. (Recall that queue density in the basic bottleneck model is infinite, which can be seen as an extreme form of hypercongestion). Some insight can be obtained from numerical analyses by Verhoef (2003), which can be compared to numerical results provided by Chu. Verhoef uses the car-following model of (3.20) to investigate optimal time-varying tolls on a road with a sudden reduction in the number of lanes by half. This model does produce hypercongestion with real effects in the unpriced equilibrium, in the form of a queue immediately upstream of the bottleneck. Optimal tolling eliminates queuing, but increases the duration of the peak and therefore also raises the generalized price. Qualitatively, the differences with the basic bottleneck model are therefore similar to what is found in Chu's model, but Verhoef's simulation results are somewhat closer than Chu's to what is predicted by the basic bottleneck model.

Thus, it appears that the basic bottleneck model overestimates the benefits from optimal tolling, and underestimate the resulting increase in generalized price, by exaggerating the extent to which travel delays can be eliminated without increasing scheduling costs. But opposite biases may exist in flow-based dynamic models, perhaps because they ignore the waste of hypercongested queuing. The simulation results of Mun (2002) support similar conclusions.

Verhoef's study also provides a potentially useful computational device for approximating the optimal toll schedule. Although no analytical solution for it is found, more than 99 percent of the possible welfare gains from tolling appear to be captured by adopting a time- and location-

[12] The toll and travel delay are both zero for the first and last driver arriving, in the unpriced equilibrium as well as in the optimum. The generalized trip price in both regimes, which is equalized for all drivers, must be equal to the sum of the values of free-flow travel time and schedule delay for the first or the last driver. The fourth result therefore follows from the lengthening of the peak under optimal tolling.

specific version of (4.30). Specifically, the toll on a link is calculated, at each point in time, as the integral over distance of:

$$\tau(t,x) = V(t,x)\alpha \frac{d[1/S(V)]}{dV}\bigg|_{V(t,x)}, \tag{4.31}$$

where $S(V)$ is the stationary-state speed flow function from which the first-order car-following equation is derived. In other words, the logic of the basic Pigouvian toll in (4.6) not only extends to a dynamic setting with flow congestion, as shown by (4.30), but apparently also to more realistic cases where congestion varies both temporally and spatially in a continuous manner. The shape of this link toll as a function of time is approximately triangular, similar to that in the basic bottleneck model, but it does not fully eliminate travel delays: in fact, eliminating all travel delays by naïvely copying the triangular toll schedule from the basic bottleneck model is a disastrous policy in Verhoef's numerical simulations, as it produces a welfare loss (compared to the unpriced equilibrium).

A practical advantage of a toll based on (4.31) would be that it can in principle be set adaptively, based on local and instantaneous traffic conditions. This allows a regulator to iterate toward the desired toll by trying a toll structure, observing $V(t,x)$, and calculating (4.31) from a previously measured speed-flow relationship.

Discussion

Dynamic models suggest that a main source of efficiency gains from optimal pricing would be the rescheduling of departure times from the trip origin. With heterogeneous users, additional gains may result from changing the order of arrivals at the trip destination, too. Despite severe practical limitations, the basic bottleneck model and its generalizations have proven useful in generating understanding of these features.

4.2 Second-Best Pricing

The rules for marginal-cost pricing discussed in the previous section are often referred to as "first-best" because there are no constraints on the pricing instrument and there are no market distortions other than the congestion externality. Although useful as a theoretical benchmark, first-best pricing is increasingly recognized to be of limited practical relevance. We therefore

need to consider explicitly how such constraints and indirect effects alter the properties tolls and the best way to designed them.

This section considers five cases: when not every link in a network can feasibly be priced; when tolls cannot be varied smoothly over time; when tolls cannot be differentiated among different classes of users; when tolls must be set before knowing what the actual demand and/or capacity will be; and when pricing affects labor supply which is already discouraged due to income or other taxes. Much of our discussion follows Lindsey and Verhoef (2001).

4.2.1 Network Aspects

Probably the earliest example of second-best road pricing considered in the literature concerns the case where not every congested link in a network is tolled. Reasons for this might include excessive cost for charging each link, political constraints requiring that road charging be implemented gradually, or an acceptability constraint that a toll-free alternative should always be available. Lévy-Lambert (1968) and Marchand (1968) were the first to address this type of problem, using a simple network featuring a toll road and a parallel untolled road between a common origin and destination.[13] This set-up is still relevant today, as it describes the main constraint faced when setting a congestion toll on an express lane with parallel untolled lanes.

But the general problem of untolled links encompasses a much wider set of practical congestion-charging mechanisms. A toll cordon as used in various Norwegian cities corresponds with a situation where tolls are in place only on the links that define the cordon. An area charge as applied in central London can be viewed as a priced "virtual" link that is added for modeling purposes to all routes that pass through the charging area, with all others unpriced. Parking charges can be modeled by adding a tolled virtual link to all routes that end within the area in which the parking charge applies.

We start our discussion with the static two-route problem just mentioned, and provide the full derivation of the second-best optimal toll for this case. We then go on to other second-best problems, which can be solved using similar Lagrangian approaches.

---

[13] Pigou (1920) in fact considered the same network, but because he assumed one road to be uncongested, the derivation of the optimal toll on the other road became a matter of first-best optimization.

Static Congestion: Two Routes in Parallel

The classic two-route problem considers static congestion on two parallel roads connecting a single OD-pair. A congestion toll can be applied on one of the two roads (route *T*), while the other (route *U*) must remain untolled. Total traffic V is equal to the sum of traffic on the two roads, V_T and V_U, and average user cost on road R is $c_R(V_R)$.[14] Users consider the roads as pure substitutes, so Wardrop's first principle applies. Users are homogeneous in all respects except that willingness to pay for trips differs across users, so that overall demand $d(V)$ is elastic, where $d(\cdot)$ is the generalized price including user cost and toll.

The optimal toll on route T, τ_T, can be found by maximizing the following Lagrangian function:

$$\begin{aligned}\Lambda = & \int_0^{V_T+V_U} d(v)\,\mathrm{d}v - V_T \cdot c_T(V_T) - V_U \cdot c_U(V_U) \\ & + \lambda_T \cdot \left[c_T(V_T) + \tau_T - d(V_T+V_U)\right] + \lambda_U \cdot \left[c_U(V_U) - d(V_T+V_U)\right]\end{aligned} \tag{4.32}$$

where the first three terms form the social objective of social surplus, the variables λ_R give the shadow price of the equilibrium constraint for route R, and capacities are suppressed as arguments in cost functions because we look at the short run.

The equilibrium value of a Lagrange multiplier reflects the marginal impact of a relaxation of the associated constraint upon the optimized value of the objective. For the Lagrangian (4.32) this means that the equilibrium value of λ_R should give the impact on social surplus of a marginal increase in the toll on route R. We therefore expect to find $\lambda_T = 0$ in the second-best optimum (τ_T should be set optimally), and $\lambda_U > 0$ when route U is congested (since surplus would be increased if a positive toll could be included in the second constraint). The first-order conditions to (4.32) confirm this:

[14] Thus c_T here refers to user cost on the tolled road (route *T*), not to travel-delay cost in the sense of Section 4.1.2.

$$\frac{\partial \Lambda}{\partial V_T} = d - c_T - V_T \cdot c_T' + \lambda_T \cdot \left(c_T' - d'\right) - \lambda_U \cdot d' = 0$$

$$\frac{\partial \Lambda}{\partial V_T} = d - c_U - V_U \cdot c_U' - \lambda_T \cdot d' + \lambda_U \cdot \left(c_U' - d'\right) = 0$$

$$\frac{\partial \Lambda}{\partial \tau_T} = \lambda_T = 0 \quad , \tag{4.33}$$

$$\frac{\partial \Lambda}{\partial \lambda_T} = c_T + \tau_T - d = 0$$

$$\frac{\partial \Lambda}{\partial \lambda_U} = c_U - d = 0$$

where primes denote derivatives. These first-order conditions can be solved to yield:

$$\lambda_U = \frac{V_U \cdot c_U'}{c_U' - d'}, \tag{4.34}$$

which is positive as expected (recall that d', the slope of the inverse demand function, is non-positive). Furthermore, λ_U increases in the congestion externality on route U (the numerator) and in the sensitivity of equilibrium demand to price changes on route U (which varies inversely with denominator). Both factors would indeed boost the welfare gains from introducing a marginally positive toll on route U.

Using the two equilibrium values of λ_R, we can substitute Wardrop's condition for route T (*i.e.*, the constraint corresponding to λ_T) into the first-order condition for V_T. We then find the following second-best toll τ_T:

$$\tau_T = V_T \cdot c_T' - V_U \cdot c_U' \cdot \frac{-d'}{c_U' - d'} \ . \tag{4.35}$$

The toll is equal to the marginal external congestion cost on route T, minus a certain fraction of the marginal external congestion cost on route U. The first term reflects the direct beneficial impacts of the toll upon congestion on route T; the second term captures its indirect spillovers on route U through induced route diversion. Under perfectly inelastic demand ($d' \to -\infty$), the fraction becomes unity and the two effects are equally important: only route choice matters for

overall efficiency with inelastic demand, and it is optimized by setting the toll on route T equal to the difference between marginal external costs on the two routes. In contrast, when demand is perfectly elastic ($d' = 0$), the second term vanishes: V_U cannot be affected by τ_T because it is fully determined by Wardrop's condition $d = c_U(V_U)$, with d constant. The best thing the regulator can then do is to ignore route U altogether and optimize the use of route T by perfectly internalizing the congestion externality.

This example illustrates two lessons that turn out to be more general. A first is that the difference between welfare gains from first-best and second-best pricing can typically be determined only when detailed information is available. The upper curve in Figure 4.5 illustrates this point by plotting, for the simulation model of Verhoef and Small (2004), the relative efficiency of second-best pricing — *i.e.*, its welfare gain (compared to no pricing) as a fraction of the welfare gain from first-best pricing. This relative efficiency (the top line in the figure) rises from 0 to 1 as the relative size of priced capacity increases from 0 to 1. The sigmoid shape of the curve underlines unless a significant portion of capacity is priced, the welfare losses from spillovers on unpriced capacity, reflected in the second term of (4.35), are substantial. The relative efficiency at one-third of capacity in Figure 4.5, approximately 0.3, is higher than the corresponding estimate by Liu and McDonald (1998) for the express lanes on the SR-91, which is about 0.1. Probably the reason is that Verhoef and Small allow for heterogeneity in values of time; the relative efficiency increases with heterogeneity because the travel time gains on the tolled lane are enjoyed by higher-value-of-time users, whereas the travel time losses on the free lane are incurred by lower-value-of-time users (see also Small and Yan, 2001). Both with and without heterogeneity, however, the gains from express lane pricing are disappointingly small in static models. As we will see, more optimistic conclusions are derived from dynamic models.

Figure 4.5

A second lesson from the example leading to (4.35) is that second-best pricing rules are usually more complex than first-best; compare equation (4.6). The reason is that second-best

rules account for the indirect effects of the price.[15] Because the rules are more complex, the chance of making mistakes in setting second-best optimal tolls is also greater; accounting for this would further diminish the expected welfare gains from pricing.

One might think a safer course is to use first-best reasoning wherever constraints do not apply and hope it is close enough to a true second-best solution. We call such a prescription *quasi first-best* tolls, defined in the current example as toll $V_T \cdot c_T'$ on route T and no toll on route U. However, this is not an attractive option either because it is often far inferior to the second-best toll. It may even lead to a welfare loss compared to no pricing, as shown by the middle line in Figure 4.5: in this example, unless at least 65% of capacity is priced, the quasi first-best toll is worse than no toll.

The third curve in Figure 4.5 shows the relative efficiency from a private (revenue-maximizing) express lane, which will be discussed in Section 6.1.

Dynamic Congestion: Two Bottlenecks in Parallel

By allowing for gains that arise from departure-time adjustments, dynamic models of the two-route problem tend to give more optimistic indications of the relative welfare gains from second-best pricing (Braid, 1996; De Palma and Lindsey, 2000). Consider two bottlenecks of equal capacity V_K in parallel, in a setting that is otherwise identical to the basic bottleneck model. Now suppose one can implement a time-varying toll on one of the two bottlenecks, but no toll on the other. In this case the quasi first-best toll, as given by (4.23) with Q taken to be only those people using the tolled road, will eliminate all travel delays for this bottleneck. Because the equilibrium price on the tolled bottleneck remains unaffected by the toll, it induces no route shift and therefore produces no spillover of congestion onto the unpriced road. As a result, the relative welfare gain of this policy is exactly 50%.

---

[15] This is a particular application of the so-called envelope theorem, which implies that indirect effects of a marginal change in a choice variable upon the objective are zero when evaluated in the full optimum. For first-best pricing on general networks, the changes in other link flows that would follow from a marginal increase in one of the link tolls have a zero net welfare effect, because all other links in the network are optimally priced and only carry traffic from OD-pairs for which marginal benefits and marginal costs are equated. The reader may verify that if the first-best general network problem of equation (4.16) were solved using the Lagrangian technique of (4.32) and using route-based tolls as policy instruments, all route-specific multipliers would indeed be equal to zero in the optimum. Indirect effects therefore vanish in first-best optima.

Even higher welfare gains occur using a true second-best toll on the priced link (Braid, 1996). This toll takes into account the fact that with equal traffic, the marginal cost on the tolled route (where only schedule delay cost is incurred) is lower than on the untolled route (where both schedule delay cost and travel delay cost are incurred) — in this example it is only half as large, as can be seen by comparing equations (4.22) and (4.27) with equal values of Q. To equalize them, the tolled alternative should carry two-thirds of all traffic and the untolled alternative one-third. This requires a time-independent subsidy to users of the tolled route, added to the time-dependent toll, such that exactly half the users of the tolled bottleneck receive a net subsidy while the other half pay a net tax. This second-best equilibrium achieves two-thirds of the welfare gain from first-best pricing.[16]

Other Networks

The two parallel routes discussed above is just one of many network settings where second-best congestion pricing is relevant. A similar setting arises when travelers face a mode choice between driving and using public transport (Tabuchi, 1993) where public transport is obliged to be self-financing for political or other reasons. Average-cost pricing of transit causes the generalized price of a transit trip to exceed its marginal social cost, due to scale economies as argued in Section 3.2.4. The second-best toll on the road then exceeds marginal external cost in order to boost demand for public transport. This can also be viewed as an example of a distortion in the economy outside the road sector, in this case making it desirable to charge road prices higher than the Pigouvian values.

Verhoef (2000a,b) derives second-best optimal tolls for any sub-set of tolled links on a network of arbitrary size and shape.[17] The toll formulae are complex; (4.35) is a special case. Verhoef shows how in larger networks, the Lagrange multipliers reflecting zero-pricing constraints may be useful in computational algorithms for finding second-best optima. A

---

[16] Total cost, for an overall demand of Q and two equal capacities V_K, can be determined by substituting $\frac{1}{3}\cdot Q$ in (4.21) and $\frac{2}{3}\cdot Q$ in (4.26) and adding the costs thus obtained, which yields a total of $\frac{1}{3}\cdot\delta Q^2/V_K$. Total cost in the unpriced equilibrium and in the first-best optimum can be found from (4.21) and (4.26), respectively, with V_K replaced by $2\cdot V_K$. The results are $\frac{1}{2}\cdot\delta Q^2/V_K$ and $\frac{1}{4}\cdot\delta Q^2/V_K$. Thus the relative welfare gain of second-best pricing is $(\frac{1}{2}-\frac{1}{3})/(\frac{1}{2}-\frac{1}{4})=\frac{2}{3}$.

[17] Chen and Bernstein (2004) address the same problem with multiple user groups.

complication in such algorithms is that the interior second-best optimum need not be unique nor even exist — especially if there are untolled links with relatively high marginal external costs.

Second-best questions can be explored by trial and error rather than by formal optimization. A typical starting point is one of several well-developed applied network models, in which route choices are determined as a user equilibrium using either Wardrop's first principle or some stochastic route-choice mechanism. The model is then applied to pricing schemes that incorporate the researcher's view of technical, political, acceptability, equity-based, or other practical considerations. In some cases a search is carried out over various parameters such as toll levels and number and locations of toll charging points.

The trial-and-error approach has been applied quite successfully to at least two questions. The first question is how to introduce pricing incrementally by applying it to just a subset of the network, perhaps in conjunction with other policies such as free passage to high-occupancy vehicles. Safirova *et al.* (2004) provide a good example, analyzing policies for the Washington metropolitan area using a model known as START that was developed in the UK. As a general rule, such studies have found that the best candidates for road pricing are expressways and major urban arterial roads, because of their high traffic volumes, high speeds, and lack of close substitutes (Verhoef, 2002b).

The other question is toll-cordon design: where to place the cordon, how high to set the toll, and how to improve incrementally over a pure cordon. (We define a pure toll cordon as one where a single charge is applied to anyone crossing a well-defined boundary enclosing a highly congested area.) We give just a few examples of such studies here. Santos, Newbery, and Rojey (2001) apply the SATURN network model to eight small English cities, finding that pure cordon tolls produce large reductions in cordon crossings with only small impacts on overall quantity of vehicle travel. They also find that the cordons can produce substantial welfare gains, but that those gains are easily lost if the toll is too high. Researchers from Leeds have applied the SATURN model to road networks depicting Cambridge (England), Edinburgh, and other cities.[18] Their results suggest that even simplified analytical rules for choosing the tolled links produce great improvements over cordons selected by expert judgment. These and other studies show that

[18] E.g., May et al. (2002); Sumalee, May, and Shepherd (2005). These two studies, as well as Zhang and Yang (2004), use a genetic algorithm to facilitate search through many possible configurations of charge points and toll levels.

substantial further welfare gains can be achieved by allowing the toll levels to differ at different charging points or by using two cordons, one inside the other, with different charge levels.Yet other studies incorporate land-use effects by considering cordon tolls within broader models of urban structure.[19]

4.2.2. Time-of-Day Aspects

A second type of constraint on congestion tolls is when they cannot be varied freely over time. The simplest but most instructive examples again involve the basic bottleneck model.

Flat Pricing of a Bottleneck

Assume first that only a single fixed ("flat") toll can be charged throughout the peak period. This toll can affect the overall use Q, provided demand is not completely inelastic, but it does not affect queue-entry rates. Therefore the reduced-form average, total, and marginal cost functions for an unpriced equilibrium, (4.20)-(4.22), remain valid. The second-best optimal flat toll τ_F is therefore:

$$\tau_F = \overline{mc}_g^0 - \bar{c}_g^0 = \delta \cdot \frac{Q}{V_K}. \tag{4.36}$$

The optimal time-invariant charge (4.36) is exactly twice the time-average of the optimal time-varying toll (4.23); intuitively, this is because adding a new traveler to the system imposes both schedule-delay and travel-time costs, instead of just the former. The relative efficiency of this second-best charge depends strongly on the elasticity of demand.

The results are a little more complex if we allow for dispersion in desired queue-exit times, as in the model of Section 3.4.3. Let V_d and q again denote the height and width of the density function describing desired queue-exits (i.e., people would like to exit at rate V_d over time interval q). Hence $Q \equiv q \cdot V_d$. The relevant average cost function is that in (3.38):

$$\bar{c}_g^0 = \delta \cdot \frac{qV_d}{V_K} \cdot \left(1 - \tfrac{1}{2} \cdot \frac{V_K}{V_d}\right). \tag{4.37}$$

[19] Mun, Konishi, and Yoshikawa (2003, 2005).

The marginal cost now depends on whether a marginal increase in use would *intensify* the period of desired queue-exits by increasing V_d, or *extend* it by increasing q.

To fix ideas, suppose one wishes to influence the decisions of developers whose buildings generate traffic according to some known process, by charging them with the marginal external cost of new traffic generated assuming that tolls cannot be differentiated by time of day. The new development might then *intensify* desired queue-exits if work-start times are the same everywhere and buildings are all about the same distance from the bottleneck's exit. In this case the marginal cost of one unit of traffic (an increase in Q holding q constant) is $(1/q)\cdot\partial\left(qV_d\bar{c}_g^0\right)/\partial V_d$; subtracting average cost (4.**Fout! Verwijzingsbron niet gevonden.**) yields the following optimal flat toll (where superscript I denotes *intensify*):

$$\tau_F^I = \delta\cdot q\cdot\frac{V_d}{V_K}, \tag{4.38}$$

which is identical to (4.36).

The new development might *extend* desired queue-exits (raise q but not V_d) if firms practice staggered work hours or if the development is at the edge of an already developed area. The marginal cost of one unit of traffic (an increase in Q holding V_d constant) is now $(1/V_d)\cdot\partial\left(qV_d\bar{c}_g^0\right)/\partial q$; subtracting average cost (4.**Fout! Verwijzingsbron niet gevonden.**) yields the following optimal flat toll (with superscript E designating *extend*):

$$\tau_F^E = \delta\cdot q\cdot\left(\frac{V_d}{V_K}-\tfrac{1}{2}\right), \tag{4.39}$$

which is smaller than (4.36). One would therefore charge a developer (or the drivers using the new development) less for extending demand than for intensifying it, because extending demand has less impact on other users.

Coarse Tolls

A more sophisticated variant of the flat toll is a coarse or "step" toll, defined as one that allows one or more non-zero toll values, each applying during a specific time interval that can be chosen to maximize welfare (ADL, 1990, 1993). The simplest has a single step: the toll is zero except for a time interval (presumably a subset of the peak period) when it is a positive constant.

Analytical derivations for this case are cumbersome, even for the basic bottleneck model. Using numerical simulation, ADL find that the second-best single-step toll produces relative efficiency gains just over one-half. This performance compares favorably to the flat toll, whose relative efficiency is less than one-third even under fairly generous assumptions about the elasticity of demand for peak-period travel. Chu (1999) also compares coarse and flat tolls, finding relative efficiencies of 38–79% for the coarse toll and 27–66% for the flat toll, depending on parameters.

Network and Time-of-Day Aspects Combined: An Application

De Palma, Kilani, and Lindsey (2005) use the dynamic queue-based METROPOLIS model, described in Section 4.1.2, to simulate a number of second-best policies on a stylized circular network with eight radial roads connecting four concentric ring roads to a central point. Origin-destination demands are distributed around the network. They find that welfare gains are substantially higher with step tolls (in half-hour steps) than with flat tolls, and also higher with area pricing (i.e. pricing all trips within a cordon) than with cordon pricing. Step tolls also have a more favorable acceptability impact than flat tolls, in that they have a higher proportion of travelers whose consumer surplus change is positive (without considering use of revenues). However, area tolls have a lower acceptability impact than cordon tolls. Thus area pricing with step tolls achieves the highest welfare gain (61% of the first-best gain); while cordon pricing with step tolls achieves the highest proportion of positive consumer-surplus changes (41%), along with a substantial relative welfare gain (44%). These results further support the cordon studies described earlier in suggesting that pure cordons can be greatly improved upon by allowing even modest flexibility in the locations and toll levels of charge points.

4.2.3 User Heterogeneity

Another second-best problem arises when the price cannot differ by user group. In this case, the second-best toll is a weighted average of the marginal external costs for the different groups, at least within a static model (Verhoef, Nijkamp, and Rietveld, 1995a). A group's weight depends positively on its price-sensitivity of demand. The welfare losses from undifferentiated prices generally increase in the price-sensitivity of demand of those user groups whose marginal external costs differ the most from the weighted average.

Undifferentiated tolls give up two advantages of optimally differentiated tolls: they fail to secure optimal use levels by all groups of travelers, and also to provide optimal incentives for choice among groups. For example, tolls differentiated by pollution emissions would encourage owners of dirty cars especially to curtail their use (securing optimal use levels across groups), and also would encourage them to buy clean cars instead (choice among groups). With undifferentiated charges, supplementary policies may become desirable. For example, inability to differentiate tolls by emissions technology might justify setting technological standards for emissions control, and inability to differentiate tolls by distance traveled might justify land-use controls or spatial planning. Doing this in a second-best optimal manner requires the regulator to possess much more information than just the marginal external costs that are required for short-run optimal pricing. The risk of mistakes in second-best policies is therefore again larger than for first-best policies.

4.2.4 Stochastic Congestion and Information

Uncertainty about actual traffic conditions may also affect the optimality of pricing rules. It is useful to distinguish between two types of uncertainty: idiosyncratic and objective.

Idiosyncratic uncertainty exists when traffic conditions are predictable, but individual travelers do not know them precisely and instead form idiosyncratic perceptions of their own travel times. The standard approach to describing behavior in this case is the Stochastic User Equilibrium (SUE), discussed in Section 3.4.4. Smith, Eriksson, and Lindberg (1995) and Yang (1999a) investigate the conditions under which a system optimum (in which total travel time is minimized) can be supported as a stochastic user equilibrium with non-negative tolls. Under certain conditions, tolls equal to marginal external cost will accomplish this.

Objective uncertainty exists when traffic conditions vary unpredictably due to accidents, bad weather, demand shocks, or other factors. Chapter 3 already discussed some aspects of this: for example, that providing real-time travel information is guaranteed to increase welfare only when optimal pricing is in place. Recent research has elaborated on such potential synergies between information and road pricing. However, the jury is still out on whether the benefits from pricing and information provision are additive — *i.e.*, equal to the sum of benefits of each instrument in isolation.

Verhoef , Emmerink, Nijkamp, and Rietveld (1996) find in a static model that if information is perfect, pricing and information provision have approximately additive benefits and are complementary: under conditions when one instrument does not yield much benefit, the other does particularly well. El Sanhouri and Bernstein (1994), using a dynamic model with endogenous trip-timing decisions, likewise find the benefits of non-responsive pricing and information provision to be approximately additive. Yang (1999b) finds that the answer depends on how many drivers receive information: super-additivity occurs beyond a certain level of market penetration. De Palma and Lindsey (1998) show that information may be welfare-reducing unless it is supplemented by *responsive* pricing (*i.e.*, the price depends on conditions that are realized *ex post*).

The combination of non-responsive pricing and perfect information provision is nearly as efficient as first-best responsive pricing for most parameter combinations in the static model of Verhoef *et al*. just described. This might make it an attractive combination in practice. It avoids the probably unpopular feature of responsive first-best pricing that the toll would be exceptionally high just when conditions are worst. More generally, unpredictability of tolls may reduce their social acceptability. Note that non-responsive pricing does not imply flat pricing: the toll could still vary continuously over time, but in a predictable fashion.

4.2.5 Interactions with Other Distorted Markets

The imposition of prices in transport, as well as the use of the revenues, may have non-marginal welfare effects in other markets when these markets do not function efficiently. Such distortions can be consider constraints, from the point of view of transport pricing, in the sense that they could be eliminated with appropriate instruments but such instruments are assumed unavailable.

An important example is labor markets. Income taxes lower the nominal wage paid to labor, while indirect taxes (e.g. sales or value-added taxes) tend to increase the price level and thereby to reduce the real wage. In both cases, the taxes create an incentive against work that will reduce labor supply below its efficient level unless the incentive is offset in other ways;[20] the resulting cost to the economy is called the *deadweight loss* of the existing tax system. If

[20] Kaplow (1996) suggests that the labor-supply distortion is in fact largely offset because people who earn more also receive more benefits from public services.

congestion tolls also discourage labor supply, then they will aggravate this deadweight loss. If congestion tolls also discourage labor supply, they will aggravate this deadweight loss. That aggravation may be partially mitigated if toll revenues are used to reduce the original distorting labor tax. Thus, revenue uses have efficiency effects as well implications for equity and political feasibility.

Mayeres and Proost (2001) evaluate the efficiency effects of transportation charges in Belgium by using a general equilibrium model to compute the marginal welfare cost of public funds for a number of tax instruments. They consider a social welfare function that incorporates a certain degree of aversion to income inequality, through a parameter that can be varied parametrically. They find that a marginal increase in peak-period road transport prices yields the highest benefit when revenue is spent on road capacity expansion. Because public transport is already heavily subsidized, they find a negative benefit from road pricing if revenues are spent on public transport unless the degree of social inequality aversion is very high.

Parry and Bento (2001) find general-equilibrium effects of road pricing that are large and sensitive to the allocation of revenues. Specifically, the interaction with a tax-distorted labor market can cause road pricing to be welfare-reducing if revenues are distributed in a lump-sum manner. On the other hand, when revenues are used to reduce the distorting labor taxes, the usual efficiency advantage of such tax reductions is magnified by the same complementarity. In that case welfare gains are twice as large as would be predicted by the standard partial-equilibrium analysis and the optimal tax is set to the Pigouvian level. Van Dender (2003) generalizes the Parry-Bento model, obtaining similar but less extreme results using arguably more realistic assumptions.

4.2.6 Second-Best Pricing: A Conclusion

First-best analysis of congestion pricing provides important insights; but, because constraints and distorted markets abound in reality, second-best analysis is the only way to translate those insights into practical advice in policy design. The resulting pricing rules are more complex, because they reflect indirect effects, and they require more information. This increases the chance of making mistakes in setting prices. Yet ignoring indirect effects can result in much smaller efficiency gains or even losses.

Happily, second-best models do provide some general guidance as to which factors are most important to consider. For example, toll cordons can be designed that achieve a high proportion of theoretically possible benefits, but they do not necessarily look much like what an expert would intuitively draw on a map. Relaxing typical constraints, such as that all cordon crossings must be tolled at the same amount or that the toll must be flat over the entire peak period, can significantly improve results. Providing information on traffic conditions may sometimes be harmful if tolls are absent or set *ex ante* based on expected traffic conditions. Finally, taxes affecting labor supply can greatly influence the welfare effects of congestion tolls, and accounting for them introduces an important efficiency factor in choosing how to use toll revenues — generally favoring using them to offset those taxes creating the distortion.

Most of the constraints we have considered are really "soft," reflecting judgment about how to account for the regulatory costs (i.e. technical, bureaucratic, or political implementation costs) of toll collection. If those costs could be incorporated into the objective function, then maximizing it would achieve what we might call *broadly first-best* pricing. By contrast, the conventional definition of first-best ignores those costs so is only *narrowly first-best* (Milne, Niskanen, and Verhoef, 2000). We cannot formulate the broadly first-best pricing problem because we cannot quantify the relevant costs. But in many cases, a second-best analysis (one assuming hard constraints) may be a good approximation; if so, we can say that it is broadly first-best to apply second-best pricing. This, of course, leaves open the possibility that future developments could change regulatory costs and thus make certain constraints obsolete.

4.3 Congestion Pricing in Practice

Even ignoring the ancient world, road pricing has a long history, with turnpikes dating back at least to the seventeenth century in England and the eighteenth century in the US (Levinson, 1998). Road pricing for congestion management is more recent. The earliest modern application is Singapore's Area License Scheme, established in 1975. Since then, other applications have appeared, varying from single facilities such as bridges or toll roads to tolled express lanes as in the US, toll cordons as in Norway, and area-wide pricing as in London. We describe here selected applications where managing congestion is explicitly or potentially a significant goal in the design of the pricing scheme. Small and Gómez-Ibáñez (1998) provide a more extended review of the earlier applications.

4.3.1 Singapore

In 1975, Singapore implemented the first operational congestion pricing scheme in the world, the Area License Scheme (ALS). A license had to be purchased and displayed at the windscreen before entering the central 'Restricted Zone' (RZ) during designated peak hours; compliance was monitored manually at control points. Peak hours and tolls were adjusted a number of times. An afternoon peak charge was implemented in 1989, and in 1994 an access fee for inter-peak day-time passages was introduced. When first implemented, the number of vehicles entering the RZ fell by an amazing 44%. While speeds rose dramatically in the zone itself, displaced traffic caused increased traffic outside the zone, and average commuting times to jobs inside the zone even increased (Small and Gómez-Ibáñez, 1998). If this increase in travel times was entirely due to the toll system, it would provide a good example of how second-best distortions and spillovers, when not properly accounted for when setting tolls, may undermine the efficiency gains from road pricing. The fee may have been (far) above the second-best optimal level, as argued for example by Watson and Holland (1978) and McCarthy and Tay (1993).

In 1995, Singapore extended the ALS with a Road Pricing Scheme, which added some pricing of expressways. Although traffic volumes on tolled expressways dropped by 17%, no less than 40% of these drivers switched to major by-pass roads (Goh, 2002), underlining the second-best character of the policy and the empirical relevance of the two-route problem of Section 4.2.1.

Singapore switched to a scheme known as Electronic Road Pricing (ERP) in 1998. Since then, charges are deducted from a smart card when passing through an ERP gantry using microwave technology. The scheme features 28 gantries that form a daytime toll cordon around the central area and 14 tolled expressways and arterial roads further out during morning peak hours only. There are no tolls on weekends. Charges vary by time of day in 30-minute steps and are adjusted quarterly, depending on average speeds realized in the past quarter. Despite the fact that the average charge for ERP is lower than it was for ALS, traffic into the CBD decreased by another 10-15% compared to the ALS scheme (Keong, 2002). One reason might have been that every entry with the same car is charged under ERP, whereas ALS allowed for unlimited access on a single permit.

4.3.2 Norwegian Toll Rings

Four toll cordons have been operated in Norway for some time: Bergen (started 1986), Oslo (1990) and Trondheim (1991), and Stavanger (2001). These toll rings were meant not to manage congestion but to raise revenues. Nevertheless we discuss them here because cordon pricing is often considered as a possible form of congestion pricing, and this possibility has been extensively discussed in Norway. Also, the Trondheim scheme, which has over the years evolved from a ring into a multi-zonal scheme, has some time-of-day differentiation.

Tolls are relatively low, and seasonal passes further reduce the marginal charge. Unsurprisingly, the impact on traffic has been modest, reducing vehicle crossings by no more than 5–10 percent (Tretvik, 2003; Ramjerdi, Minken, and Østmoe, 2004). Time-differentiated charging in Trondheim nevertheless caused substantial shifts in timing for car trips: a decrease by 10% for charged periods and an increase by 8-9% for uncharged periods (Meland, 1995).

4.3.3 Value Pricing in the US

Congestion pricing has gained momentum in the US since federal legislation began funding congestion pricing pilot projects under the so-called Value Pricing Program, and allowing limited pricing on Interstate highways.[21] By the end of 2004, the US Federal Highway Administration listed 39 projects, divided in various categories and further sub-divided into operational projects and projects under development.[22]

One type of project introduces congestion pricing on existing, previously untolled infrastructure. The conversion of High-Occupancy Vehicle (HOV) lanes into High-Occupancy/Toll (HOT) lanes is an example. HOT lanes are free of charge for HOV's, and allow other vehicles to use the lane by paying a toll. Two examples are San Diego's *FasTrak* scheme, implemented in 1999 on Interstate 15 (I-15), and Houston's *QuickRide* scheme on the Katy Freeway and on US290, starting 1998 and 2000, respectively. In the San Diego scheme, the toll is varied in real time depending on congestion, attempting to maintain a target speed, a

---

[21] In particular the Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991 and the Transportation Equity Act for the 21st Century (TEA-21) of 1998.

[22] See the Value Pricing website of the Hubert H. Humphrey Institute of Public Affairs, University of Minnesota, at www.hhh.umn.edu/centers/slp/projects/conpric. We discuss here only the first three categories of value pricing distinguished by FHWA, which are those involving time-varying congestion pricing.

procedure often called "dynamic pricing" although it has nothing to do with the dynamic models discussed in this chapter. The toll is usually between $0.50 and $4.00 but can be as high as $8.00. Half the revenues are used to support transit services. Brownstone *et al.* (2003) analyze users' responses, finding among other things that HOT-lane use is higher for commuters, higher-income people, women, people between 35 and 45 years old, highly educated people, and homeowners.

Another form of tolling existing infrastructure, cordon tolls, is under consideration in Lee County (Florida) and central New York City.

Somewhat more popular in the US is the idea of using tolls to finance new infrastructure. Most such cases are conventional toll roads, with no time differentiation. But one relatively early and well-studied example was initiated in 1995 as a time-varying toll on a new set of express lanes added to the Riverside Freeway (State Route 91, known as SR-91). This is a heavily peaked commuter route in Orange County, leading to employment centers in Orange and Los Angeles Counties. By the end of 2005, the preset tolls varied hourly and by day of the week, with tolls reaching a maximum of $8.00 for an outbound afternoon trip from 4:00-6:00 p.m. on Thursdays. Apart from travel-time savings, users perceive greater comfort and safety from using the Express Lanes (Sullivan, 2000).

While starting out as a private undertaking, the so-called "91 Express Lanes" reverted to public ownership and control in 2003. The main reason was a non-compete clause in the original contract that precluded expansion of competing capacity. This clause became problematic after congestion had grown worse on the untolled lanes, presumably much sooner than the government had envisaged when signing the original contract. This experience emphasizes the differences in interest between public and private road operators (see also Chapter 6): whereas heavy congestion on parallel connections is good news for a private operator because it improves profits, it is generally bad news from the public perspective. Such fundamental conflicts in interests pose great challenges for policies aimed at facilitating private road investments for purposes of relieving congestion.

A third category of value pricing projects involves the introducing time-of-day variation to tolls on existing toll facilities. Examples of such variable tolls include two bridges in Lee County (introduced 1998), on bridges and tunnels crossing the Hudson River between New York and New Jersey (2001), the New Jersey Turnpike (2000), and the San Joaquin Hills Toll Road in

Orange County, California (2002). In many of these cases the variation in rates is small. Such policies appear to be successful in rescheduling trips (DeCorla-Souza, 2004). For example, surveys indicate that over 71% of motorists with a transponder have shifted their travel time at least once a week in the Lee County project in response to a toll difference of just 25 cents. And for the Hudson River crossings, one year after giving off-peak toll discounts of 20%, morning peak traffic had fallen by 7% and evening peak traffic by 4%, with overall traffic stable.

The impacts of value pricing depend of course on local circumstances and project design. But the applications are generally believed to have demonstrated that variable pricing can have substantial impacts on trip timing, vehicle occupancy, and modal choice. For recent reviews, see Ward (2001), Supernak (2001), and DeCorla-Souza (2004).

4.3.4 London Congestion Charging

In 2003, London introduced a congestion charging scheme in its central area covering 22 square kilometers. A charge of £5, later raised to £8, applies to all vehicles driving or parking on public roads in the area between 7:00 a.m. and 5:30 p.m. on working days. Vehicles are identified by number plates using video cameras, an expensive technology but one that could be implemented quickly.

In terms of congestion reduction, the scheme is considered successful. After a year, traffic circulating within the zone had decreased by 15%, and traffic entering the zone by 18%, during charging hours (TfL, 2004). Congestion, measured as the actual minus free-flow travel time per kilometer), decreased by 30% within the zone, also leading to an improvement in reliability. Travelers primarily switched to public transport (50%-60%), and furthermore changed routes around the cordon (20%-30%) or made other changes like carpooling, destination changes, and trip timing adjustments. The surprisingly large mode-switching effect was presumably caused by the good initial coverage of London's public transport network and by further improvements financed in part by charge revenues. The scheme's initial effectiveness was at the high end of the original projections, but financially it was less successful due to an unexpectedly large reduction in traffic and high cost of collecting the charge.

The politics behind the London scheme, described in detail by Richards (2006), are quite unusual. As of this writing, attempts to implement similar schemes elsewhere in the U.K. have

not been successful. Meanwhile the London scheme is scheduled to expand westward, nearly doubling the size of the charged area.

4.3.5 Other Applications

There are more instances of congestion pricing than the cases addressed above. Some recent examples are the Highway 407 north of Toronto, introduced in 1997, and the M6 motorway in Birmingham, in 2003. These have only minimal time-of-day differentiation. Less recent is the Sunday time-varying pricing on route A1 near Paris (Small and Gomez-Ibanez, 1998). Area-wide schemes have been considered several times in Stockholm and in the Netherlands, with Stockholm launching an eight-month trial in January 2006.

4.3.6 Technology of Road Pricing

Decisions on what kind of pricing to adopt are strongly influenced by the capabilities of technologies available for charging motorists. This is a field that is changing very rapidly. Here, we give a brief history of some of the main developments, followed by an analysis of the key tradeoffs facing system designers.[23]

Brief History

As already mentioned, Singapore's original Area Licensing Scheme began in 1975 with the paper stickers mounted on the windshield. Soon after, existing toll roads began introducing electronic transponders mounted in the vehicle and read by roadside equipment. Singapore's current ERP system also uses such transponders, as do most of the pricing implementations discussed above. Communication between transponder and roadside reader is by dedicated short-range communications (DSRC), i.e. radio-frequency signals whose sole purpose is toll charging.

Enforcement of DSRC systems is usually done by taking video images of license plates. Usually these images are processed automatically using character-recognition software, a technology known as automated number plate recognition (ANPR). At least two pricing systems — on Highway 407 in Toronto and in Central London —apply ANPR to all vehicles, rather than

[23] We rely especially on Sorensen and Taylor (2005), Samuel (2005), Bertini and Rufolo (2004), and PRoGRESS (2004).

just those with missing or invalid transponder signals. However, accuracy remains a problem, causing high costs for follow-up enforcement and/or significant loss of revenues from unreadable plates.

As authorities have become more interested in charges that vary by location and distance traveled, they have turned to global positioning systems. A GPS allows a vehicle to determine its location using satellite signals, and this information can be recorded or sent to a central processor. GPS is already being introduced by vehicle manufacturers to provide travel information, route guidance, or logistics control for fleets. Germany installed a GPS system to charge trucks on its motorways starting in 2005, and the UK is studying GPS as part of a future system of distance-based charges on motorways throughout Great Britain (UK Department for Transport, 2004a). Proposals for more widespread GPS-based charging have raised vigorous debate about the potential for government surveillance of individual travel patterns.

Another system, not widely implemented, uses a mobile telephone network to handle communications between the vehicle and the charging system, thereby obviating the need to locate expensive roadside communication devices wherever charging is to take place.

Any of these systems may use a plastic card with a magnetic stripe or embedded computer chip, in conjunction with an on-board card reader and transponder, to store charging information anonymously. Such a card is called a "smart card" if it has sufficient brain power.

Issues and Tradeoffs

A partial list of significant issues that affect technology choice, and hence the degree of potential flexibility in pricing, is as follows.

- *Cost and lifetime of on-board devices*: Because there are so many vehicles, it is very expensive to install and maintain high-priced in-vehicle devices. This factor favors the use of relatively simple transponders (unit cost as low as $8 in 2006)[24] and is creating interest in using even cheaper radio frequency identification (RFID) tags. Of course, these advantages foreclose capabilities that might be useful to meet other objectives. Cheap in-vehicle equipment (or

[24] Peter Samuel, personal communications, 6 July 2006.

ANPR) also increases the market penetration of electronic tolling and thus reduces the expense of providing toll booths and coin machines.

- *Required vehicle speed and location*: Many older devices require vehicles to slow down and stay in a well-defined channel while being charged. Roadside readers have improved so that it is routine to read transponders at highway speeds, but doing so during lane-change maneuvers in normal traffic has proven a challenge and requires more costly and visually intrusive roadside readers. A charging system that requires neither speed reduction nor limitations on lane movements is known as open-road tolling.
- *Location of account information*: Information about stored or available funds may be stored cheaply in a central system, or more expensively in an on-board magnetic or smart card. This affects the user's privacy and flexibility in shifting funds from one vehicle or user to another.
- *Interoperability*: Most users would like to have a single device work for any toll collection point they may come across. The market, however, has produced a variety of incompatible technologies. In many areas agencies within a region have coordinated to adopt the same technology, but this limits their flexibility to innovate. One solution being developed commercially is transponders that are compatible with two or more systems.
- *Integration with other services*: Some users and planners would like to use a single device, such as a smart card, to pay for parking, transit, and even retail transactions. This rules out the least sophisticated devices and requires coordination among agencies.
- *Original Equipment on New Vehicles*: Road-charging capability would be cheaper and easier to disseminate widely if manufacturers built in devices at time of manufacture. Of course, this requires foresight as to what capabilities will be desired throughout the vehicle's life. Such equipment is already sometimes provided for other purposes: for example, a GPS for navigation. This could encourage toll authorities to opt for GPS-based charging systems.
- *Point versus distance-based charging*: As noted, there is interest in pricing schemes that are proportional to distance traveled, sometimes called "variabilization" (e.g. Proost and Van Dender, 1998). It is motivated by several factors including a desire to promote economic efficiency, a desire to reduce vehicle use, and dissatisfaction with fuel taxes as a financing mechanism. GPS and mobile-telephone networks are amenable to distance-based charging, and many DSRC systems already can track individual vehicles as they pass through the network.

- *Automated vehicle classification and occupancy measurement*: Pricing schemes may vary the charge by type of vehicle, occupancy, or vehicle emissions characteristics. Enforcement of these distinctions is difficult, and development is underway to use on-board or roadside devices to record these characteristics automatically.

Obviously decisions about these tradeoffs are sensitive to local conditions and objectives. This makes it likely that no one technology will emerge as the best, and that interoperability will be incomplete. We suspect that just as technological progress has made the toll booth nearly obsolete, it will overcome many limits to interoperability, allowing agencies to meet more of their goals while maintaining transparency to the user. We also believe that even though any given implementation of pricing is constrained by available technology, a far-sighted approach will design the pricing regime to evolve along with technology. The more important source of constraints is not technology, but rather the kind of complexity that users and political representatives will tolerate.

4.4 Pricing of Parking

As shown in Chapter 3, parking accounts for a major part of the social costs of automobile trips to central cities. Furthermore, parking is heavily subsidized by many governments, employers, and businesses. One might suspect, then, that pricing it at either marginal or average cost would make a substantial difference in travel behavior. This suspicion is confirmed by many demand studies (Young et al., 1991; Willson, 1992) and by direct comparisons of the commuting choices of people with and without free parking at work. Eliminating free parking reduced the number of solo drivers by 19 to 81 percent in four sites in Los Angeles, and by 20 percent in Ottawa (Willson and Shoup, 1990, Table 1). In eight California case studies, the number of solo commuters fell by 17% for employers affected by a 1992 law requiring them to "cash out" employer-paid parking by offering an allowance in cash as an alternative (Shoup, 1997).

All this provides ample evidence to support Shoup's (1982) claim that free parking is one of the major factors distorting mode choices in high-density urban core areas. This distortion is especially perverse because subsidized employee parking not only misallocates resources by creating too many parking facilities; it also exacerbates the congestion externality caused by

underpricing peak-hour highways. Furthermore, cities that insist on free downtown parking may undermine the ability of the downtown area to fully capture the advantages of agglomeration because the land needed to produce a high density of activities is instead used for parking (Shoup, 2005, ch. 5). Shoup notes that parking subsidies are encouraged by US municipal zoning ordinances, which typically require builders to provide enough parking to satisfy the demand expected at zero price, and by the greater deductibility of employer-paid parking costs relative to other transportation subsidies under US tax laws governing employer-provided benefits.

The situation for parking at shops is more complex, making it especially important to consider spillover effects before advocating pricing reform. Low parking prices encourage shoppers to drive rather than ride transit, but by reducing turnover (because vehicles will remain parked for a longer time) could actually reduce the total number of auto trips (Glazer and Niskanen, 1992). Street parking also interacts with traffic flow in complex ways (Arnott, 2004). Generally, parking charges probably would reduce congestion but they do not discourage through trips and do not differentiate by how much a given driver adds to congestion.[25]

Parking charges may help to internalize other social costs besides the resource value of the parking spaces themselves. Examples include visual intrusion of parked cars, search externalities in finding parking places, congestion cost imposed by parked cars upon moving traffic, and externalities imposed during the trips made before and after parking.

The parking models considered in Section 3.4.5 have various pricing implications. Arnott, de Palma and Lindsey (1991b) find that a spatially differentiated parking fee is necessary to induce the optimal parking pattern, in which parking spaces furthest from the CBD are taken first instead of last. Anderson and De Palma (2004) similarly find that the optimum in their model can be achieved using a spacially differentiated parking charge, which falls by distance from the centre (as in Verhoef *et al.*, 1995b). Provided searching causes no road congestion, this charge could be achieved through private ownership of parking spaces in a monopolistically competitive market. In the model of Arnott and Rowse (1999), the existence of multiple equilibria implies that a marginal-cost tax on the parking externality cannot guarantee an optimum. Arnott (2004)

[25] See Glazer and Niskanen (1992) and Verhoef, Nijkamp, and Rietveld (1995b). One study calibrated on Brussels suggests that parking pricing might achieve 70 percent of the benefits of congestion pricing, but it excludes departure-time adjustments and route choice (Calthrop, Proost, and Van Dender, 2000).

identifies a potential triple dividend to be reaped from the pricing of parking: through reduced search, reduced traffic congestion, and use of parking revenues to lower other taxes.

Calthrop and Proost (2004) consider the interaction between publicly provided on-street parking and privately provided off-street parking. In their model, wasteful searching for on-street parking occurs when it is cheaper than private parking, to the point where the generalized prices in the two parking markets are equalized. The optimal on-street parking charge is equal to the off-street charge if off-street parking is supplied competitively, and somewhat lower if its suppliers have market power.

Given the cost figures in Section 3.5.3, it seems clear that eliminating subsidized parking for employees in high-density business districts is a high priority for improved efficiency in urban transportation. Doing so would reduce expenditures on parking facilities, free up land for other uses, and favorably alter the modal mix on congested roads. Charging more for on-street parking would have similar advantages and would in addition reduce congestion from cruising.

4.5 Pricing of Public Transit

Setting prices for transit service involves at least three issues: the average fare level, the fare structure, and the incentive effects of transit subsidy programs. We treat each in turn.

4.5.1 Fare Level

We noted in Section 3.2.4 that optimal transit fares might not fully cover average costs because of increasing returns to scale when the value of user inputs is taken into account. In fact, in the simple model presented there, the optimal fare is zero whenever there are empty seats. This can be viewed as an example of public-goods pricing: once it is decided to offer a public good with high enough quality (i.e., frequent enough service that that there are empty seats), it costs nothing to serve an extra passenger. This, of course, abstracts from costs the passenger may impose by slowing the bus, and from adverse effects of taxes that finance the service.

The problem is more complex if a competing mode is not priced optimally. The case most often considered is underpricing of peak-hour automobile travel. Using straightforward models, Glaister (1974) and Henderson (1977, chap. 7) derive the second-best solution and confirm our intuition that it calls for further subsidy of transit service: the more so the higher the cross-

elasticity of demand relative to transit's own-price elasticity. Dodgson and Topham (1987) add several features including distributional preferences, distorting taxes, and cost-sharing by higher levels of government.

We can understand these two arguments for transit subsidies — scale economies and underpriced automobile travel — through a concrete model, adapted in part from Glasiter (1974). Let q_A and q_R be the numbers of automobile and rail transit trips over a specified time period. Recall that the generalized price of mode k (k=A,R) is defined as the average user cost c_k plus toll or fare payment τ_k:

$$\begin{aligned} p_A &= c_A + \tau_A \\ p_R &= c_R^{user} + \tau_R \,. \end{aligned} \tag{4.40}$$

Then the (interdependent) demand functions for auto and rail can be written as q^k=Q^k(p_A,p_R) and the inverse demand functions as p_k=P^k(q_A,q_R), k=A,R, each giving the marginal willingness to pay for an additional trip. Let $C^A(\cdot)$ and $C^R(\cdot)$ be the total cost functions for auto and rail including user costs for auto and both user and agency costs for rail:

$$\begin{aligned} C_A &= C_A^{users} = q_A \cdot c_A(q_A) \\ C_R &= C_R^{agency} + C_R^{users} = q_R \cdot c_R^{agency}(q_R) + q_R \cdot c_R^{user}(q_R) \end{aligned}$$

Note that the average agency and user costs, c_R^{agency} and c_R^{user}, play important roles in transit finance and in user satisfaction, respectively. Both may be decreasing functions of q_R due to scale economies, as derived in Chapter 3. We costs on the two modes are independent of each other.

Social welfare can be defined as benefits less costs:

$$\begin{aligned} W &= (B_A + B_R) - (C_A + C_R) \\ &= \left(\int_0^{q_A} P^A(\tilde{q}_A,0)d\tilde{q}_A + \int_0^{q_R} P^R(q_A,\tilde{q}_R)d\tilde{q}_R \right) - \left(C_A(q_A) + C_R(q_R)\right) \end{aligned} \tag{4.41}$$

First, consider the first-best solution in which both auto and rail are be priced optimally. The first-order conditions for maximizing W are:

$$0 = \frac{\partial W}{\partial q_A} = P_A(q_A,0) + \int_0^{q_R} \frac{\partial P^R}{\partial q_A} d\tilde{q}_R - mc_A = p_A - mc_A \tag{4.42a}$$

$$0 = \frac{\partial W}{\partial q_R} = p_R - mc_R \,. \tag{4.42b}$$

The last step in (4.42a) uses the approximate symmetry of the Jacobian matrix of the inverse demand functions (valid if income effects are small, due to the symmetry of the Slutsky matrix) to write $\partial P_R/\partial q_A = \partial P_A/\partial q_R$, which makes its integral equal to p_A-$P_A(q_A,0)$.

Thus, the first-best optimum is attained when price is set to marginal cost in each market. As we have seen in Chapter 3, this means that auto users "pay" their own user cost c_A plus a toll that reflects marginal external cost of congestion, $\tau_A = mc_A - c_A = q_A \cdot c_A'(q_A)$; and transit users pay less than the average agency cost due to economies of scale in transit provision.

Now consider the second-best solution, where the auto toll is fixed at zero. We could solve this by adding to (4.41) a Lagrangian term, $\mu\cdot[c_A(q_A)- P^A(q_A,q_R)]$. But it turns out to be easier to reformulate the objective as a function of generalized prices instead of quantities. To do this, we first subtract user payments and the value of user-supplied inputs from the first term in parentheses, and add them to the second; this changes each of the integrals in (4.41) to an area below the inverse demand curve but above a line at price p_k. We then rewrite the integrals by integraing horizontally instead of vertically, which means using the direct demand functions $Q^k(\cdot)$:

$$\int_0^{q_k} \left(P^k(\cdot) - p^k\right) d\tilde{q}_k = \int_{p_k}^{\infty} Q^R(\cdot) d\tilde{p}_R \,.$$

The result is that W is now rewritten as the sum of consumer surplus and profits:

$$\begin{aligned} W &= \left(CS_A + CS_R\right) + \left(\Pi_A + \Pi_R\right) \\ &= \left(\int_{p^A}^{\infty} Q^A(\tilde{p}_A, p_R) d\tilde{p}_A + \int_{p^R}^{\infty} Q^R(\infty, \tilde{p}_R) d\tilde{p}_R \right) + \left(\tau_A q_A + \tau_R q_R - C^R_{agency}\right). \end{aligned} \tag{4.43}$$

Our path of integration is now $(p_A,p_R)\rightarrow(\infty,p_R)\rightarrow(\infty,\infty)$.

We want to maximize W subject to the definitions (4.40), and to the constraint τ_A=0. Solving (4.40) for τ_A and τ_R and substituting into (4.43) gives

$$W = \int_{c^A(q_A)}^{\infty} Q^A(\tilde{p}_A, p_R) d\tilde{p}_A + \int_{p^R}^{\infty} Q^R(\infty, \tilde{p}_R) d\tilde{p}_R + p_R Q_R\left(c_A(q_A), p_R\right) - C_R(q_R) \tag{4.43a}$$

The first-order condition is then:

$$0 = \frac{\partial W}{\partial p_R} = \left[\int_{p_A}^{\infty} Q_R^A(\tilde{p}_A, p_R)d\tilde{p}_A - q_A \cdot c'_A(q_A) \cdot Q_R^A\right] - \left[Q^R(\infty, \tilde{p}_R)\right]_{p^R}^{\infty} + \left[q_R + p_R Q_R^R\right] - \left[C'_R(q_R) \cdot Q_R^R\right]$$

where each term in square brackets is derived from one of the four terms in (4.43a). We invoke $Q_R^A = Q_A^R$, so the first integral is $Q^R(\infty,p_R)$-q_R, which cancels two other terms in the equation. We can assume that $Q^R(\infty,\infty)$=0 (given finite budgets) and we further note that $q_A c'_A$ is the marginal external cost of an automobile trip, mec_A. Thus:

$$-mec_A \cdot Q_R^A + p_r \cdot Q_R^R - C'_R \cdot Q_R^R = 0$$

or

$$p_R = C'_R + mec_A \cdot \frac{Q_R^A}{Q_R^R} = mc_R - mec_A \cdot \left(\frac{\partial q_A}{\partial q_R}\right)_{p_A}. \qquad (4.44)$$

The optimal generalized price of rail is set equal to marginal cost plus a downward adjustment for the effect of rail price on automobile congestion. This latter adjustment is computed as the marginal external cost of an automobile trip times the proportion of new transit trips (i.e. those induced by a change in transit price) that are diverted from automobile. The downward adjustment is large if auto congestion is very costly at the margin, if the lower transit fare is effective at luring automobile drivers to transit, and if the lower fare does *not* induce too many brand-new trips (since the latter are accommodated at below marginal cost but have no offsetting congestion-reduction benefits).

Equivalently, (4.44) can be written in terms of the second-best bus fare as:

$$\tau_R \equiv p_R - c_R^{user} = \left(mc_R - c_R^{user}\right) - mec_A \cdot \left(\partial q_A / \partial q_R\right)_{p_A}$$

or in terms of a transit subsidy,

$$\sigma \equiv c_R^{agency} - \tau_R = \left(ac_R - mc_R\right) + mec_A \cdot \left(\partial q_A / \partial q_R\right)_{p_A}. \qquad (4.45)$$

Equation (4.45) shows clearly the two sources of second-best transit subsidies in our model. The first is scale economies: because average cost exceeds marginal cost (for agency and user costs combined), it is desirable to subsidize the difference. The second is automobile congestion: insofar as lowering transit price is effective in reducing congestion costs by drawing away automobile users, it is desirable to use subsidies to encourage that result.

This type of model is presented more rigorously and fully by Glaister (1974), who also considers two time periods with cross-period demand substitutability. Glaister finds that under plausible conditions, the automobile externality may be so strong as to warrant reverse peakload pricing on transit: i.e., setting the fare lower during the peak than during off-peak, even higher agency expenses and capacity constraints.

This and other models of second-best transit pricing have been used to investigate optimal transit subsidies numerically for specific cities. They typically encompass some but not all pertinent factors such as variation in conditions over times and locations, substitutability of demand across times and locations, transit-agency operating policies, externalities from transit vehicles, and crowding on transit vehicles. As a result, results vary greatly. For example, Glaister and Lewis (1978) estimate optimal rail and bus fares for London at about 50-60% of marginal operating costs.[26] Viton (1983) finds optimal fares for the San Francisco Bay Area and for Pittsburgh to be virtually zero. Winston and Shirley (1998) find quite the opposite for the United States as a whole, with optimal bus and rail fares covering 84% and 97% of marginal operating costs, respectively. For Brussels and London, Van Dender and Proost (2004) estimate optimal transit fares to be nearly zero in peak periods, yet double the current fares in off-peak periods. Probably these conflicting results will be resolved only by building a model that incorporates all the factors considered separately in the various studies, and applying it consistently to several cities.

4.5.2 Fare Structure

A complete schedule of marginal-cost prices would distinguish many trip characteristics including distance, time of day, direction, and density of loadings and boardings. Mohring (1972) provides a comprehensive analysis. In practice, only time of day and trip distance are normally considered as potential bases for price differentials, and even they are often ignored for simplicity. Cervero (1986) claims that the handful of U.S. transit operators that charge peak-hour premia, in order to reflect the higher vehicle and operator costs attributable to peak operations, reap substantial efficiency and financial benefits from doing so.

---

[26] From line 3b of their Table 4.

The fare structure may also be designed to pursue distributional goals, with narrowly defined population subgroups often receiving discounts. As argued by Starrs and Perrins (1989), this is a better approach to distribution than subsidizing transit fares across the board, since many transit users are well off financially – especially users of high-quality radial commuting services.

4.5.3 Incentive Effects of Subsidies

In practice, most transit systems in the world are subsidized, whether for reasons discussed here or for other reasons.[27] However, subsidy programs inevitably have rules that distort the decisions of the transit operators.

One such distortion occurs in programs that subsidize capital but not operating costs. Unsurprisingly, recipients tend to use a higher ratio of capital to other inputs than would be cost-minimizing. For example, Armour (1980) calculates that an 80-percent federal capital subsidy cuts in half the bus retirement age that minimizes local costs in Seattle. Frankena (1987) verifies the same effect empirically for municipal bus systems in Ontario, Canada, by observing the effect of a province-wide capital-subsidy program on scrappage rates. He also finds that an accompanying monitoring program, designed to prevent this result, was ineffective.

Another form of capital bias can occur in the choice among types of transit. It is widely believed that in the U.S., at least, capital subsidies have encouraged local authorities to build rail systems, which are very capital-intensive, in locations where corridor volumes do not justify them. Interest in these systems persists, even for small metropolitan areas, despite evidence of extremely high costs compared to buses, and a record of severely overpredicting demand and underpredicting cost (Pickrell, 1989; Flyvbjerg *et al.*, 2002, 2006).

The capital bias could be reduced by subsidizing operating costs as well. But increasing the total subsidy has its own incentive problems. Several studies have found that subsidies cause costs to increase.[28] This appears to be largely due to labor costs, as higher wages and lower

---

[27] For example, the European Commision's UNITE project computed cost-recovery ratios (fare revenues divided by operating costs) averaging 50% for ten European nations in 1998, varying from 25% for Italy to 91% for The Netherlands (UNITE, 2003).

[28] For example, Anderson (1983) and Savage (2004).

productivity absorb much of the subsidy funds (Pickrell, 1983). Another effect of US subsidy programs has been to encourage inefficient expansion of service to low-density suburbs.[29]

These results are discouraging for the prospects of achieving optimal pricing of public transit. We can offer two responses. First, subsidy programs can be designed to minimize adverse incentives, for example by making the subsidies a fixed proportion of fare revenues or basing them on ridership. Second, the low price-elasticity of transit use (typically measured at –0.3 to –0.4) mitigates the force of our first argument for subsidies, that of scale economies. Indeed, evidence on privately provided transit service suggests that unsubsidized transit is already viable in many markets. Simulation studies suggest it would become much more so if congestion and parking were priced anywhere near their marginal cost, and in that case the other argument for subsidies (to relieve traffic congestion) would also be diminished or eliminated. Thus the solution to institutional difficulties with transit subsidies may be to let prices rise for both transit and competing forms of urban transportation.

4.6 Conclusions

How important is it to "get the prices right"? And what does it mean? Two recent policy statements offer perspectives on what we can learn from the concepts reviewed in this chapter. Both define "right" as reflecting marginal social costs.

The European Commission's transport research program (EXTRA, 2001) suggests that the concept of marginal-social-cost (msc) pricing "can be translated into practical pricing or taxation measures using existing technology. Moreover, simple 'second best' approaches such as cordon tolls … can achieve nearly as much as a theoretically optimal solution." It also states that "pricing measures are effective in changing people's behaviour and travel patterns."

Delucchi (2000) argues that msc pricing is good for economic efficiency, but that its effects would be too modest to solve pressing problems of traffic congestion and ballooning transit deficits. Furthermore, there are other important social concerns such as "distributive fairness, equal opportunity, uncertainty and risk, ecological stability, future generations, [and]

[29] Meyer and Gómez-Ibáñez (1981), Pucher (1984).

quality of life" that may alter the prescription. He also regards certain second-best constraints as barriers to implementation of msc pricing.

Our analysis suggests that the concept of msc pricing is very flexible and can rigorously account for many factors that make naïve calculations of msc unsuitable. It is by now well accepted that social costs formerly thought to be unquantifiable, such as pollution and safety externalities, can be incorporated. We have shown here that many feasibility constraints can also be accommodated, at some cost in complexity and data requirements. Furthermore, tractable models can be built that account for decisions, such as trip scheduling, that have often been taken as given; and doing so considerably changes the nature of optimal pricing. Finally, we find that expanding slightly the bounds placed by certain heuristic constraints, such as equal tolls at all entry points through a cordon, can allow second-best pricing to reap major welfare gains, although perhaps not quite the size suggested by EXTRA.

This view of msc pricing defines marginal cost broadly, well beyond the naïve value that we have called "quasi first-best." Theoretically it includes not only externalities but any effects of a particular priced activity that tend to exacerbate the welfare losses caused by constraints. For example, deficit finance may require higher labor taxes and thereby worsen their adverse effect on labor supply; this raises the marginal cost of any subsidized transit trip and so increases the prescribed msc price. Analytically, it doesn't matter whether such considerations are counted as part of marginal social cost or whether they cause optimal price to diverge from marginal social cost; the end result is the same.

Does pricing change people's behavior? We agree with Delucchi that under currently foreseeable constraints, msc pricing will not eliminate congestion and transit deficits. And we certainly agree that it will not, nor should it, reverse the long-standing trend of growing reliance on automobile travel. However, we believe it can substantially tame these problems by reducing trip times, increasing their reliability, reducing air pollution and traffic accidents, reducing transit deficits, and facilitating targeted improvements to transit service. The problems are significant, so the potential role of pricing is important.

Our views are well within the range of professional opinion of economists. However, it would be a mistake to view such opinion as monolithic. Consider, for example, congestion pricing. Nearly all economists would argue that some type of congestion pricing would be good policy. But there is a wide range of opinion about specific features: how complex to make it,

how politically feasible it is, how well it would guide investment, whether it should be accompanied by privatization, its distributional effect, and how revenues should be spent (Lindsey, 2006). In our view, such diversity is a healthy sign that economics does not neglect practical considerations but rather is struggling with the best way to account for them rigorously.

Figure 4.1 Optimal tolls in static models: smooth average cost (left panel) and piecewise linear average cost (right panel)

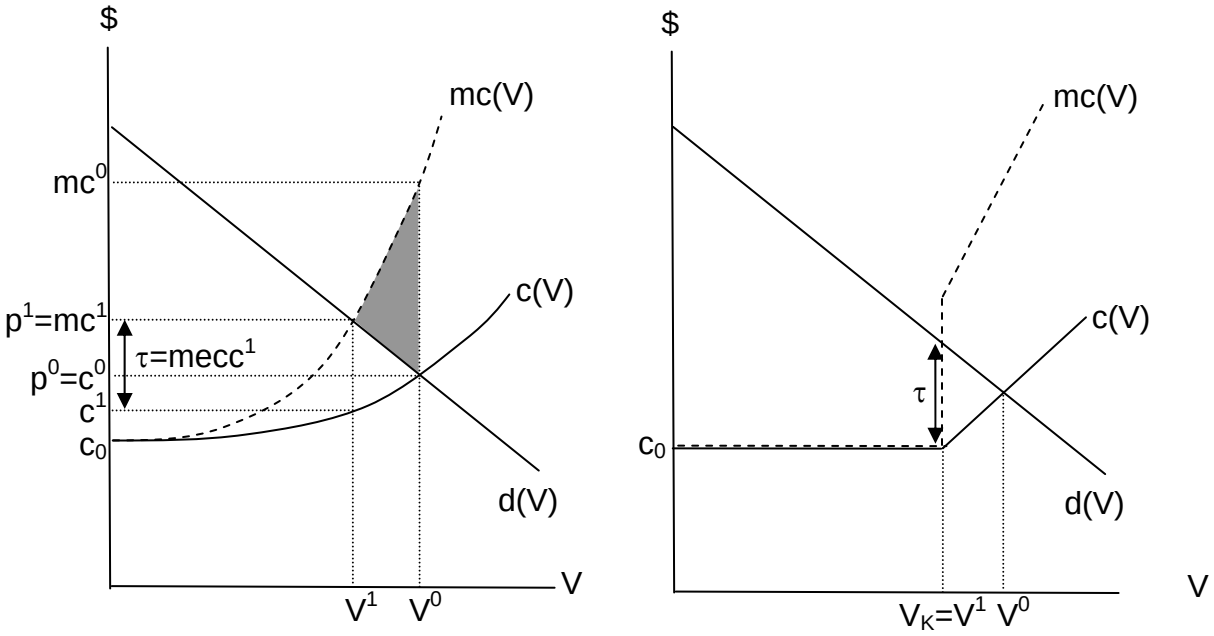


Figure 4.2 Average cost components and optimal tolls by queue-exit time in the basic bottleneck model for the unpriced equilibrium and under first-best tolling

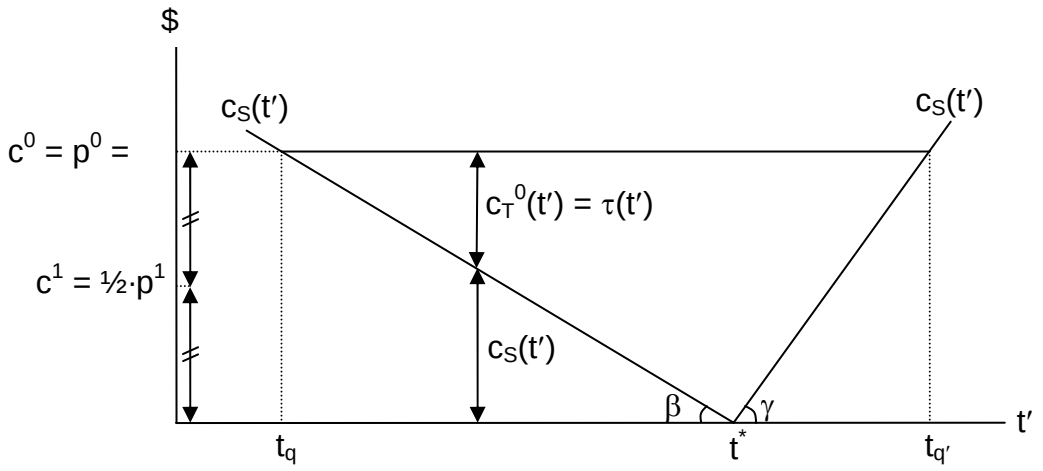


Figure 4.3 Queue entries, desired and actual queue exits

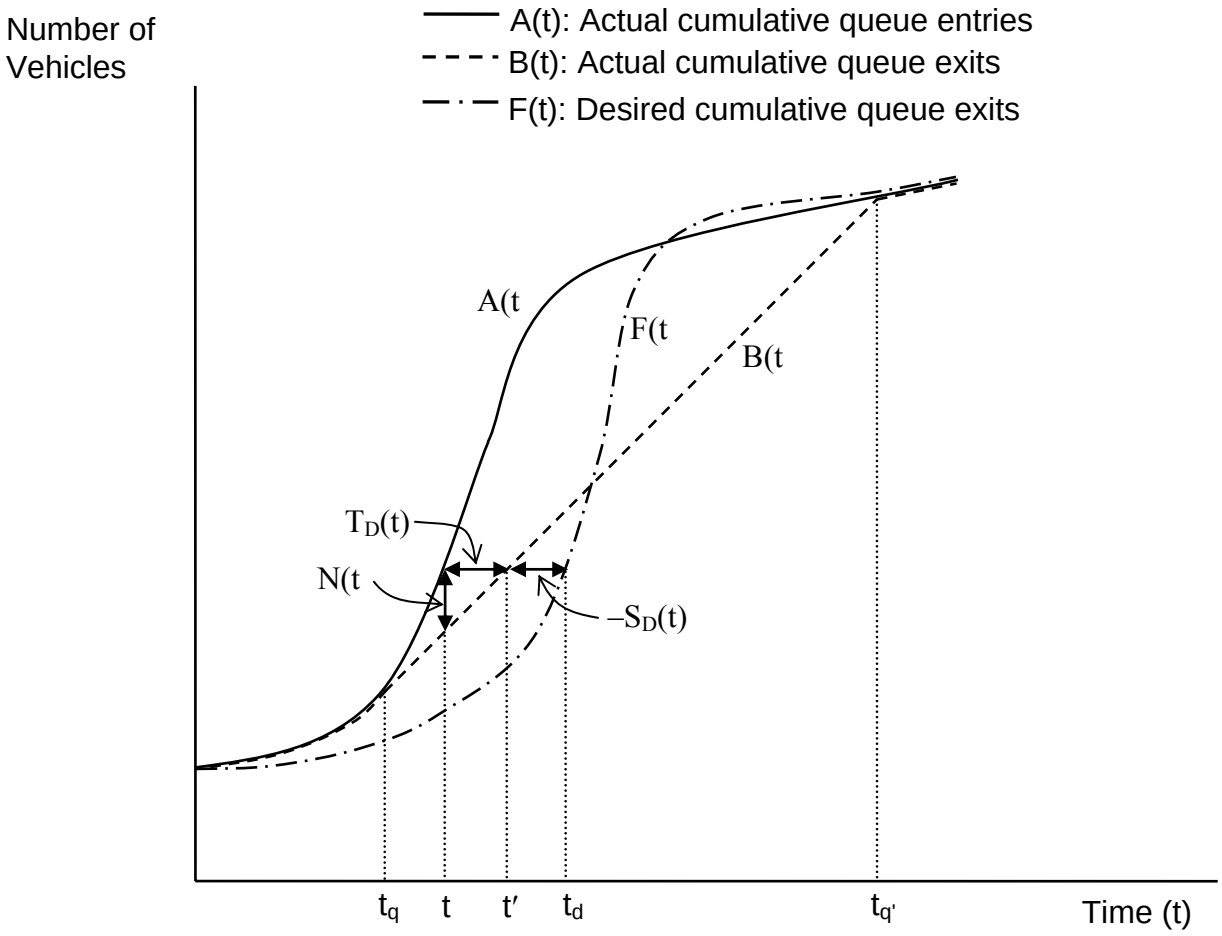


Figure 4.4 Optimal tolling in the bottleneck model with heterogeneous users

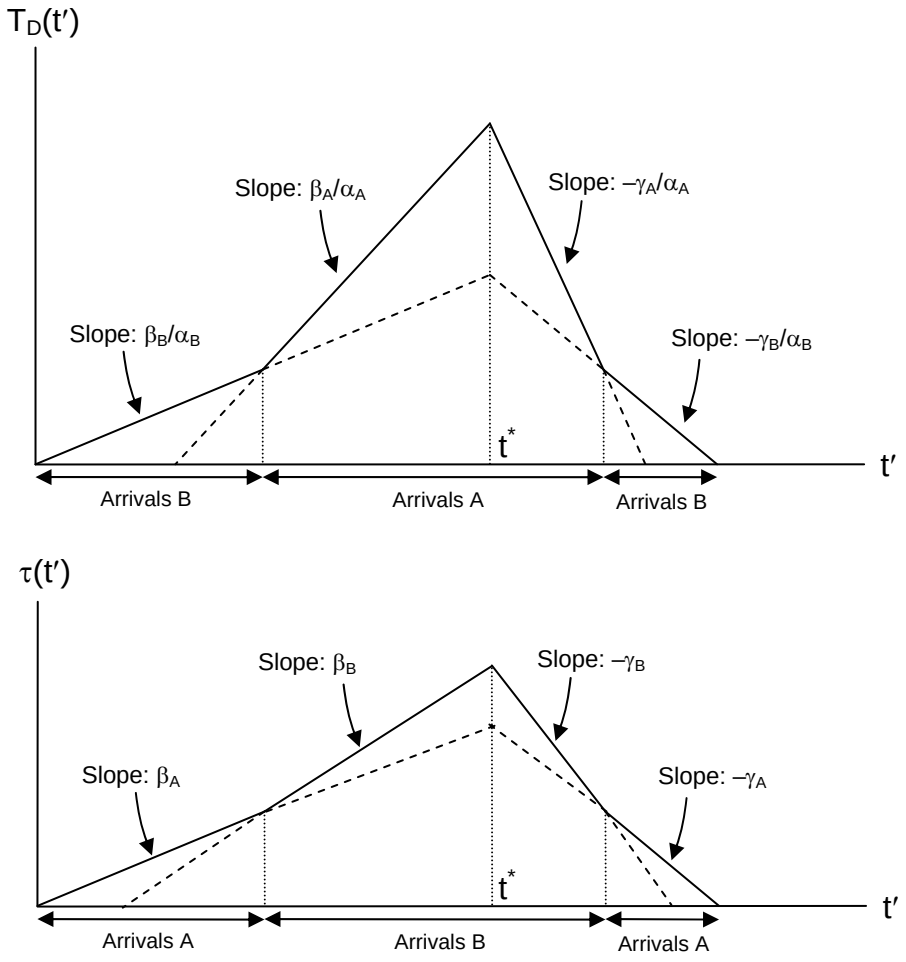
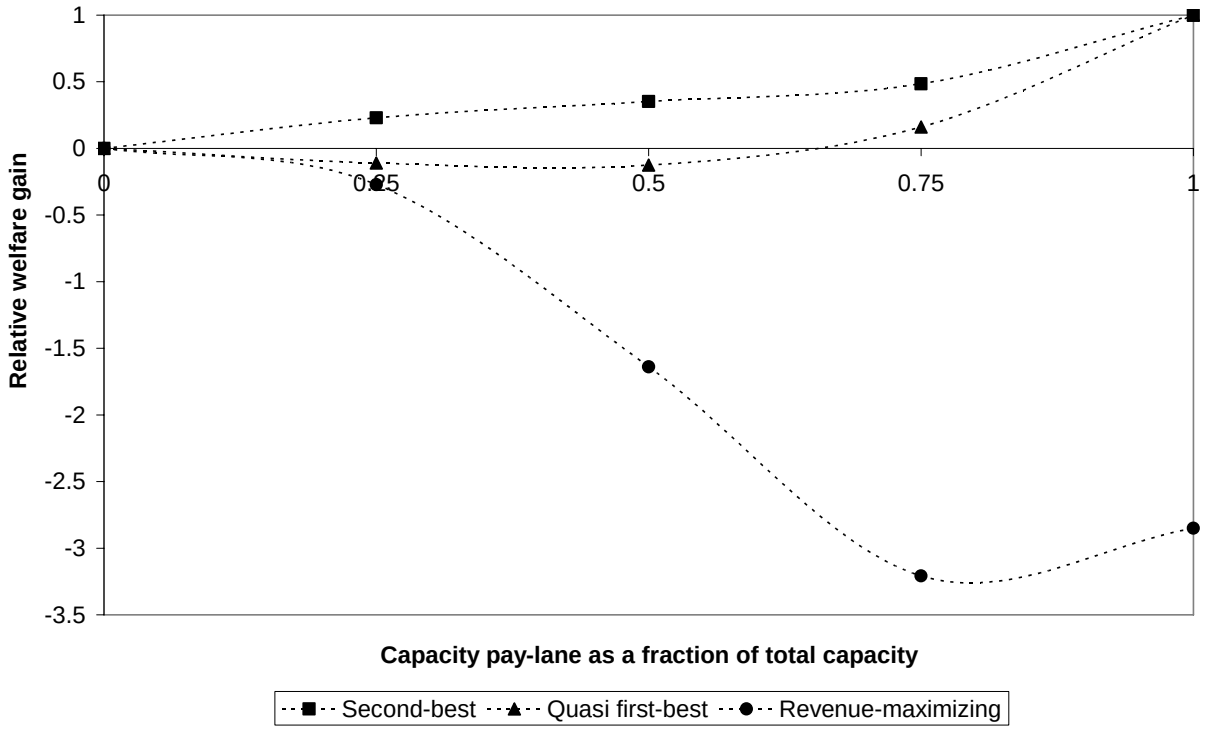


Figure 4.5 Relative efficiency of three pay-lane pricing regimes



Source: Computed using the base-case numerical model of Verhoef and Small (2004). The "quasi first-best" case presents previously unpublished calculations.

5. INVESTMENT

While pricing as discussed in the previous chapter may be the economist's knee-jerk reflex to transportation problems externalities, most policy makers appear to consider capital investment as a more natural response. From the economic perspective, pricing and investment are both important instruments in managing transportation. We can ask: what are the consequences of optimizing investment, with or without pricing in place?

This chapter adds choice of capital stock to our analytical framework. In the terminology introduced in Chapter 3, this means moving from a short-run to a long-run analysis. Section 5.1 discusses capacity choice for highways. We focus on the interactions and interrelations between pricing and capacity choice, and between toll revenues and capacity cost, under first-best and second-best conditions. We also address the perennial question: is it possible to build our way out of congestion? Section 5.2 extends the framework to lumpy investments or other discrete policy changes, introducing a very general technique known as cost-benefit analysis.

5.1 Capacity Choice for Highways

We return to the framework of Section 4.1, in which we sought to maximize the difference between benefits and cost, and now apply it to the choice of highway capacity as well as short-run usage. In this way we can generalize the capacity results of Section 3.5, obtained as part of finding long-run cost functions, to more complex situations. Treating optimal pricing and capacity within the same framework also makes transparent the close relationship between revenues and capacity costs, a relationship of great importance for financing capacity. It also reflects that from the economic perspective, pricing decisions and capacity choices can be evaluated in essentially the same way, namely in terms of their contribution to social welfare.

We first re-derive the rule for determining first-best optimal highway capacity under optimal road pricing, and discuss the relationship between capacity cost and the revenues from optimal pricing — first within a basic model and then considering more complex settings. Next

we turn to second-best capacity choice when pricing is constrained to be sub-optimal. We will see, for example, that with underpricing, the threat of "induced traffic" causes a downward adjustment of second-best capacity. We also consider the impacts of unpriced congestion elsewhere in a network upon second-best capacity choice for a priced link. We conclude this section by discussing the potential effects of using "naïve" investment rules that ignore behavioral responses to capacity expansion.

5.1.1 Basic Results: Capacity Choice with First-Best Pricing and Static Congestion

A basic difference between pricing and capacity choice is that road capacity cannot be varied over the day. We therefore consider the multi-period static model of Section 4.1.1 as our starting point.

Optimizing Capacity

We earlier defined social welfare $W=B-C$ for this model as equation (4.13), which we now rewrite making explicit its dependence on road capacity V_K:

$$W = \sum_h q_h \int_0^{V_h} d_h(v)\mathrm{d}v - \sum_h q_h V_h \cdot c_h(V_h;V_K) - \rho \cdot K(V_K)\,. \tag{5.1}$$

Earlier, we maximized W with respect to the vehicle flows V_h and obtained the marginal-cost pricing rule (4.14) for each time:

$$\tau_h = V_h \cdot \frac{\partial c_h}{\partial V_h}\,. \tag{5.2}$$

Now, we also maximize W with respect to capacity V_K to obtain an investment rule. Doing so is identical to minimizing C, since B (the first term in W above) does not depend on V_K. This is therefore the case already considered when deriving the long-run cost function of Section 3.5, except that here we do not limit $K(V_K)$ to be linear. The first-order condition for maximizing W therefore yields the same optimal investment rule derived before as equation (3.44):

$$\rho \cdot K'(V_K) = -\sum_h q_h V_h \cdot \frac{\partial c_h(\cdot)}{\partial V_K}. \tag{5.3}$$

The marginal cost of adding capacity is equated to the resulting marginal user-cost savings due to from lower congestion. These cost savings are calculated holding flows V_h constant: the envelope theorem assures this is correct because we are dealing with a marginal change from a first-best optimal starting point. That is, because marginal benefits and marginal social costs of flows are equalized though optimal pricing according to (5.2), any indirect marginal benefits or costs of capacity changes via traffic flows have zero net welfare effects. We will see later, in Section 5.1.3, that such indirect effects do have to be taken into account when first-best pricing is not in place.

Self-financing of capital cost

The congestion fees derived in Section 4.1 may be viewed as charges for the use of capacity, which is scarce because it is expensive. It is natural, then, to ask whether the fees will bring in enough revenue to cover the cost of capacity.

By combining investment rule (5.3) with pricing rule (5.2), we can relate total revenue to total cost. To do so, we need first to relate $\partial c_h/\partial V_h$ to $\partial c_h/\partial V_K$. This is easy for any model in which c_h depends only on the volume-capacity ratio V_h/V_K, a condition interpretable as *constant returns to scale in congestion technology*. Under this assumption we can use the quotient rule and chain rule of differentiation to show that:

$$V_K \cdot \frac{\partial c_h(\cdot)}{\partial V_K} = -V_h \cdot \frac{\partial c_h(\cdot)}{\partial V_h},$$

which is also an example of Euler's theorem.[1] Under optimal pricing and capacity choice as defined by equations (5.2) and (5.3), we get the following equation for total revenue R:

[1] Euler's theorem states that if a differentiable function $c(\mathbf{x})$ of variables $\mathbf{x}\equiv(x_i)$ is homogeneous of degree k, i.e. if $c(t\mathbf{x})=t^k c(\mathbf{x})$ for all positive scalars t, then $\Sigma_i x_i \cdot (\partial c/\partial x_i) = k \cdot c(\mathbf{x})$. In this case, c_h is homogeneous of degree $k=0$ in its two arguments.

$$R \equiv \sum_h q_h V_h \tau_h = -\sum_h q_h V_h V_K \cdot \frac{\partial c_h}{\partial V_K} = V_K \rho \cdot K'(V_K) \,. \tag{5.4}$$

This can be simplified by using the economies-of-scale indicator s_K introduced in Section 3.5.3, which was defined as the ratio of average to marginal capacity cost:

$$s_K = \frac{K(V_K)}{V_K \cdot K'(V_K)} \,. \tag{5.5}$$

Using (5.5), we find that (5.4) implies the following "degree of self-financing":

$$\frac{R}{\rho \cdot K(V_K)} = \frac{1}{s_K} \,. \tag{5.6}$$

The ratio of total revenue to total capacity cost is equal to $1/s_K$, which is also the elasticity of capacity cost with respect to capacity. We refer to this as the *self-financing result*.

Optimal fees therefore exactly cover the cost of providing capacity if there are *neutral scale economies in capacity provision* (s_K=1); we call this equality *exact self-financing*. A deficit arises under economies of scale (s_K> 1), and a surplus under diseconomies of scale (s_K< 1). This was demonstrated for transportation by Mohring and Harwitz (1962), and is simply an extension of the result in equation (3.4) to this more complex formulation of capacity costs and (congestion) pricing. Note that the self-financing result applies no matter how $K(V_K)$ is constructed, so long as s_K is defined as above. For example, $K(V_K)$ may or may not entail constant input prices. Of course, for the result to represent actual financial balance for a highway provider, $K(V_K)$ must reflect the provider's actual costs.

The benchmark result of exact self-financing when s_K= 1 thus requires three technical assumptions to be fulfilled: (I) constants returns to scale in the congestion technology; (II) neutral scale economies in capacity provision; and (III) perfect divisibility of capacity. How likely are these assumptions to be satisfied? As far as assumptions (I) and (II) are concerned, one can freely choose the units of capacity, which means that what eventually matters is only the combined effect of the two assumptions. Our earlier discussion in Section 3.5.3 suggests that overall, there are probably mild economies of scale in major cities, which may disappear altogether in very large cities. Thus the degree of self-financing in (5.6) may be close to one.

Condition (III) seems unrealistic on the face of it, given that roads have an integer number of lanes. But, as noted in Section 3.5, capacity can be fine-tuned in various ways such as widening lanes, adding shoulders, or straightening the road, so it is less discrete than might first appear. Nevertheless, as we shall see shortly, the exact results of this section break down when capacity is imperfectly divisible, although the amount of deviation from exact self-financing is not necessarily very large as long as conditions I and II remain fulfilled.

Implications of the Self-financing Result

It is useful to note some policy implications of the self-financing result, before we describe how it holds up under various generalizations of the model. Our discussion focuses on situations where assumptions I – III are reasonable approximations so that exact self-financing applies.

First, when applicable, the exact self-financing results provides a necessary condition for optimal capacities and prices — namely, financial balance when capital costs are properly accounted for — that is more readily observable than variables entering the first-order conditions for optimal capacities and prices. Thus it provides a practical check on whether the road system is efficient. Second, the self-financing result implies that other taxes are not needed to sustain the road sector, which is good for economy-wide efficiency because other taxes are likely to cause distortions. Third, it promotes public acceptability of road pricing because road finance can be perceived as fair (since roads are paid for by users) and transparent (since it depends on money flows). Finally, under certain conditions, the exact self-financing result provides an iterative way for a road authority (or private providers) to achieve first-best optimal capacity: namely, by expanding road capacity whenever short-run optimal congestion pricing yields revenues that exceed the incremental capital cost of capacity.[2] The market thus indicates whether expansion is

[2] To see why, observe that for a given demand function, the short-run optimal congestion toll and the associated road use per unit of capacity are both decreasing in capacity. This can be verified graphically in the left panel of Figure 4.1, by imagining how a larger capacity would imply equality of $d(V)$ and $mc(V)$ at a some price p lower than p^1; this in turn would imply an equilibrium user cost c lower than c^1, hence an equilibrium ratio V/V_K lower than V^1/V_K and a lower optimal toll τ. With exact self-financing for the optimal capacity and toll, short-run optimal toll revenues per unit of capacity therefore exceed the unit cost of capacity with a below-optimal total capacity, and fall short of it with an above-optimal total capacity.

socially warranted, just as in competitive private markets. The proviso of optimal pricing is an important one, however; unless there is perfect competition among road operators, an unlikely scenario, they will have incentives to price according to other objectives, which destroys the equality between optimal revenues and capacity cost.

However it would be an erroneous to conclude that when assumptions I–III are fulfilled, all revenues from optimal congestion tolling should be used to finance further capacity expansions. This interpretation of the self-financing result confuses current investment expenditures with the annualized cost of capital. Furthermore, the annualized cost of capital coincides with financial outlays only if the financing mechanisms for capital recovery exactly reproduce social costs, which is typically not the case due to taxes on capital, explicit or hidden capital subsidies, financial mechanisms for risk-sharing, inadequate accounting standards, separation of ownership and control of private corporations, and even financial fraud.

Highway improvements designed to increase free-flow speeds or improve safety, independent of congestion, do not engender congestion fees under optimal pricing and therefore are not included in the self-financing result. But many such improvements also increase capacity. It is possible that a substantial portion of highway investment in rural or even suburban areas falls into this category, and hence should be subsidized (Larsen, 1993). Unfortunately, this question has received little attention. Jansson (1984, ch. 10) models it as a type of scale economy, while Larsen treats it as a quality variable that is produced jointly along with capacity. To better understand this issue, we need empirically estimated models in which safety and free-flow speed, as well as capacity, are affected by capital investment. Such models might reveal that uniform design standards, such as characterize the US Interstate Highway System, result in excessive free-flow speeds in urban areas where their production is presumably more expensive.

5.1.2 Self-Financing in More Complex Settings

This section considers a number of complications in which the self-financing result carries over in some form, either exactly or as an approximation under certain conditions. Our discussion

follows and sometimes draws from earlier reviews such as Lindsey and Verhoef (2000), De Palma and Lindsey (2005), and Verhoef (2005).

Discrete Capacity

Especially for smaller roads, assumption III involving continuous capacity will often be unrealistic. Discreteness of capacity generally causes the self-financing result to break down. This is most easily seen for a road for which demand is so small that it will never be congested once constructed at minimum feasible capacity. The revenues from optimal road pricing are then zero, and the road cannot be self-financing.

Figure 5.1 illustrates the problem more generally. The heavy solid line marked *lratc* shows the long run average total cost, where the "total" means that it includes both average user cost c and per-user capacity cost $\rho \cdot K/V$ (with K optimized, in a discrete fashion, with respect to V). This *lratc* curve is the lower envelope of many U-shaped short-run average total cost $atc(l)$ curves, each valid for a different discrete number of lanes (denoted l). A different short-run marginal cost function $srmc(l)$ applies for every segment of the *lratc* curve, and each $srmc(l)$ curve cuts the *lratc* curve through its relevant local minimum. In any short-run optimum, the optimal toll is the difference between *srmc* and c (with c not drawn in the Figure), and the per-user capacity cost is the difference between *atc* and c. Thus if V is such that $srmc < atc$ at the optimum, a deficit occurs; if $srmc > atc$, there is a surplus. Only by coincidence would the inverse demand function (also not drawn in the figure) cut the relevant *srmc* curve at its intersection with the associated *atc*, so that exact self-financing would apply.

FIGURE 5.1

Nevertheless, if the general trend of *lratc* is neither up nor down, we have a discrete analog of neutral scale economies and it is more or less equally probable to have a deficit or surplus with optimal pricing and capacity. In that case, the self-financing result will approximately hold under certain conditions likely to prevail in practice. First, if number of

possible capacity values is large and the difference between them small, then *lratc* will approach a horizontal line and self-financing will approximately apply. Second, if demand grows steadily over time, periods of deficit and surplus will tend to alternate, causing the discounted net deficit or surplus to be small. Third, on a network of many roads, deficits and surpluses on individual roads can be pooled, causing most of them to cancel. These considerations make the discreteness issue less important in practice than would appear from Figure 5.1.

Short-run Dynamics

We next consider dynamic congestion. Does the self-financing result of equation (5.6) remain intact? Arnott and Kraus (1998a) have shown that this is indeed the case under rather general conditions.

The basic bottleneck model for a single period, extended to incorporate elastic demand, can illustrate this. Recall from equation (4.25) that the average user cost under optimal pricing for this model amounts to $\bar{c}_g^1 = \bar{c}_S^1 = \frac{1}{2}\delta \cdot Q / V_K$. This equation is proportional to total number of trips Q, implying that the congestion technology exhibits constant returns to scale. The appropriate social objective can next be written as a single-period variant of (5.1), with q_h normalized to 1 and Q replacing V_h. Because the average optimal time-varying toll $\bar{\tau}$ is equal to $\bar{c}_g^1$, it is also equal to $Q \cdot \partial \bar{c}_g^1 / \partial Q$. Equation (5.2) therefore remains valid (with adapted notation). Because (5.3) also remains valid (again with adapted notation), the self-financing result of equation (5.6) again applies.

Surprisingly, the result goes through even if the time pattern of tolls is not optimal, so long as the toll level is optimized subject to whatever constraint on the time pattern applies. Recall from equation (4.36) that second-best optimal flat pricing for the basic bottleneck model with elastic demand can be determined by solving its reduced-form time-independent representation as if it were a static model with an average user cost function $\bar{c}_g^0 = \delta \cdot Q / V_K$. Again this function exhibits constant returns to scale in the congestion technology, and the result is derived just as for the case of an optimal time-varying toll. The same is true for coarse tolls. We see, then, that

the essential aspect of a pricing scheme that makes the self-financing result applicable is not that it prices be first-best, but rather that prices equal marginal costs. As explained in Chapter 4, flat pricing of a bottleneck involves a toll equal to the marginal external cost given that departure times are not affected by the toll. It thus qualifies as marginal cost pricing, although it is second-best in nature. To summarize, "if a road system should be self-financing when a sophisticated tolling system is employed, it should also be self-financing when only [a crude tolling system such as] a flat parking fee is applied" (ADL, 1993, p. 173).

Long-run Dynamics

Arnott and Kraus (1998b) consider a variety of situations when capacity need not be fixed over the period of analysis, accounting for depreciation, maintenance, adjustment costs, and irreversibility of investments. They find that the self-financing theorem remains valid in present value terms, provided that the size of capacity additions is optimized conditional on the timing of investments. This holds irrespective of whether capacity is added continuously or in discrete units, and whether or not the timing of investments is optimal.

Networks

Yang and Meng (2002) demonstrate that if the self-financing result holds for every individual link in an optimally priced network, it holds also for the network as a whole, despite the demand interdependencies across links that result from user equilibrium. This is another example of the envelope theorem at work: we can ignore the indirect effects of policy variables (link tolls, link capacities) upon other markets (other links) in welfare analysis provided optimality conditions on these other markets are fulfilled. In other words, one can then analyze each link individually, and derive (5.6) for it, without worrying about network spillovers. Section 5.1.3. will show that this is no longer true when some links are not optimally priced.

Heterogeneity

Arnott and Kraus (1998a) address heterogeneity across users and find that, as long as marginal cost pricing applies to all users, it does not undermine the self-financing theorem. But if practical considerations prevent tolls from being differentiated optimally among users, for example with flat pricing in the model of Figure 4.4, the self-financing result generally breaks down.

User heterogeneity in relation to capacity choice also raises the question of whether segregation of traffic onto different roads or lanes is desirable. Such segregation generally becomes more attractive as cross congestion effects (congestion that one group inflicts on another) become stronger relative to own congestion effects (congestion inflicted on members of one's own group). If cross effects are high in all directions, appropriate segregation can arise spontaneously, but it they predominate in only certain directions regulation or pricing may be required to bring about an optimal degree of segregation. Consider, for example, the case where cars travel faster than trucks, due for example to safety regulations or vehicle capabilities. Separation may then be desirable, depending on discreteness of capacity and any resulting cost premium. But these cross effects are asymmetric, since passenger cars prefer to avoid sharing lanes with trucks but not vice versa. Thus lane restrictions for trucks, truck-only roads, or lane-specific truck tolls may be desirable in order to enforce separation.

Another possibility is to segregate users by value of time, enabling the planner to optimize toll and capacity separately for two facilities. Simulations by Verhoef and Small (2004) and Small, Winston, and Yan (2006) suggest that the benefits of such segregation are negligible if first-best pricing can be practiced, but can be substantial if one is restricted to no pricing or to second-best pricing of just one roadway (i.e. priced express lanes). Small, Winston, and Yan also find that segregation permits a compromise pricing scheme, involving differentiated tolls on both roadways, that achieves higher welfare than second-best pricing of an express lane, with much lower direct impacts on users (measured as consumer surplus loss before accounting for revenue uses). All these results depend critically on the existence of substantial differences among users in value of time, differences which seem to exist empirically both in the Netherlands and in southern California according to empirical evidence presented in the above-mentioned papers.

Yet another possibility is to segregate vehicles with different numbers of passengers by reserving certain lanes for high-occupancy vehicles (HOVs). These lanes may be on the highway itself or on metered entry ramps. Mohring (1979) and Small (1983a) estimate the welfare gains from such restricted lanes on city streets and on expressways, respectively, finding substantial welfare gains under somewhat ideal conditions. Dahlgren (1998), however, finds that only in rather exceptional circumstances does an HOV lane outperform a general-purpose lane: namely, when there is high initial congestion and when the initial proportion of vehicles that are HOVs falls within a rather narrow range. Small, Winston, and Yan (2006) get more favorable results for HOV lanes, but they also depend strongly on the factors identified by Dahlgren and on relatively highly elastic demand. The results stated above assume no pricing is possible; if instead marginal-cost pricing is in effect, then the welfare gain from segregating HOVs is essentially negligible (Small, 1983a; Yang and Huang, 1999).

Road Maintenance

So far, we have ignored road damage, road maintenance cost, and the choice of durability (*e.g.* thickness of the pavement) in construction. These questions are treated in some detail by Small, Winston, and Evans (1989) and De Palma and Lindsey (2005). Here we focus on implications for self-financing.

Newbery (1989) considers whether the combination of optimized congestion charges and road-damage charges would cover the cost of constructing and maintaining roads, taking into account maintenance costs due to pavement damage and the extra construction costs undertaken to limit such damage. Newbery shows that self-financing holds in this broader sense, provided that heavy vehicles affect user and maintenance costs on the entire width of the road uniformly. (This extra condition may be thought of as an extension of neutral scale economies of road use). Under these circumstances, capital costs, even those incurred to make the road stronger, are proportional to capacity and so are recovered through congestion charges just as in the basic self-

financing theorem.[3] Maintenance costs incurred due to heavy vehicles (e.g. periodic repaving) are recovered through road-damage charges, following the same logic —road strength is optimized by balancing the extra capital cost with the maintenance-cost savings. User costs due to road damage (e.g. damage to vehicle tires) do not create an additional externality on average over a network, given certain rather strong technical assumptions about traffic growth and the criteria for undertaking maintenance; therefore they create no additional charges or revenues under optimal pricing (Newbery, 1988b).

Small, Winston and Evans (1989) reach a similar conclusion empirically, without adopting Newbery's assumptions. They treat congestion and road damage in a multi-product framework, where the products are number of trips and the total weight of their loads (more precisely, the number of standardized units of road damage done by the heavy loads). Even though they find substantial economies of scale in providing for each product separately, this is balanced by diseconomies of scope in that the presence of heavy trucks makes it more expensive to handle large volumes of passenger cars. (Diseconomies of scope imply that if there were no indivisibilities, it would be cheaper to segregate trucks and cars.) Overall, neutral scale economies approximately hold.

Variable Input Prices

Contrary to the conventional assumptions, in urban areas the supply function of land for roads (i.e., its price as a function of amount purchased) is likely to be upward-sloping, since taking more land for roads drives up its scarcity value for other uses. The distinction between returns to scale (a property of production functions) and economies of scale (a property of cost functions) then becomes important. A general-equilibrium analysis by Berechman and Pines (1991) demonstrates that constant returns to scale in the production function for roads implies that optimal revenues equal imputed costs, which include the amount of land multiplied by its current price. For a road authority that has to take prices as given, imputed land costs would coincide

[3] They also cover any portion of maintenance cost that is non-allocable to usage, which enters the model just like capital cost, so long as these are proportional to capacity.

with actual land costs. But a road authority with market power in the land market will account for the rising supply function in its cost function, which will then show a degree of scale economies, s_K, that is less than the degree of returns to scale in production. Small (1999a) shows that optimal revenues are then determined by degree of scale economies, s_K, of the actual cost function, just as in equation (5.6). Thus self-financing still applies: if s_K=1, revenues equal costs when prices and capacity are optimal.

This point has practical relevance. As discussed in Section 3.5.3, many studies ignoring the rising supply price of land find economies of scale in road building. To the extent that these economies are offset by the rising supply price of land, the end result is closer to one where self-financing applies.

Other Externalities

The self-financing result equates revenues from congestion charges to capacity costs. That equality remains even if other road charges are levied to cover externalities such as traffic accidents, noise, and air pollution. These charges do not change the user-cost or capacity-cost functions, so the relationships already established for congestion remain valid although they will now describe equilibrium at a different (lower) traffic volume. Thus if conditions I-III hold, optimal congestion charges exactly cover capacity costs and the revenues from other charges create a financial surplus.

5.1.3 Second-best Highway Capacity

Capacity choice under first-best pricing provides an important benchmark, but in practice investment decisions are made under more constrained conditions. We now ask how this affects capacity choice and the degree of self-financing. We first consider sub-optimal pricing on the road whose capacity is being chosen, and then on other roads within a network.

Capacity Choice with Sub-optimal Pricing

Several authors consider capacity choice in a second-best world where the optimal congestion fee cannot be charged.[4] The problem can be analyzed within the multi-period static model by fixing arbitrarily determined tolls τ_h^A for each period *h*. The objective function (5.1) is then augmented by adding a Lagrangian term for this constraint, becoming:

$$\begin{aligned}\Lambda = &\sum_h q_h \cdot \int_0^{V_h} d_h(v)\mathrm{d}v - \sum_h q_h V_h \cdot c_h(V_h;V_K) - \rho \cdot K(V_K) \\ &+ \sum_h \lambda_h \cdot \left[c_h(V_h;V_K) + \tau_h^A - d_h(V_h)\right] \ .\end{aligned} \tag{5.7}$$

The first-order conditions for maximizing (5.7) with respect to V_h, V_K and λ_h can be solved to yield the following second-best investment rule:

$$\begin{aligned}&\rho \cdot K'(V_K) = -\sum_h \left(q_h V_h - \lambda_h\right) \cdot \frac{\partial c_h}{\partial V_K}, \qquad h = 1,...,H \\ &\text{with}: \ \lambda_h = q_h \cdot \frac{V_h \cdot \dfrac{\partial c_h}{\partial V_h} - \tau_h}{\dfrac{\partial c_h}{\partial V_h} - d_h'(V_h)}.\end{aligned} \tag{5.8}$$

The second-best policy rule deviates from the first-best rule unless the Lagrange multipliers are all zero. That would require, in each period *h*, that toll τ_h^A be set optimally or that demand be perfectly inelastic ($d_h' = -\infty$). We assume finite d_h' for purposes of discussion. Then if τ_h is below its optimal value, λ_h must be positive and the marginal benefits of capacity expansion are calculated as if fewer than $q_h V_h$ travelers benefit from the expansion. (With overcharging, the opposite occurs.) This reflects that with undercharging, the social benefit of an additional user is smaller than the private benefit because of the congestion externality. Limiting capacity is one way to discourage this additional user from entering the road.

[4] Examples include Henderson (1977, chap. 7), Wheaton (1978), Wilson (1983), and D'Ouville and McDonald (1990b).

While the Langrangian multipliers in equations (5.8) reduce second-best capacity below its first-best value at any given value of flows $\boldsymbol{V}$, those flows are of course different from their first-best values. Typically in the period of undercharging, the flow will be greater than the first-best value, which has the opposite effect on (5.8).Therefore, whether in the end second-best capacity is smaller or larger than first-best capacity depends on demand and cost elasticities and their impact on equilibrium use levels (d'Ouville and McDonald, 1990b). Still, Wheaton (1978) shows that if the toll starts at its first-best level and is reduced marginally, optimal capacity increases — at least in a single-period model. Wilson (1983) derives more general conditions, which he judges to be reasonable, under which second-best road capacity with underpricing exceeds the first-best level. Thus it seems likely that one should compensate for lack of pricing by building more and bigger roads. Or to put it another way: optimal pricing would probably allow us to get by with less pavement.

The self-financing result of equation (5.6) generally breaks down under arbitrary pricing and second-best capacity choice. This is of course obvious for the case where tolls are zero.

Induced Traffic

Whether or not second-best capacity is greater than first-best, equation (5.8) shows clearly that a naïve application of cost-benefit analysis to an incremental capacity increase will mislead. Capacity should not be expanded just because travel-time savings to existing users would be valued at more than the cost of expansion. Rather, the desirability of such expansion is limited by the fact that it reduces the equilibrium generalized price and therefore attract new traffic, known as *induced traffic* or *induced demand*.[5]

The problem is especially acute if demand is highly elastic. To see why, consider again Figure 4.1a, illustrating equilibrium congestion in a simple static model. If the inverse demand

[5] These terms, plus "induced travel" and "latent demand," tend to be used synonymously. Lee (1999) suggests the following useful distinction: induced traffic is a change in traffic resulting in movement along a short-run demand curve, whereas induced demand is a shift in the short-run demand curve (perhaps also a movement along a long-run demand curve). Here "short-run" should be defined as a time period over which demand factors, such as vehicle fleet or land uses, are kept constant. We do not attempt here to maintain this distinction rigorously, partly because we use "latent demand" for a particular kind of potential induced traffic.

curve were nearly flat, neither shifting it to the left (demand-limiting policies) nor shifting the cost curve to the right (capacity-augmenting policies) would make much difference to the equilibrium average cost. The second-best investment rule in (5.8) reflects this ineffectiveness as follows: with perfectly elastic demand and zero tolls, $\lambda_h = q_h V_h$ and the right-hand side of (5.8) — which expresses the marginal benefit of capacity expansion — is zero. The reason is that any expansion will be filled up with new traffic whose marginal benefit (height of the demand curve) equals its average cost, so that the new traffic yields no net gain in social surplus. Furthermore, inframarginal traffic does not benefit because with perfectly elastic demand, equilibrium average cost does not change.

Even in less extreme situations, induced traffic in a situation like that of Figure 4.1a represents the release of potential traffic that is deterred by congestion itself. Such potential traffic, often called *latent demand*, consists of people who, because of congestion, now choose an alternative route, mode, time of day, or home or workplace location, or do not travel at all. Unfortunately, latent demand is prevalent in just those areas where capacity expansion seems most needed — namely, in high-density urban areas during times when congestion is severe. Under such conditions, the release of latent demand is likely to undo much of the congestion relief that capacity expansion or demand reduction might otherwise bring about.

The undesirable filling up of new unpriced capacity by induced traffic is sometimes called the "law of highway congestion." It forms the basis for important debates concerning highway investment policy.[6] While not usually true in its extreme form (where new capacity is completely filled by induced traffic), it can come uncomfortably close to the truth, as argued persuasively by Downs (1962) and Thomson (1977). Smeed (1968) states that in British cities, "the amount of traffic adjusts itself to a barely tolerable speed" (p. 41); he estimates that "if it were not for the inhibiting effects of congestion, we might well have 4 to 5 times as much traffic in Central

[6] This "law" exemplifies a more general principle, known as the "tragedy of the commons," applying to any public good or environmental amenity that is available at little or no charge and whose quality deteriorates with intensity of use (Hardin, 1968).

London as we have now" (p. 58). More formal modeling of the phenomenon is exemplified by Holden (1989).

Empirical evidence for induced traffic is considerable. According to Kroes, Antonisse, and Bexelius (1987), a telephone survey in urbanized western Holland measured latent demand for automobile travel during the evening peak period at 27 percent of current volume. Mackie and Bonsall (1989) give several examples of new roads whose traffic far exceeded that attributable to simple route shifts. A report by the Standing Advisory Committee on Trunk Road Assessment (SACTRA, 1994) caused a major rethinking of road-expansion policies in the UK by demonstrating empirically that traffic responds significantly to road capacity.

More recently, several econometric studies have measured the elasticity of traffic with respect to road capacity. (A positive elasticity indicates some induced demand, and an elasticity of 1.0 would imply the fundamental law of highway congestion.) The best studies account, among other things, for the possibility that road capacity is endogenous (with road authorities assumed to respond, logically enough, to actual or anticipated traffic); accounting for endogeneity somewhat reduces the measured elasticity, but still it seems to be quite high. Goodwin (1996) reviews several studies with estimates of short- and long-run elasticities averaging 0.20 and 0.77. Noland (2001) estimates short-run (five-year) and long-run elasticities of around 0.5 and 0.8, whereas Cervero and Hansen (2002) find an elasticity of 0.79 that applies over a six-year period.[7]

As already noted in connection with Figure 4.1a, the problem of releasing latent demand occurs not only with policies that shift the cost curve to the right, but also with those that shift the demand curve to the left. Thus demand management policies and improved transit, when implemented as measures to relieve congestion, are also vulnerable to induced traffic. Empirical evidence supports this conclusion. Sherret (1975) analyzes trip rates on the Bay Area Rapid

[7] By contrast, Mokhtarian *et al.* (2002) find no evidence of induced traffic when analyzing nine matched pairs of California road segments. They speculate that the reason might be that induced traffic is greater for new roads than expansion of existing roads, or it may tend to reflect increases in trip length more than in number of trips. Their measures would miss these types of induced demand. However they also think that econometric studies using aggregate data are plagued by omitted variables that may bias their results.

Transit (BART) line between Oakland and San Francisco during the first few months after it opened, and on the parallel San Francisco-Oakland Bay Bridge; he finds that the diversion of 8,750 automobile trips to BART was soon followed by the generation of 7,000 new automobile trips, so that "traffic levels at the busiest hours showed only small reductions" (pp. xii-xiii).

One reason for high latent demand during peak periods is substitutability among travel at different times. For example, nearly half the latent demand reported by Kroes *et al.* was from "travel taking place at times other than the desired time" (p. 237). Other evidence comes from before-and-after studies of major capacity changes. The opening of a section of the M25 London Orbital Motorway apparently caused peak narrowing at a main Thames River crossing (Mackie and Bonsall, 1989, p. 415). When a bridge opened near Vianen in The Netherlands, morning peak-period traffic on the corridor feeding it increased by 33 percent, one-fourth of which was attributed to shifts in travel schedules — presumably causing peak narrowing (Kroes *et al.*, 1987, pp. 236-237). By the same token, peak periods tend to spread when demand grows faster than road capacity. In London between 1962 to 1981, a time when downtown traffic increased dramatically with little new street capacity, midday speeds actually became slower than peak speeds (Mogridge *et al.*, 1987, p. 297). A shifting peak is, of course, precisely what is predicted by equilibrium models with endogenous scheduling such as we reviewed in Section 3.4.3. Obviously the phenomenon reduces the attraction of static models that assume pre-defined periods of fixed duration with constant demands.

In our example using Figure 4.1a with a perfectly elastic demand curve, capacity expansion has no benefits. In more realistic cases, there are benefits from expansion because some of the induced traffic previously used other roads or time periods that were underpriced. Relieving conditions at those alternate locations or times may confer substantial benefits to other travelers. Thus, for example, if a road expansion causes the peak period to narrow, those who travel at adjacent times may see big reductions in travel times.

Some people think the "fundamental law of traffic congestion" negates all policies aimed at reducing traffic, including pricing. But this is not true even if the "fundamental law" is true, because pricing operates on the demand and supply curves differently that other policies. Again

returning to Figure 4.1a, pricing does not shift the demand or cost curves to the right or left, but rather inserts a wedge between the cost curve and the generalized price perceived by users. Thus it does not attract latent demand and it is able to reduce congestion and produce welfare gains. Furthermore, if pricing is in place, capacity expansion is no longer frustrated by latent demand, as can be seen from the investment rule in (5.8): the closer τ_h^A approaches its optimal value, the smaller is λ_h and therefore the higher are the marginal benefits from capacity expansion for given flows. Thus pricing is especially well suited to overcome the policy limitations imposed by induced traffic.

Network Spillovers

Another type of second-best situation occurs when it is not the toll on the road itself that is imperfect, but the tolls on other roads in the network. We consider the case where tolls on other roads are completely absent, and discuss how this affects a road's second-best optimal capacity, as well as the degree of self-financing. It turns out that the answer depends on whether the unpriced capacity is parallel to or in series with the road whose capacity is under consideration. (More generally, the question is whether traffic on the unpriced links is a substitute or complement to that on the link in question.) We therefore consider two extreme cases, defining for each a simple two-link network in which both toll and capacity can optimized for one link (*T*), while an unpriced link (*U*) of fixed capacity exists either parallel to or in series with link *T*. We simplify by considering a single period only.

Consider first the case where the links are parallel. We already treated pricing for this case using the Lagrangian function (4.32); we need only modify it to include capacities and their costs:

$$\begin{aligned}\Lambda = &\int_0^{V_T+V_U} d(v)\,\mathrm{d}v - V_T \cdot c_T(V_T;V_{KT}) - V_U \cdot c_U(V_U;V_{KU}) - \rho \cdot K_T(V_{KT}) - \rho \cdot K_U(V_{KU}) \\ &+ \lambda_T \cdot \left[c_T(V_T;V_{KT}) + \tau_T - d(V_T+V_U)\right] + \lambda_U \cdot \left[c_U(V_U;V_{KU}) - d(V_T+V_U)\right]\end{aligned} \quad . \tag{5.9}$$

where V_{KT} and V_{KU} are the capacities of the two links (with V_{KU} fixed). The first-order conditions

(with respect to V_T, V_U, V_{KT}, λ_U, and λ_T) can be solved to yield:

$$\tau_T = V_T \cdot \frac{\partial c_T}{\partial V_T} - V_U \cdot \frac{\partial c_U}{\partial V_U} \cdot \frac{-d'}{\left(\frac{\partial c_U}{\partial V_U} - d' \right)} \tag{5.10}$$

$$\rho \cdot K_T'(V_{KT}) = -V_T \cdot \frac{\partial c_T}{\partial V_{KT}} \ . \tag{5.11}$$

The second-best toll (5.10) is the same as second-best toll (4.35), while the second-best capacity rule (5.11) is identical to the first-best rule (5.3). Thus the toll on link *T* is set to account for the underpriced parallel link, while the capacity of link *T* is set to minimize the cost incurred by its users.

Because the investment rule for V_{KT} is the same as the first-best rule, we already know that revenues under a first-best toll would balance capacity cost under the conditions for the self-financing result. But revenues under the second-best toll (5.10) are smaller than this. Therefore the self-financing result breaks down. This is basically because the second-best distortion leads to a downward adjustment in pricing compared to marginal external cost pricing.

How does the second-best highway capacity compare to first-best capacity in this example? We are not aware of any systematic analyses of this question. We can say something about the case where capacities on both links can be optimized. Suppose users are homogeneous and the links are equally long; then the optimal capacity of the untolled link is zero so that, effectively, the link is eliminated while the tolled link is expanded to first-best tolled capacity. This result would change with sufficient dispersion in values of time, since it may then be desirable to leave an untolled link available for the lowest-value-of-time drivers. However, the relatively small numerical difference that Verhoef and Small (2004) find between optimal tolls on parallel links for heterogeneous drivers suggests that the second-best capacity for an untolled link would still be very small.

Now supposed the unpriced link *U* is in series with link *T*, with a single origin at one end and destination at the other so that all traffic must use both links. The Lagrangian function is now:

$$\Lambda = \int_0^V d(v)\mathrm{d}v - V \cdot \left[c_T(V,V_{KT}) + c_U(V,V_{KU})\right] - \rho \cdot K_T(V_{KT}) - \rho \cdot K_U(V_{KU}) + \lambda \cdot \left[c_T(V,V_{KT}) + c_U(V,V_{KU}) + \tau_T - d(V)\right] \tag{5.12}$$

The first-order conditions with respect to V, V_{KT}, and λ imply:

$$\tau_T = V \cdot \left(\frac{\partial c_T}{\partial V} + \frac{\partial c_U}{\partial V} \right) \tag{5.13}$$

$$\rho \cdot K_T'(V_{KT}) = -V \cdot \frac{\partial c_T}{\partial V_{KT}} \tag{5.14}$$

Not surprisingly, the toll rule (5.13) perfectly internalizes the congestion externalities on both links jointly, and therefore exceeds the first-best toll for link T whenever link U is congested. The investment rule again has the familiar first-best form. With a first-best investment rule and a toll higher than the first-best value, we can see that now a surplus occurs at link T, compared to the degree of self-financing as given in (5.6).

Unpriced congestion elsewhere in the network therefore does not seem to affect the optimal investment rule for a tolled road. Of course, because flows will differ between first-best and second-best optima, the equilibrium *size* of the second-best capacity for link T is generally different from first-best. Also, the self-financing result breaks down. The two examples above suggest that, for links in larger networks, surpluses would result if unpriced complements dominate, and deficits for unpriced substitutes.

5.1.4 Naïve Investment Rules

We finally address a question of considerable importance to the interpretation of conventional investment analysis, but rarely analyzed. This is the question of whether and how a planner's misperceptions of the true preference structure of travelers, or the true congestion technology, might lead to systematic biases in capacity choice. We discuss two examples. One involves the neglect of induced traffic, the second of rescheduling.

Ignoring Induced Traffic

Applied investment analysis often overlooks the problem of induced traffic. How this affects investment decisions depends on whether the neglect stems from a misperception of the actual demand elasticity or from the erroneous use of the conventional first-best investment rule (5.3) where the second-best rule (5.8) is appropriate.

Let us first consider, as a benchmark, the case with optimal pricing in place — admittedly irrelevant to current practice. In this case, an erroneous mixing up of the conventional investment rule (5.3) and the second-best rule (5.8) would be harmless because, with all $\lambda_h = 0$ in (5.8), the rules are identical. But if demand is erroneously assumed to be highly inelastic, optimal flow and capacity will be underpredicted, leading to an unpleasant surprise once the new capacity is opened. Even so, the mistake could be corrected iteratively by further adjusting capacity to the newly revealed level of traffic demanded, presuming such iterations would be convergent.

Now consider the more realistic case where no toll can be charged and the regulator is aware of the second-best rule (5.8) but erroneously assumes that demand is completely inelastic. Again there will be an unexpected increase in demand after an expansion. There are now two possibilities. The first is that the regulator learns from the previous experience, adjusts the estimate of demand elasticity to the correct value, and ends up in the second-best optimum. The second is that the regulator is unable or unwilling to learn, instead treating the newly observed flow V as an exogenous shock to an inelastic demand function. This regulator ends up in an equilibrium where (5.8) is satisfied under the incorrect assumption that $\partial d_h/\partial V_h = -\infty$, hence $\lambda_h=0$ — that is, the regulator applies the first-best rule because in the mistaken belief that demand is inelastic the first-best and second-best rules coincide. This equilibrium is the same as that attained in yet another case, where the regulator knows about induced traffic but mistakenly applies conventional first-best rule (5.3) to analyze investment. This is perhaps the most common case, since the second-best investment rule is not widely known in practice.

To determine the impacts of mistakenly applying the first-best investment rule, let us define two alternate expressions for the marginal benefit of capacity expansion: the true value MB_K, equal to the right-hand side of (5.8), and the naïve value MB_K^n, equal to the right-hand side

of (5.3). The regulator expands capacity until MB_K^n equals the marginal cost of capacity. But since $MB_K > MB_K^n$ at any given set of traffic volumes, this means the road will be overbuilt for the situation: the last increment of capacity was not worth its cost, given the restriction on tolling.

Ignoring Rescheduling of Trips

Another type of naïve cost-benefit analysis would ignore endogenous scheduling and the implied departure-time adjustments that a capacity expansion may induce. Intuitively, one may expect two errors in opposite directions from this mistake. On the one hand, capacity expansion leads to a stronger concentration of trip completion times (peak narrowing), causing aggregate scheduling cost to fall — a benefit that is ignored in the naïve analysis. On the other hand, peak narrowing limits the ability of the capacity expansion to reduce congestion, so savings in travel-delay cost are likely to be less than predicted in the naïve analysis.

To illustrate, suppose the world operates according to the bottleneck model of Section 3.4.3, with uniformly dispersed desired queue-exit times. What happens when the regulator is ignorant of scheduling costs and tries to optimize capacity subject to a static congestion model such one based on the BPR congestion function of equation (3.9)?

To keep it simple, we assume that total demand Q is fixed, and this is correctly perceived by the regulator. We observe the system in equilibrium when capacity is V_K^0 (assumed less than the desired queue-exit rate V_d), and we wish to assess the benefits of expanding it slightly to V_K. The true marginal benefit is then found by differentiating the average cost $\bar{c}_g$ of equation (3.38) and multiplying by Q:

$$MB_K = -Q \cdot \left. \frac{\partial \bar{c}_g}{\partial V_K} \right|_{V_K = V_K^0} = \delta \cdot \left(\frac{Q}{V_K^0} \right)^2 . \tag{5.15}$$

But under our assumptions, the regulator instead uses a naïve measure of marginal benefit, MB_K^n, computed under the following assumptions:

(i) average congestion cost (call it c_g^n) consists only of travel-time cost;

(ii) for a given demand, c_g^n is inversely proportional to $(V_K)^b$;

(iii) the initial value of c_g^n is the observed average travel-delay cost, which we know from the true model is $\bar{c}_T = \frac{1}{2}\delta \cdot Q / V_k^0$ (equation 3.36).

Thus the regulator uses the following naïve average cost as a function of V_K:

$$c_g^n = \left(\tfrac{1}{2}\delta \cdot \frac{Q}{V_K^0}\right) \cdot \left(\frac{V_K^0}{V_K}\right)^b . \tag{5.16}$$

Differentiating, evaluating the result at $V_K = V_K^0$, and multiplying by Q yields the naïve marginal benefit:

$$MB_K^n = -Q \cdot \left.\frac{\partial c_g^n}{\partial V_K}\right|_{V_K = V_K^0} = -\tfrac{1}{2} b\delta \cdot \left(\frac{Q}{V_K^0}\right)^2 = \tfrac{1}{2} b \cdot MB_K . \tag{5.17}$$

Comparing (5.15) with (5.17), we see that the naïve calculation underestimates true marginal benefits when $b < 2$ and overestimates them when $b > 2$. Which of the two mistakes dominates — ignoring scheduling benefits or overestimating travel time benefits — therefore depends on the curvature of the naïve cost function. This is because the convexity determines how seriously travel-delay savings are overestimated: with small b that error is small and overshadowed by the naïve neglect of scheduling-cost savings, but with large b the mistaken forecast of reduced congestion is the more serious mistake.

It seems quite likely that the error in computing marginal benefits could be large, in either direction. If the conventional BPR value of b=4 is used, marginal benefits would be overestimated by 100 percent. But if the BPR function were fit using observed data generated by the bottleneck model (here assumed to be the true one), we know from equation (3.36) that it would appear linear, i.e. b=1; then marginal benefits would be underestimated by 50 percent.

Other examples can be considered. What if the observed pattern of queue-exit times is thought to be fixed, when it really is determined by the bottleneck model? Small (1992a) finds in

that case that $MB_K^n = \frac{1}{2}(\alpha/\delta)\cdot MB_K$. Using the empirical values (α/β)=1.631 and (α/γ)=0.417 from equation (2.24), $\frac{1}{2}(\alpha/\delta)$=1.02, so the marginal benefit is overestimated by just 2 percent. But Henderson (1992) considers the same question in the context of the no-propagation model already mentioned in Section 3.3.3. He finds that under a mild parameter restriction, the marginal benefit is always overestimated, causing the road to be overbuilt.

It appears, then, that ignoring trip scheduling in investment analysis can cause serious mistakes. But there seems no general rule as to which direction those mistakes will take.

5.2 Cost-Benefit Analysis

The investment analysis we have described thus far depends on the possibility of incremental investments whose purpose is to enlarge a well-defined measure of capital such as "capacity." It also assumes that the relevant benefits and costs can be described as continuous functions. Sometimes these conditions are not met. More generally, we often want to analyze policy initiatives that are arbitrarily defined and may have little to do with any describable optimality conditions. It then becomes useful to have a more general method for comparing benefits and costs of proposed projects.

Cost-benefit analysis is a set of tools for making such comparisons. As its name implies, it focuses on economic effects; thus it is not by itself a complete decision mechanism. Nevertheless it can incorporate many factors that are sometimes considered non-economic — just as we showed in Chapter 3 that cost functions can incorporate air pollution, noise, and risk of injury and death. Indeed, the way that was done illustrates a central principle of cost-benefit analysis: namely, that benefits can be measured as the willingness of individuals to pay for them.

We begin by explaining some basic principles embodied in cost-benefit analysis and why they provide a useful reference point for summarizing project impacts. We then consider just a few of the many measurement issues that afflict cost-benefit analysis. The topic is vast, but our

goal is modest: to help the reader understand the implications of particular analytical choices that have been used in, or are suggested for, particular applications.[8]

5.2.1 Willingness to Pay

The starting point for measuring costs or benefits is *willingness to pay*: that amount of money that an individual or firm could pay after a proposed change and still be equally well off (by his or her own evaluation). This concept incorporates consumer sovereignty, i.e. the belief that individuals are the best judge of the value to them of their consumption decisions. This does not mean that externalities are to be ignored — on the contrary, the willingness-to-pay principle allows one to assess how much people care about relieving the externality. The principle does, however, exclude analysts' or governments' beliefs about the inherent worth of activities unless those beliefs can be articulated in terms of benefits that people, as individuals, will appreciate. For example, fostering a healthy ecology is often thought to have its own moral value; cost-benefit analysis can account for this but only insofar as that value can be translated into privately valued improvements such as better health, more pleasant living conditions, or more reliable resource availability for current and future residents of the earth.

The willingness-to-pay principle has actually been present throughout the analysis of this book. For example, in Chapter 2 we measured the values of travel time and reliability from the way individuals trade them against price in travel-demand models; they are simply travelers' willingness to pay for time savings or reliability improvements. The costs defined in Chapter 3 measure the collective willingness to pay by individuals and transportation providers for any savings in their required inputs of time, fuel, labor, and other things. The area under a demand curve, introduced in Chapter 4 as a benefit measure, is an approximate indicator of travelers' collective willingness to pay for an increase in the quantity of travel.

[8] A fuller treatment is provided by many references including Little and Mirrlees (1968), Mishan (1988), Layard and Glaister (1994), Moore and Pozdena (2004), Small (1999), and Boardman *et al.* (2006). Much of this section is adapted from Small (1999).

Figure 5.2 illustrates the relationship between willingness to pay and the demand curve. The sloping line depicts the demand for bus trips by users with identical values of time (but who differ in other ways, causing some but not all to choose bus at a given generalized price). Suppose a service improvement speeds up the buses and thereby lowers the full price from p^0 to p^1. There are Q^0 existing users, each willing to pay $(p^0\text{-}p^1)$ for the improvement; their aggregate willingness to pay is therefore the rectangular area $p^0\text{AF}p^1$. The improvement also attracts $Q^1\text{-}Q^0$ new users; some (those most easily attracted) are willing to pay almost $(p^0\text{-}p^1)$ for the improvement, while others (those just barely attracted) are willing to pay very little. Adding them together, aggregate willingness to pay for new users is the triangular area ABF. The total willingness to pay is therefore the trapezoid $p^0\text{AB}p^1$, which is the change in consumers surplus (i.e. in the area under the demand curve and above the price).

FIGURE 5.2 (adapt from Small 1999, Fig. 5-1)

Equivalently, we could define net benefits as total benefits (trapezoidal area under the demand curve) less total costs (rectangular area under the price line); this yields the same quantities. The difference is that the willingness to pay is now calculated for the trips themselves, instead of for trips associated with certain costs; the required inputs of users' time and money (as reflected in p^0 or p^1) are then counted as costs to be subtracted from willingness to pay.[9]

The diagram illustrates that benefits cannot all be classified in terms of measurable travel characteristics such as travel-time savings. In this case the benefits to existing users are solely from travel-time savings. But new users did not use bus prior to the change, so we don't know what their prior travel times might have been; in fact some of them did not previously travel at

[9] In this formulation, total benefits of the change are given by area $\text{GB}Q_1\text{O}$ less $\text{GA}Q_0\text{O}$, while total costs are given by $p^1\text{B}Q_1\text{O}$ less $p^0\text{A}Q_0\text{O}$. Their difference is again trapezoid $p^0\text{AB}p^1$.

all. So their benefits are best described simply as gains in consumer surplus, indicating how much they perceive themselves as being better off than in their previous situations.

Some inputs that are part of the users' price p are not freely traded in markets. We have already seen how to deal with these in the case of users' time, for which we imputed a value. But even money costs may not reflect free markets or, more generally, social values. For example, during the oil crises of 1973 and 1979, various price controls on gasoline were imposed in the US, resulting in informal fuel rationing. As a result, many consumers would have been willing to pay more for fuel than they actually paid. In such cases, the analyst may be able to calculate a *shadow price*, different from (in this case higher than) the market price, that reflects willingness to pay. As another example, a project may provide work for people who otherwise would be involuntarily unemployed; those people would have been willing to work for less than the actual wage, so the shadow wage to be applied in calculating the costs of that project would be lower than the actual wage. (A warning, however: this is valid only if the project truly causes a net change in employment, as opposed to a shift of employment from one sector to another.)

Most projects have diverse effects that will benefit some people and harm others. Granted that we can measure each person's willingness to pay (negative if he or she is harmed), why should we be interested in the result of adding them all together? By doing so, cost-benefit analysis can identify projects that are *potential Pareto improvements*, i.e. projects for which winners could compensate losers, leaving all better off. For example, a new highway interchange may save enough people time that they could compensate those harmed by extra noise and traffic. Mechanisms exist for such compensation but they are far from perfect; as a result, few if any projects can be actual Pareto improvements and thereby achieve unanimous support. Indeed, one can argue that it is precisely the absence of such unanimity that creates the need for cost-benefit analysis to aid public decision-making.

Nevertheless, if projects were consistently analyzed and undertaken only when they are potential Pareto improvements, one could expect that the losers from some projects would be winners from others. Given enough randomness in the effects of different projects, there would

be "a strong probability that almost all would be better off" eventually (Hicks, 1941, p. 111). Polinsky (1972) provides one attempt to formalize the kind of randomness required.

5.2.2 Demand and Cost Forecasts

The most critical component of a cost-benefit analysis is simply predicting what will happen. The desirability of a new or improved facility depends especially on how much it will cost and how many people will use it. Yet as already noted in Section 4.5.3, there are wide disparities between forecasted and realized results of transportation infrastructure projects, for both costs and demands. Furthermore, the discrepancies are generally not random or neutral in sign, but rather appear to be strategically related to their use, through cost-benefit or other types of analysis, in decision-making about the projects.

Flyvbjerg, Skamris Holm, and Buhl (2004, 2006) analyze the factors determining the size and direction of forecasting errors based on worldwide samples of over 200 projects. For cost projections, they find that errors grow rapidly with the time it takes to build the project, and for bridges and tunnels they grow more than proportionally to the size of the project. Perhaps surprisingly, they find no evidence that errors are systematically worse with public than private projects; they do seem to be worse, however, for state-owned enterprises that lack financial transparency. For demand projections, they find forecast errors for roads that are significant but evenly balanced in sign; whereas for rail projects, the errors are enormous and positively biased, with 90 percent of projects overestimating demand by an average of 106 percent.

Various proposals have been made to reduce such discrepancies and especially to eliminate the biases that systematically result in poor investments being undertaken. These include performing sensitivity analysis, limiting the time span over which projections are made, and subjecting forecasting procedures to peer review (Pickrell, 1989). Because the biases appear to be strategic, and the main use of cost-benefit analysis is to inform political decision-making, any remedies must be consistent with an overall plan to improve the political process for choosing investments.

5.2.3 Discounting Future Costs and Benefits

We already mentioned how costs or benefits incurred over time can be converted to an equivalent present value or an equivalent constant annualized flow (Section 3.4.6). The basic principle is that any expenditure C undertaken in year t could be funded by investing some smaller amount PV today (year 0) and regularly reinvesting its financial yield. Assuming a constant interest rate r, compounded annually, they are related by $PV\cdot(1+r)^t = C$, i.e. $PV=C/(1+r)^t$. The quantity PV is called the *present value* of the future expenditure. More generally, a series of expenditures C_t in years $t=1,\ldots,T$ has present value

$$PV = \sum_{t=1}^{T} \frac{C_t}{(1+r)^t}. \tag{5.18}$$

The same principle works for benefits if one assumes that people have access to capital markets at interest rate r. This is because if they were required to pay an amount PV in conjunction with receiving a stream of benefits $\{B_t\}$ over time, they could borrow or lend in the capital markets in order to be just as well off in each year as before, provided PV is given by (5.18) with C_t replaced by B_t.

Perfect access to capital markets is far from universal; as a result, many people are willing to trade time against money but at some rate different from (typically lower than) r. Such a rate, called the *social rate of time preference*, may be appropriate for discounting costs or benefits accruing to capital-constrained individuals. Meanwhile, businesses pay taxes on investment returns, sometimes determined under very complex rules, and these taxes cause the return to capital (representing production that people are willing to pay for) to exceed the market interest rate. When a proposed investment displaces private investment subject to such taxes, the appropriate rate for cost-benefit analysis would then be the before-tax rate of return to private capital, sometimes called the *value of marginal product of capital*. Finally, when inflation is expected, the value of future consumption or cost savings brought about by investing capital today will be less than the nominal returns indicated by financial accounts; we account for this

by defining a *real rate of interest* approximately equal to $r=n-\pi$, where n is any particular nominal rate of interest and π is the rate of inflation.[10]

Defining an interest rate, or a set of interest rates, to apply to the elements in a cost-benefit analysis therefore requires considerable sophistication. Unfortunately, the outcome of such an analysis is often very sensitive to the interest rate chosen, especially in the case of long-lived capital investments as are common in transportation. All the more so if benefits are small at the beginning and grow over time, as is typical of new transportation facilities in areas undergoing population growth. Partly because of this, cost-benefit practitioners have often found it easy to manipulate the results to favor an outcome that they desire. As a damper to such practices, government agencies sometimes promulgate rules, or at least guidelines, for how other agencies and private parties should choose interest rates and other key analytical elements when presenting cost-benefit analyses of government projects.[11]

Most commonly, the recommended real interest rate is a weighted average of the marginal rate of time preference in consumption and the pre-tax real value of marginal product of capital. One careful estimate of such a rate, applying to the US in 1989, is 9.6 percent per year (???) (Boardman *et al.*, 2006, p. ???) [Ken will update]. US OMB (2003) instructs government agencies to use a rate of 7 percent unless there is compelling reason otherwise. Transport Canada (1994, p. 66) recommends 10 percent. It is universally recommended that sensitivity analysis be undertaken by recomputing results at different interest rates.

When benefits of a project occur very far in the future, conventional discounting makes them virtually irrelevant. For example, the quantifiable benefits of measures taken now to reduce global warming are small for this reason. Many observers find this situation unpalatable, as it

[10] More precisely, the real interest rate r is implicitly defined by the equation $(1+r)\cdot(1+\pi) = 1+n$. If π is small, the solution to this equation is approximately $r=n-\pi$.

[11] Manuals or directives for practitioners include US OMB (2003), American Association of State Highway and Transportation Officials (2003), European Commission DG Regional Policy (2002), Japan Research Institute Study Group on Road Investment Evaluation (2000), Austroads (1996), and Transport Canada (1994). For access to these and others, see the web site of the Economic Development Research Group, *Specialized Benefit-Cost Guides*, http://www.edrgroup.com/edr1/library/lib_guides_special/index.shtml (accessed July 26, 2006). For a critique of US practices, see US GAO (2005).

appears to ignore the welfare of future generations. Actually it does not, but discounting such benefits does make an implicit assumption that even greater benefits can be produced by using the funds in some other way today. For example, if we forego expensive development of hydrogen vehicles and instead invest in technology for climate manipulation or adaptation, we might be able to counteract or ameliorate the future effects of that portion of greenhouse-gas emissions that would otherwise have been eliminated — leaving future generations doing at least as well as they would if we had developed hydrogen vehicles.

Several arguments can be made for instead using a lower interest rate to discount economic transactions far in the future. One is that the rate of return on capital is likely to decrease in the future as cumulative investment makes capital less and less scarce; thus, we should use an interest rate that declines gradually over time (Dasgupta, 1994). Working against this is a long history of technological progress that has continually widened the scope for productive capital investments; if that can be relied upon, then expected new technologies become yet another reason for not worrying too much about problems we are creating for the distant future.

Weitzman (2001) gives another argument for using a lower interest rate to discount the distant future. There is no single number agreed upon by experts for the interest rate that is appropriate for cost-benefit analysis. Opinion is both varied and skewed, with a long "right tail" (i.e. a few experts think the appropriate rate is quite large). Weitzman estimates the distribution of expert opinion from a survey and uses the results in a Monte Carlo simulation of cost-benefit calculations under alternative interest rates. He finds that the result is the same as he would get using an interest rate that is smaller the farther in the future we look. Thus both Dasgupa and Weitzman, using quite different approaches, conclude that events in the very far future should be discounted using a lower interest rate than that prevailing today, and hence these events will have more influence over present decisions than they otherwise would.

5.2.4 Shifting of costs and benefits

One appeal of cost-benefit analysis is that it can identify the economic actors affected by an improvement and measure the value to each of them, thereby permitting some description of the distribution of these benefit across various groups.

However, price changes in ancillary markets may drastically alter this distribution by shifting benefits or costs from one group to another. A classic example is that land rents and land prices in areas made more accessible will increase through the normal working of land markets, as these locations become more desirable compared to others. Thus a change that appears to create benefits for, say, transit users who can take advantage of a new transit station is more likely to end up as an advantage to landowners in the area. Boyd (1976) gives a cogent account including some specific examples for urban land markets. Other markets that may shift transportation benefits from one party to another are those for labor and for retail goods and services.

5.2.5 External Benefits

Promoters of projects often like to highlight various benefits such as jobs created, real-estate development induced, economic activity attracted, or industrial activities made more efficient by a transportation improvement. Economic analysis has shown that most such benefits are either transfers from other locations (e.g. economic activity moved from one location to another) or transformations of transportation benefits, already measured in conventional cost-benefit analysis, to another form.

To understand such transformations, consider the benefits of "industrial reorganization," by which transportation improvements enable firms to create more efficient systems of distribution or production.[12] Mohring and Williamson (1969) show that these benefits, while quite possibly real, are already captured in the demand curve for transportation that is the basis for conventional benefit measurement. To illustrate, consider again Figure 5.2 and suppose it represents demand for urban freight travel by industrial firms. The ability to reorganize distribution or production through cheaper transportation is represented by the amount the firms

[12] For a rigorous study defining and measuring such benefits, see Shirley and Winston (2004).

are willing to pay for the Q^1-Q^0 new trips that make this reorganization possible. Thus area ABF already captures their value to the firm. Even if the firms are competitive, so that all their savings are shifted to customers or suppliers, the ultimate benefits are still measured by the area under the demand curve for transportation (Jara-Díaz, 1986).

Nevertheless, some benefits may be "external" in the sense that they accrue to people other than the decision-maker responsible for them. Such *external benefits* are likely to be important in at least two situations.

The first is when transportation improvements reduce the market power of firms by promoting trade among previously isolated locations. Jara-Díaz (1986) illustrates this process with a model of two firms, each initially a monopolist within its own region, when a transportation improvement lowers the cost of shipping the good from one location to the other. Because the initial monopoly power caused prices to be higher than optimal, thereby creating a deadweight loss to the economy, welfare is increased when that power declines. Neither firm can capture all this benefit in its own use of transportation, so the demand curve for transportation does not fully reflect it. Indeed, this example amplifies to our comment made in opening this book, that transportation is central to economic activity because it permits trade; we now can say that this benefit of transportation is partly external and so may justify government intervention to promote transportation. We hasten to add, as does Jara-Díaz, that this point applies mainly to developing economies or isolated rural areas where such benefits have already been exhausted.

Second, urban areas are thought to exist because they enable firms to take advantage of *economies of agglomeration*, which are cost reductions made possible when a large number of firms exist in proximity. When an additional firm locates in such an area, it reaps some of these advantages but creates others that are external to itself, by adding to the density experienced by others. This phenomenon involves a reciprocal externality, very much like congestion except it conveys positive rather than negative benefits. Thus, just like congestion, free markets will not produce an optimal amount of agglomeration —they will tend to produce too little.

Any improvement in transportation enlarges the number of firms whose mutual interaction create these economies, so should increase their size. The upshot is that economies of

agglomeration potentially create another external benefit of transportation. Again, we hasten to add that in well-developed urban areas, incremental transportation improvements are likely to have only small effects on urban efficiency. By way of analogy, Fernald (1999) finds evidence that the US Interstate Highway System created substantial productivity gains in its early years, but that more recent additions have had much smaller effects on the US economy.

5.2.5 Conclusion: Use and Misuse of Cost-Benefit Analysis

In the end, cost-benefit analysis is undertaken to inform decision-makers. One can hope that by making the impacts of a project transparent and by quantifying them in consistent ways, the decision-making process — which in most cases is political — can be guided toward decisions that are better for most people. Indeed, this is the main purpose of recent legislation in several nations that subjects new regulations to cost-benefit analysis: the presumption is that exposing any negative economic effects will force regulators to articulate what they are trying to achieve, and allow the public to decide whether it is worth those negative effects.

As a political process, cost-benefit analysis will inevitably be subjected to pressures to misuse it for the benefit of particular interest groups. Professionals can reduce such strategic misuse by developing clear procedural guidelines and articulating them clearly to the public. Furthermore, the cost-benefit analyses themselves need to be transparent as well as technically sound so that such misuse can be identified.

Sensitivity analysis is one good way to be sure that users can understand the implication of particular assumptions made in a cost-benefit analysis. One type of sensitivity analysis, sometimes called *risk analysis*, postulates specific probability density functions for uncertain parameters, then uses Monte Carlo simulation to compute the corresponding frequency distribution of any particular result of interest. There is a risk in formalizing uncertainty in this way, however, because it may lead to a false sense of precision about how uncertainty can be characterized; in many cases, especially developing nations, the outcome may be more affected by administrative competence, sabotage, or breakdown of related markets than by the elements formalized in risk analysis (Jenkins, 1997).

5.3 Conclusions

Long-lived transport investments can be analyzed either as a problem of optimizing over a continuously varying quantity representing capital stock, or as a problem of choice among well-defined discrete packages. If done the first way, the analysis completes the characterization of long-run cost functions. If done the second way, it becomes part of a more general technique, cost-benefit analysis, which can be used to analyze other proposals such as pricing or regulatory changes. Indeed, it has become common for governments to mandate such analyses of a wide range of government action in an attempt to increase the burden of proof on proposed actions that have adverse economic effects.

Like pricing, investment analysis can be considered under first- or second-best conditions. The latter are of course far more prevalent, and in several cases we can identify the way investment rules should be modified to account for constraints, especially constraints on pricing the facility in question or other facilities that interact with it. Furthermore, we can sometimes describe the potential misallocations due to applying first-best rules (which might be thought of as common-sense cost-benefit analysis) to what are really second-best situations.

Public investment decisions are heavily influenced by political considerations. The goal of cost-benefit analysis is not to create rigid criteria that must be followed to the exclusion of all others, but rather to inform decision makers about how a project looks when judged by one set of criteria that are intended to be consistent with widely accepted economic principles. To be useful, these criteria need to be defined by a transparent process, and to this end it is especially important that transportation professionals understand the technical implications of proposed cost-benefit techniques and be able to explain them to others in non-technical language.

Figure 5.1. Surpluses and deficits with discrete capacity

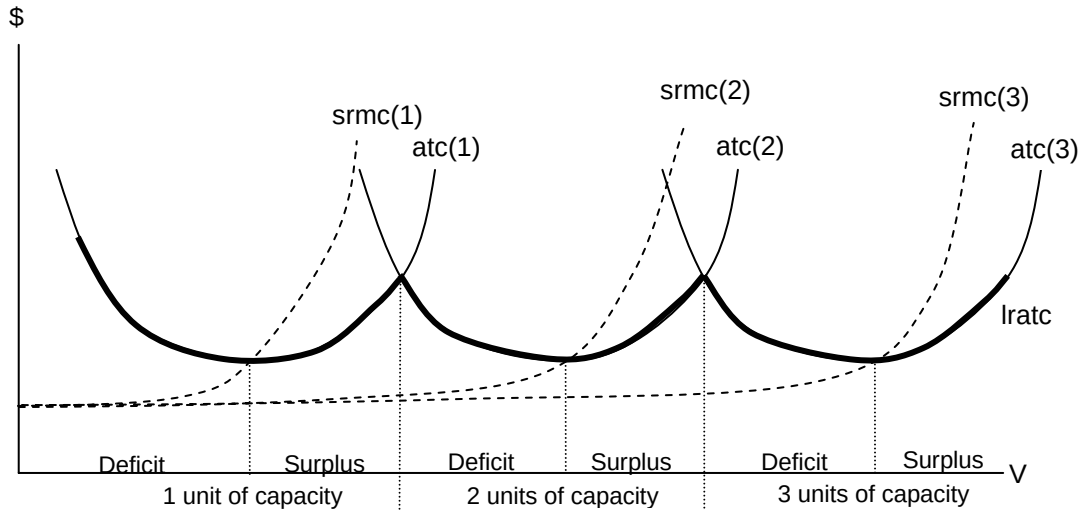
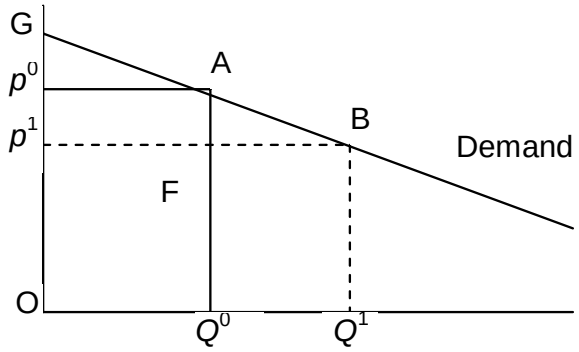


Figure 5.2. Benefits to Existing and New Users



6. INDUSTRIAL ORGANIZATION OF TRANSPORTATION PROVIDERS

So far we have discussed desirable investment and pricing policies, but not what institutional structures can best bring them about. The dominant organizational form for providing urban transportation services to individual users, especially in developed nations, is public ownership. This is supplemented by regulation of those firms allowed to operate privately. Observers in many nations increasingly question the efficacy of these arrangements. This chapter reviews issues related to privatization. Section 6.1 discusses profit-maximizing price and capacity choice for private highway operators, and compares these to welfare-maximizing choices. Section 6.2 discusses regulation and franchising of private highways. The chapter then moves on to considering private transit services in Section 6.3 and paratransit in Section 6.4. The discussion focuses on the deregulation of transit and paratransit services, and will briefly describe the results of some recent innovations in public policy.

6.1 Private Highways

Some observers, such as Geltner and Moavenzadeh (1987), Poole (1988) and Roth (1996), suggest that many of the recent difficulties with financing and pricing publicly owned highways could be overcome through a return to private ownership, which was common in past eras. Private-sector participation in road supply is reviewed by the OECD (1987), Beesley and Hensher (1990), Gómez-Ibáñez and Meyer (1993) and World Bank (2006).

Some of these authors stress congestion pricing as a main advantage of privatization. DeVany and Saving (1980) show that this occurs under perfect competition. However, the pricing and investment strategies of a private profit-maximizing highway owner under the more

realistic conditions of monopoly or oligopoly are, of course, not necessarily optimal.

6.1.1 Static Congestion, Single Road, Multiple Periods

We can see what happens in a private monopoly case by modifying the benefit-cost framework developed in Chapters 4 and 5. Instead of maximizing *B–C*, as in equation (@5.1), an unregulated monopolist chooses capacity and tolls so as to maximize profits Π, equal to its revenues minus its own costs. For the single-road, multiple-periods set-up of (@4.13), we may define:

$$\Pi = \sum_h q_h \cdot V_h \cdot \tau_h - \rho \cdot K(V_K)\,. \tag{6.1}$$

The equilibrium condition from equation (@4.1), equating marginal benefit $d(V)$ to the generalized price $c(V)+\tau$, of course remains valid for each period h. It is convenient to directly substitute this condition into the objective function:

$$\Pi = \sum_h q_h \cdot V_h \cdot (d_h(\mathbf{V}) - c_h(V_h;V_K)) - \rho \cdot K(V_K)\,. \tag{6.2}$$

We maximize (6.2) with respect to capacity V_K and flows V_h. Perhaps surprisingly, maximizing (6.2) with respect to capacity V_K reproduces the first-best condition encountered in (@5.3):

$$\rho \cdot K'(V_K) = -\sum_h q_h \cdot V_h \cdot \frac{\partial c_h(\cdot)}{\partial V_K}\,. \tag{6.3}$$

The investment rule (6.3) implies that for given flows V_h, the monopolist chooses capacity to minimize total social cost, including user cost. This is important as it shows that the monopolist is cost-conscious, even with those resources supplied by its customers. The intuition is that for any given equilibrium demand level V_h and corresponding equilibrium generalized price level $d_h = p_h = c_h + \tau_h$, the monopolist can increase toll revenues without losing demand when he succeeds in reducing the average user cost c, and replacing this reduction by an equivalent increases in the

toll τ. Every dollar reduction in total user cost can therefore be turned into an extra dollar of toll revenues for given flow levels. The monopolist therefore faces the optimal incentive to minimize the sum of user cost and capital cost, as is also required for the optimization of welfare.

The resulting capacity would be identical to the first-best capacity if the investment rule (6.3) were evaluated at first-best demand levels. This, however, is not generally the case. Instead, maximizing (6.2) with respect to V_h yields the first-order condition:

$$\tau_h = V_h \cdot \frac{\partial c_h}{\partial V_h} - V_h \cdot \frac{\partial d_h}{\partial V_h} - \sum_{i \neq h} \frac{q_i}{q_h} \cdot V_i \cdot \frac{\partial d_i}{\partial V_h}. \tag{6.4}$$

The first term on the right-hand side is equal to the first-best toll in (@4.14) and (@5.2), and therefore shows that the profit-maximizing toll internalizes the congestion externality. However, two extra terms are added that take into account demand elasticities (note that both terms are positive overall with downward-sloping demand functions). When demands in different time periods are independent of each other, (6.4) can be simplified to:

$$\tau_h = V_h \cdot \frac{\partial c_h}{\partial V_h} - V_h \cdot \frac{\partial d_h}{\partial V_h} \quad \Leftrightarrow \quad p_h \cdot \left(1 - \frac{1}{\varepsilon_h}\right) = mc_h, \tag{6.5}$$

where ε_h is the absolute value of the own-period price-elasticity of demand. Equation (6.5) looks like the familiar monopoly rule equating marginal revenue to marginal cost, but here the price and marginal cost both include user costs. As usual with monopoly solutions, it is valid only for elastic demand; *i.e.*, $\varepsilon_h > 1$.

This result is equivalent to that of Mohring (1985) for a monopolist owner of a congested port. Mohring's result (his equation 4, p. 31, for one time period only) applies to a demand curve stated as a function of fee τ rather than of price p. The equivalence between the two approaches is easiest demonstrated for a single-period setting and by writing the inverse demand function as perceived by the monopolist, relating τ to V, as $\tau(V) = d(V) - c(V)$. With total revenues defined

as $V\cdot\tau(V)$ and marginal cost for the monopolist equal to zero, profit maximization requires revenue maximization or $\tau(V) + V\cdot\tau'(V) = 0$. Given the earlier definition of $\tau(V)$, this implies $\tau = V\cdot(c'-d')$, the single-period equivalent of (6.5).

This also provides the intuition behind the monopolist internalizing the congestion externality. Because the toll is equal to the difference between inverse demand and average user cost, rising average user cost affects the marginal revenue in exactly the same way as a falling inverse demand function does in conventional uncongested monopoly models (in absence of price discrimination). The two terms $V\cdot c'$ and $-V\cdot d'$ therefore enter the optimal pricing rule symmetrically. They jointly represent the loss in revenues extracted from inframarginal users when lowering the toll such as to allow one additional user on the road. The marginal external cost term reflects one among two reasons for this loss: the average user cost is rising – besides the usual reason that marginal willingness to pay is falling in output.

The monopolist therefore takes marginal social cost into account in setting price, and although adding a mark-up, thus may be said to practice a form of congestion pricing. Note that the bracketed term in the second equation in (6.5) multiplies the entire user-perceived price p_h, not just the toll τ_h received by the monopolist. As a result, a substantial fee may be applied by the monopolist even to time periods when the optimal congestion fee is zero; indeed, the monopolist would never charge less than required to bring the absolute value of the elasticity, ε_h, above unity. These observations are consistent with our usual understanding of monopoly ownership of a resource with zero production cost.

In the special case of perfectly elastic demand, as in the two-road example analyzed by Pigou (1920, p. 194) and Knight (1924) where an uncongested road runs parallel to the road under consideration, the demand-related monopolistic mark-up disappears and the monopolist undertakes socially optimal pricing and investment. Although important as a benchmark, this

observation is of limited use in practice where demand elasticities (in absolute value) well below one typically apply (see also Section 2.1.4).

Note also from (6.5) that the fractional mark-up on marginal social cost, as given by the 'Lerner index' $(p_h - mc_h)/p_h$, is simply the inverse of the absolute value of the demand elasticity $(1/\varepsilon_h)$. Again this is consistent with conventional microeconomics. A similar relationship holds in Ramsey pricing, in which social welfare is maximized subject to a minimum-profit constraint (Ramsey, 1927; Baumol and Bradford, 1970). We can derive the Ramsey result in this case by maximizing $B - C$ subject to a constraint that profit Π defined in (6.1) exceed some value $\Pi^{\#}$ which cannot exceed the monopoly profit level. To do so, we maximize the Lagrangian function:

$$\begin{aligned}\Lambda = &\sum_h q_h \cdot \int_0^{V_h} d_h(v)\mathrm{d}v - \sum_h q_h \cdot V_h \cdot c_h(V_h;V_K) - \rho \cdot K(V_K) \\ &+ \lambda \cdot \left(\sum_h q_h \cdot V_h \cdot \left(d_h(V_h) - c_h(V_h;V_K)\right) - \rho \cdot K(V_K) - \Pi^{\#} \right),\end{aligned} \tag{6.6}$$

where λ is the Lagrangian multiplier associated with the budget constraint. The first-order condition for capacity V_K is unchanged (except for weighting the conventional condition by $1+\lambda$): once again, V_K is chosen to minimize total social cost. The first-order condition for volume V_h may be solved to yield:

$$\tau_h = V_h \cdot \frac{\partial c_h}{\partial V_h} - \frac{\lambda}{1+\lambda} \cdot V_h \cdot \frac{\partial d_h}{\partial V_h} \quad \Leftrightarrow \quad p_h \cdot \left(1 - \frac{\lambda}{1+\lambda} \cdot \frac{1}{\varepsilon_h}\right) = mc_h \,. \tag{6.7}$$

The 'Lerner index' $(p_h - mc_h)/p_h$ is that for a monopolist, $1/\varepsilon_h$, multiplied by $\lambda/(1+\lambda)$. This yields the first-best result when $\lambda=0$ (indicating that the constraint is non-binding), and the monopoly result as $\lambda \to \infty$.

6.1.2 Dynamic Congestion

In models with endogenous scheduling, demands at different times within the peak period

are determined by individual travelers' tradeoffs between travel delay and schedule delay. In general it is not obvious whether this sets up varying time-specific elasticities to be exploited by a monopolistic road owner, causing the time-varying fee to be distorted from the optimum. But whatever the time-varying fee turns out to be, a time-invariant fee can be analyzed, as in ADL (1991), using the steady-state model, thereby yielding the results just derived.

For the basic bottleneck model, defined in Chapter 3 and also used in Section 4.1.2, we can be more explicit. To avoid trivial results involving infinite profits with purely inelastic demand, consider a downward sloping inverse demand function $d(Q)$. Next, it is convenient to follow De Palma and Lindsey (2002) and distinguish between two toll components that together add up to the total time-varying toll $\tau(t')$ for an exit at t'. These are the time-independent 'base toll' τ_0, paid by all users, and a time-varying component $\tau_v(t')$. Hence, $\tau(t') = \tau_0 + \tau_v(t')$. If toll differentiation over time is costless, a revenue maximizer would not leave any queuing in existence, because queuing time can be replaced by toll revenues without affecting the generalized price p. The toll schedule is therefore at least as steep as the optimal one. But in absence of queuing, the revenue maximizer would also not make the toll schedule steeper than the optimal one. This would create periods within the peak where the bottleneck remains idle. But then it is possible to increase toll revenues by having the earliest (or the last) driver taking up an empty slot within the peak, and increasing his toll by the reduction in schedule delay cost he enjoys from this rescheduling.

The time-varying toll component will consequently follow the same pattern as the first-best time-varying toll of (@4.23) and $\tau_v(t')$=0 for the first and last driver. With linear scheduling cost, the revenues R_v from this time-varying toll component are equal to the total congestion-related user cost under optimal time-varying pricing, $\overline{C}_g^1 = \frac{1}{2} \cdot \delta \cdot Q^2 / V_K$; compare (@4.26).

The generalized price under optimal time-varying tolling with an additional base toll τ_0

amounts to $p = \tau_0 + \delta \cdot Q/V_K$; compare equation (@4.25). The revenues R_0 from the base toll τ_0 will therefore be $Q \cdot (d(Q) - \delta \cdot Q/V_K)$. Total profit can thus be written as:

$$\Pi = Q \cdot \left(d(Q) - \delta \cdot \frac{Q}{V_K} \right) + \tfrac{1}{2} \cdot \delta \cdot \frac{Q^2}{V_K} - \rho \cdot K(V_K) , \tag{6.8}$$

where the first term gives R_0, the second R_v, and the third term gives the capital cost. Maximizing (6.8) with respect to Q gives the following special case of the base toll derived by De Palma and Lindsey (2000) for the case of two bottlenecks in parallel (see Section 6.1.4 below):

$$d + Q \cdot \frac{\partial d}{\partial Q} - \delta \cdot \frac{Q}{V_K} = 0 \quad \Leftrightarrow \quad \tau_0 = -Q \cdot \frac{\partial d}{\partial Q} . \tag{6.9}$$

As argued in Section 4.1.2, the time-varying toll component $\tau_v(t')$ when set according to (@4.23) is equal to the time-varying marginal external congestion cost *mecc* for a user exiting at t'. With the base-toll τ_0 set according to (6.9), the total toll $\tau(t')$ is therefore equal to *mecc* (t') plus a conventional demand-related mark-up – and this is a straightforward time-dependent generalization of the static revenue-maximizing toll of (6.5).

Next, the first-best investment rule remains profit-maximizing also in the basic bottleneck model. To see this, observe that we can rewrite (6.8) as $\Pi = Q \cdot d(Q) - \overline{C}_g^1(Q, V_K) - \rho \cdot K(V_K)$. The first-order condition with respect to V_K minimizes the sum of total congestion-related user cost and capital cost, as it does in the first-best investment rule. The parallel with the result for the static model is again complete: deviations from first-best capacity only occur because the optimal investment rule is evaluated for a below-optimal level of road use.

The main insights from the static model on profit-maximizing tolling and capacity choice therefore survive in the basic bottleneck model.

6.1.3 Heterogeneous Users

When individuals' values of time differ, another deviation between conditions for optimality and those for profit maximization is created, as shown by Edelson (1971). This is because the non-discriminating monopolist considers the value of time only for marginal travelers (those nearly indifferent to using the highway in question); whereas conditions for optimality include inframarginal travelers as well. As a result, the monopolist may in fact allow either too much or too little congestion. The situation is analyzed elegantly by David Mills (1981), who models heterogeneity such that the value of time may vary, in either direction, with the individuals' reservation prices; or, in other words, when moving along the horizontal axis for an inverse demand function. He shows that monopoly private ownership tends to be superior to an unpriced equilibrium if users' reservation prices are either relatively similar or highly sensitive to congestion. Too much congestion occurs when, perhaps counter-intuitively, the profit-maximizing toll is lower than the first-best toll. This can only happen when users with relatively high reservation prices have relatively high values of time – which, as such, seems more plausible than the reversed relation. While these users' high value of time would affect the first-best toll that internalizes the congestion externality, it does not affect the marginal revenue of the operator via the "$V{\cdot}c'$-channel" discussed below equation (6.5). These users' values of time are therefore ignored by the profit-maximizer. Note that the underlying cause is the operator's (realistically) assumed inability to differentiate tolls by value of time.

Verhoef and Small (2004) extend the framework by modeling demand such that there is a marginal traveler for each value of time. Although under-pricing by a revenue-maximizer is not impossible in their model, it is not found in any of their numerical results, for which elasticities of around –0.4 apply for all values of time.

6.1.4 Second-best Pricing and Capacity Choice of Private Operators: The Two-Route Problem

Revisited

One way to limit a private road operator's market power is to supply an untolled public substitute. With the emerging interest in private pay-lanes, the welfare economic properties of the resulting second-best problem have received quite some attention in the recent literature. An unrestricted profit-maximizer on route T, facing an unpriced substitute U, would determine the toll and capacity by solving the following Lagrangian (note the similarity with problem (@5.9)):

$$\begin{aligned}\Lambda = {} & V_T \cdot \tau_T - \rho \cdot K_T(V_{K,T}) \\ & + \lambda_T \cdot \left(c_T(V_T; V_{K,T}) + \tau_T - d(V_T + V_U)\right) + \lambda_U \cdot \left(c_U(V_U; V_{K,U}) - d(V_T + V_U)\right).\end{aligned} \tag{6.10}$$

The set of first-order conditions (w.r.t. V_T, V_U, $V_{K,T}$, τ_T, λ_U and λ_T) can be solved to yield:

$$\tau_T = V_T \cdot \frac{\partial c_T}{\partial V_T} + V_T \cdot -\frac{\partial d}{\partial V} \cdot \left(\frac{\dfrac{\partial c_U}{\partial V_U}}{\dfrac{\partial c_U}{\partial V_U} - \dfrac{\partial d}{\partial V}} \right), \tag{6.11}$$

$$\rho \cdot K_T'(V_{K,T}) = -V_T \cdot \frac{\partial c_T}{\partial V_{K,T}}. \tag{6.12}$$

The investment rule (6.12) again has the familiar first-best structure indicating minimization of total social cost, which we also found, *inter alia*, for a private road without a substitute in (6.3), and for a public pay-lane with an unpriced substitute in (@5.11). The toll formula in (6.11) matches the one derived in Verhoef, Nijkamp and Rietveld (1996) for given capacities. The first term shows that the profit-maximizer still fully internalizes the congestion externality on the road under control. The second term gives the demand-related mark-up. It is a fraction, defined by the term in large brackets, of the mark-up in the profit-maximizing toll (6.5) for a single road. As for the second-best pay-lane toll of (@4.35), the fraction can be best interpreted by considering its extreme values. When the competing route U is uncongested, the profit-maximizer's demand is effectively perfectly elastic. The fraction falls to zero, and the profit-

maximizing toll becomes equal to the marginal external congestion cost, just as we found for a single road with perfectly elastic demand. On the other hand, when route U operates near capacity and V_U becomes insensitive to marginal changes on route T, the fraction becomes equal to unity. This reflects that the profit-maximizer can ignore the competition of route U, and set prices as if no competition from the parallel route exists.

Whereas the second-best one-route toll of equation (@4.35) subtracts a positive term from the marginal external cost on route T, to optimize the congestion spill-over upon route U, the profit-maximizing one-route toll adds a positive term in order to extract as much revenue as possible. No surprise then, that the welfare gains from private revenue-maximizing tolling is typically below the already small gains from second-best tolling, and may well be negative when compared to the unpriced situation for the same capacity.[1] This is illustrated by the lowest curve in Figure @4.5 in Chapter 4. The relative efficiency of revenue-maximizing pricing is negative for all possible capacities of the pay-lane. It reaches its minimum when 75% of the road's capacity is tolled. This illustrates that it may sometimes be better to have a private profit-maximizer operating an entire road, rather than some of its lanes. The profit-maximizer creates two distortions: he exploits market power, and aggravates congestion on the unpriced capacity. Increasing the share of capacity under the operator's control may reduce the distortion from the congestion spill-over more strongly than that it increases the market power distortion, in which case it is beneficial to increase the operator's capacity (Verhoef, Nijkamp and Rietveld, 1996).

It does now not come as a surprise that Liu and McDonald (1998), in their study of partial pricing on the Californian SR-91, find a relative efficiency "gain" of –0.85 (*i.e.*, a loss) for

[1] When comparing privately priced *additions* to existing capacity, relative efficiency cannot be negative as long as no deficits occur for the operator, no Braess paradox or similar problems arise elsewhere on the network, and no other externalities exist. The users must benefit from reduced travel times if the new capacity is to be used voluntarily. The operator cannot lose if no deficits are incurred. Therefore, no actor can lose, and social welfare cannot fall.

revenue-maximizing pricing on 1/3 of the road's capacity, while a relative gain of +0.09 would apply for second-best pricing. Small and Yan (2001) and Verhoef and Small (2004) find that these welfare losses become smaller, but certainly need not turn into gains, when allowing for heterogeneity in values of time. The reason is the same as for public pay-lane pricing, discussed in Section 4.2.1, and involves socially beneficial self-selection of users according to value of time.

Another reason, besides user heterogeneity, why a private pay-lane may be more efficient than anticipated in the studies mentioned above would be the favorable impact of revenue-maximizing pricing on departure times. Section 6.1.2 explained why revenue-maximizing in the basic bottleneck model entails time-differentiation of fees such that all queuing is eliminated. This remains true when an unpriced alternative exists (De Palma and Lindsey, 2002), and the elimination of queuing may drastically raise the relative efficiency of revenue-maximizing one-route tolling; *e.g.*, from –0.66 to +0.29 in the base-case of De Palma and Lindsey (2000).

Finally, Viton (1995), although not explicitly investigating welfare effects, presents a model that may (or may not) produce less pessimistic predictions on the efficiency impacts of private roads with unpriced substitutes. The roads, in his model, are imperfect substitutes, with a logit model describing auto users' route choice, and a probit model that of truckers. Imperfect substitutability might increase the distortions from revenue-maximizing pricing because the operator's market power is undermined less strongly than with a perfect substitute. But it might decrease the distortions from revenue-maximizing pricing because the congestion spill-overs upon the competing route are likely to be smaller. Whether either of these effects is more likely (or even certain) to dominate, and under which conditions, has not yet been analyzed.

6.1.5 Competition in Networks

Private ownership has also been analyzed in various other network configurations than the classic two-route problem with a single, unpriced substitute. De Palma and Lindsey (2000) for example consider different ownership regimes on a parallel-bottlenecks network, where either bottleneck can be free of toll, subject to a private revenue-maximizing toll, or to a public welfare-maximizing toll. They for example find that two competing private bottlenecks can yield most of the potential efficiency gains from first-best pricing (over 90% in their base case with time-varying tolling), and is nearly as efficient as a mixed duopoly with one public and one private operator. This matches conventional results on the relative efficiency of oligopoly under Bertrand competition.

On a similar network but with static congestion and distinguishing between local and transit traffic, De Borger, Proost and Van Dender (2005) study tax competition between two governments (*i.e.*, not private firms), each controlling one link. Their results suggest that appropriate tolling of transit traffic is at least as important for overall welfare as avoiding tax competition between the two governments, while the type of transit tolling (differentiated from tolls for local traffic or not) is less important. Also in the context of tax competition, but then (implicitly) on serial networks, Levinson (2001) presents evidence that jurisdictions are more likely to use tolls rather than fuel taxes when the share of non-resident workers increases.

With multiple private operators on a network, an important question is whether they operate predominantly substitute (parallel) roads, or complementary (serial) roads. The analysis of Economides and Salop (1992) for network markets in general suggests that competition on two parallel roads would reduce the tolls compared to monopolistic ownership of both links (as also found by De Palma and Lindsey, 2000), whereas competition on two serial roads would raise the total toll, for both roads together, compared to monopolistic ownership of both links.

It is instructive to illustrate and generalize these results for the case of more than 2 roads.

Suppose that we have a road of given overall capacity and length, serving a single origin-destination pair, and allow a number (F) of identical private firms (with possibly F=1) to charge revenue-maximizing tolls. To focus on the difference between parallel and serial competition, we will derive the Nash-Bertrand equilibrium toll levels for two contrasting cases; both symmetric for reasons of tractability. One is where the firms are organized such that they own equally sized parallel capacities over the full length of the road; *i.e.*, each firm operates $L^{\#}/F$ lanes where $L^{\#}$ denotes the total number of identical lanes (we ignore integer problems). The other is where they own equally sized serial capacities; *i.e.*, they each operate all lanes on a road segment of length L/F where the road's full length is L. We solve the question imposing a symmetric outcome in toll levels beforehand, and assume that all capacities are given. Primes will be used as in Chapter 4 to denote partial derivatives of cost and inverse demand functions w.r.t. V.

With individual firms occupying parallel segments of the same road, the equilibrium toll for any of these operators, indexed as firm f, can be derived by solving the following Lagrangian:

$$\begin{aligned}\Lambda_f = {} & V_f \cdot \tau_f + \lambda_f \cdot \left(c_f(V_f) + \tau_f - d(V_f + \sum_{g \neq f} V_g) \right) \\ & + \sum_{g \neq f} \lambda_f^g \cdot \left(c_g(V_g) + \tau_g - d(V_f + \sum_{h \neq f} V_h) \right)\end{aligned} . \tag{6.13}$$

The set of first-order conditions (w.r.t. V_f, V_g for all $g \neq f$, τ_f, λ_f and λ_f^g for all $g \neq f$) can be solved to yield the following toll:

$$\tau = \tau_f = V_f \cdot \left(c_f' - D' \right) + \frac{(F-1) \cdot D'}{c_{-f}' + (F-1) \cdot D'} \cdot V_f \cdot D' , \tag{6.14}$$

where τ denotes the toll that each user will pay for a trip; τ_f the toll for a specific firm (these are identical across firms and equal to τ because of symmetry); c_f the user cost for firm f and c_{-f} that for any other firm (where symmetry of course implies that c_f will be equal to c_{-f}). Equation (6.14) shows how the toll is equal to the monopolistic toll of (6.5) when F=1 and a monopoly situation indeed applies, while it approaches the first-best toll of (@4.6) when $F \to \infty$ and perfectly

competitive conditions apply.[2] These results are intuitive, and suggest that the equilibrium toll level, with symmetric firms, approaches the first-best level closer as the number of firms increases and the degree of market power therefore decreases.

When individual firms occupy serial segments, the following Lagrangian applies for a given firm *f*:

$$\Lambda_f = V \cdot \tau_f + \lambda \cdot \left(c_f(V) + \tau_f + \sum_{g \neq f} \left(c_g(V) + \tau_g \right) - d(V) \right). \tag{6.15}$$

The set of first-order conditions (w.r.t. V, τ_f, and λ_f) yields the following firm-specific toll:

$$\tau_f = V \cdot \left(c'_f + \sum_{g \neq f} c'_g - D' \right), \tag{6.16}$$

This toll is similar to the conventional revenue-maximizing toll of (6.5) and its interpretation is the same, but note that the firm not only internalizes the congestion on its own road segment, but also that on other firms' segments. The reason is that congestion on the other firms' segments affects the firm's marginal revenues in exactly the same way as the congestion on its own segment. It will be clear that if every firm internalizes the congestion on the entire road, there will be over-internalization of congestion when $F>1$. There is another upward pressure in the toll because every firm considers charges the same demand-related mark-up $V \cdot D'$. The total toll τ, for the full trip, amounts to:

$$\tau = \sum_f \tau_f = F \cdot V \cdot \left(c' - D' \right), \tag{6.17}$$

where c' gives the derivative of average user cost over the entire trip. With $F=1$, we obtain, as expected, the monopoly toll (6.5), which already exceeds the first-best toll because of the demand-related mark-up. The tax rule (6.17), and therewith the relative degree of overpricing $\tau/mecc$, otherwise increases in F. We therefore now find the opposite result to the parallel competition case: the lower the number of firms, the closer the overall toll approaches the

[2] The effects of varying c'_{-f} and D' on τ will not be discussed here for reasons of space, but can be verified in a comparable way.

efficient level – although even with one firm, over-pricing through the demand-related mark-up will remain unless demand is perfectly elastic.

There are therefore no general conclusions on the effect of the number of private road operators upon the efficiency of the resulting outcome. Much will depend on the network configuration in relation to the distribution of firms over that network. Our example confirms economic intuition and suggests that an increase in competition between substitutes would bring equilibrium tolls closer to first-best levels, while the opposite applies for an increase in the number of operators serving complementary roads. This would imply that private operators, if allowed on a network, should serve full-length corridors, but should face competition when doing so.

6.2 Regulation and Franchising of Private Roads

The local monopoly power that a private road operator would often obtain provides a potentially strong economic rationale for regulation. Moreover, given the physical nature and size of typical infrastructure investments, including its network aspects, its lumpiness, its irreversibility, the necessity to acquire right-of-way, and the long-lasting spatial implications, it would seem impractical to allow unrestricted free entry of private road operators. This raises the question of under what institutional set-up, encompassing the choice of a regulatory regime, private roads would contribute most to social welfare. The question gains relevance with the growing interest in and importance of private involvement in road operations throughout the world – with applications in Asia, Africa, Eastern and Western Europe, and North and South America (Estache, 2001; World Bank 2006).

Private involvement is typically motivated by the desire to bring in private capital when public budgets are tight, and by the hope that private management would be more efficient than public operation. A quite different type of motivation might be that the public would sooner accept tolling from a private company than from a public government, so that private roads may

be seen as a means to gradually implement road pricing. The most common format in recent years has been Build-Operate-Transfer (BOT) schemes. Under such a scheme, the concessionaire finances, builds, operates and maintains the road, and collects tolls for a certain period, usually around 30 years, after which the road is transferred to the government. With a Build-Own-Operate (BOO) scheme, the firm would operate the road for an unlimited period. Also leasing is possible, where the government remains in possession of the road and receives a predetermined fee from a private operator, who in turn has the right to charge the users. A Rehabilitate-Operate-Transfer (ROT) scheme concerns the rehabilitation of an existing road instead of the construction of a new one, but is otherwise close to a BOT-scheme.

As an alternative to actual tolls, 'shadow toll' systems have also been used in countries like the UK, Finland and The Netherlands. In this case, users do not pay actual tolls, but the authority remunerates the concessionaire depending on the degree of utilization. Whether the absence of actual tolls harms or improves social welfare of course depends on the question of whether a private toll would otherwise have improved or deteriorated social welfare, as discussed in Section 6.1 above. These contracts are often used in the context of Design-Build-Finance-Operate (DBFO) – which, however, can also have conventional tolls.

The experiences with highway franchising have not always been positive. Engel, Fisher and Galetovic (1997) mention as the two most prominent pitfalls the frequent use of government guarantees, which reduces the incentives to control construction costs, and renegotiations and government bailouts for almost every franchise that faces financial trouble. They attribute such problems mainly to the fact that franchises are typically awarded for a fixed period, and propose to use a variable-term contract instead, where the franchise is awarded to the bidder that requires the least-present-value-of-revenue (LPVR) from tolling. Once the accumulated toll revenues are such that the present value specified by the bidder is realized, the franchise ends. Such an

approach is likely to limit the scope for renegotiations of contracts, and therewith the likely distorting impact that the prospect of renegotiations may have upon the conditions that the private concessionaire initially accepts ('lowballing'), and thus upon the effectiveness and disciplining impacts of auctions for concessions.

The LPVR scheme, like all other schemes discussed above, would require some form of regulation of tolls and/or capacity, at least if the aim is to have tolls and capacities be set close to socially optimal levels. In fact, when the government would award the franchise to the firm that requires the smallest government subsidy or offers the highest payment for the right to operate the toll road, without otherwise specifying the required toll and capacity, the rules of the auction would actually push the bidders towards the profit-maximizing combination of capacity and toll of equations (6.5) and (6.3). Alternative criteria for auctions that have been used in practice include the total capacity cost, the construction period, the toll rate at opening, and the length of the concession (World Bank, 2006). Verhoef (2006) finds that toll rates and capacities may in fact vary strongly by the criterion used in the auction. His analysis considers static congestion, homogeneous travelers, neutral scale economies, and competitive auctions; and allows for unpriced congestion elsewhere on the road network. Perhaps surprisingly, a traffic flow maximizing auction appears to be capable of reproducing the zero-profit second-best outcome – but not always: it may lead to a minimum social surplus when the network configuration is such that a Braess-type of paradox plagues the efficiency impacts of the road under consideration.

The regulation of private road operators in relation with the design of franchises and auction procedures are of paramount importance for the eventual social desirability of private highways, for a variety of reasons. These include the potentially strong sensitivity of welfare with respect to toll levels and capacities, especially when untolled parallel capacity is present; the oftentimes substantial discrepancy between welfare and profit maximizing tolls; the complex

and sometimes hard-to-predict network effects that may result from private roads, both for single roads and for multiple roads (operated by either a single or multiple private operators); the potential risks of renegotiations on the eventual budgetary implications of private road provision as well as on the disciplining impact of auctions for concessions; and the uncertainty over future demand. A skeptic might well conclude that, given the associated complexities in defining optimal institutional arrangements, public road provision and tolling should remain the preferred option, to be abandoned only when case-specific compelling arguments apply. A less extreme viewpoint would be that private road operation should be no goal in itself, and would require a solid motivation either through the government's inability of raising funds at an interest rate as low as the private sector can (which seems unlikely in many cases), or through the government's inability to operate as efficiently and innovatively as a private road operator could – which might, however, also lead to less drastic measures than privatization of roads. But libertarians might take a rather different perspective and argue that it is, for a start, not the private provision of road capacity that would require a solid motivation, but its public provision instead.

Whether private or public road provision would be more desirable in the long run can, in conclusion, therefore not be said in general. The main factors increasing the support for private provision include the following: sufficient competition from substitutes and no competition from complements; particularly inefficient planning and operation from public road suppliers, including inability to implement (congestion) pricing; the existence of disadvantages for the public sector in capital markets, or other fundamental reasons for a lack of funds for public investment; relatively small external effects other than congestion (emissions; noise); and the availability of an efficient and effective auction mechanism. The extent to which these conditions apply, and therewith the relative desirability for private road supply, can be expected to vary strongly across nations, regions, and over time.

6.3 Privately Provided Transit Services Untouched

A variety of organizational forms for providing transit service are of interest, ranging from full responsibility by public agencies, through various coordinating or regulatory roles, to fully deregulated private provision. Comparing these options has become an urgent priority for public policy because of the problems with public subsidies noted earlier.

Cost Comparisons between Public and Private Ownership. Perry et al. (1988) review numerous studies attempting to determine whether public or private transit operators are more efficient. The weight of evidence seems somewhat on the side that private operators are more efficient, although there are several counter-examples. The conclusion of Perry et al. is that cost efficiency is more closely related to management incentive systems than to the form of ownership. In particular, there is evidence that subsidies, whether to public or private operators, raise the cost of providing service, as well as encourage lower fares and higher levels of service (Anderson, 1983). Other comparative studies not reviewed by Perry et al. have found substantial cost advantages for private operators in the U.S. (Morlok and Viton, 1985a) and Australia (Hensher, 1988a). Walters (1987b) finds a profitability advantage for private operators over public ones charging identical fares in the same city, for a variety of cities in the developed and developing world.

This type of research is hampered by several kinds of spurious effects that can affect the comparisons. As noted by Perry et al., during a time of transition from private to public ownership, the most inefficient private operators may be the ones first to be socialized, biasing the results from cross-sectional samples against finding efficient public firms. (The opposite

holds in an environment of gradual privatization.) A publicly owned firm often has tax advantages and, if it is not an independent authority, it may cover overhead expenses from other budgets, in both cases producing the false appearance of lower costs. Walters' profitability comparisons run the risk that the private operators may have found ways to serve just the most profitable markets, or contrariwise (as Walters argues) that the public authorities may prohibit private operators on the more profitable routes. More generally, low profitability on the part of public operators can arise either from high cost or from offering a less-utilized type of service; the latter might or might not be in the public interest. Taken together, these caveats suggest that the comparative evidence for private operators being more efficient is inconclusive.

Potential for Profitable Private Provision of Conventional Mass Transit. Given the incentive problems with subsidies, one may ask whether it would be better to forego the theoretical welfare advantages of low fares and simply let private firms operate unrestricted. Several authors have found that there is a considerable range of conditions under which profitable operation is possible. Viton (1980c), using the stylized corridor model described earlier with endogenous service frequency and route density, finds profitability to be possible for a range of corridor densities. Harker (1988) takes a quite different approach, formulating a model in which several types of transit (rail, luxury bus, standard bus, school bus, and minibus), each with a simple linear cost function, compete with each other and with (uncongested) auto on part or all of a network; the Nash equilibrium leads to profitable bus service in one of three Philadelphia-area corridors considered, which has relative high density and low incomes. Harker's result contrasts somewhat with case studies analyzed by Morlok and Viton (1980, 1985b), which find a niche for expensive, high-quality service by commuter rail (Chicago), rapid transit (Lindenwold line into Philadelphia), and bus (express service into Manhattan). Cervero

(1990), reviewing individual transit routes in twenty-five U.S. cities, finds that those actually operating at a profit serve mostly high-density areas with low-income people taking short trips.

It seems that there are two potentially profitable markets for conventional transit. One is high-quality express commuting service from affluent suburbs to large employment centers, either by rail (if density is very high) or by express bus. The other is local bus service serving low-income people in high-density areas.

Competitive Practices and Welfare Effects. If the market for mass transit services is opened to private firms, in competition with each other and perhaps with a subsidized public agency, they may adopt strategies that can be modeled as noncooperative games. Several authors have investigated this possibility.

One line of inquiry is the nature of imperfectly competitive (unregulated) equilibria in which each firm assumes that the offerings of other firms (usually in terms of service frequency, timetable, and fare) are fixed. Foster and Golay (1986) extend the standard Hotelling (1929) model of product differentiation, showing that under a wide variety of conditions, stable equilibria can occur in which each firm finds a different product niche. They argue from this that such "curious old practices" as buses racing each other for the next load of passengers are exceptional rather than normal. However, they do argue that collusion, predatory pricing, and "gaming" behavior arising from congestion must be regulated.

Evans (1987) compares four situations: monopoly, unregulated non-cooperative oligopoly with free entry (a form of imperfect competition), welfare maximization subject to a breakeven constraint, and unconstrained welfare maximization. The oligopolistic equilibrium exhibits features reminiscent of monopolistic competition: higher fares and higher service frequency than either constrained or unconstrained welfare maximization would produce. The inefficiency

occurs because each firm expands its offerings regardless of adverse effects on other firms' service quality. However, in Evans's simulations (p. 23), welfare in the oligopolistic case falls only slightly short of that resulting from constrained or unconstrained welfare maximization, whereas it far exceeds (at most demand levels) that resulting from monopoly. Hence Evans results are supportive of deregulation as a viable policy when welfare maximization is not achievable.

Dodgson and Katsoulacos (1988b) also consider entry in an imperfectly competitive world, finding a wide range of conditions for which just two firms share the market and differentiate their product. This result is consistent with a finding by Viton (1981a) that two firms, modeled after the major bus and rail operators in the eastern San Francisco Bay Area, would significantly differentiate their products if they engaged in Cournot-like competition. Glaister (1986) simulates free entry on five bus routes in Aberdeen, Scotland, finding that most entry would occur with smaller vehicles offering more frequent service than presently.

Would such entry occur even if a subsidized public operator remained? Obviously it depends on the policies of that operator. Viton (1982) finds that under plausible U.S. conditions, the presence of the subsidized firm deters entry in many markets. Teal and Nemer (1986) describe a case in which a newly legalized jitney service in Los Angeles was driven out of business because of a system-wide fare reduction by the large public operator, which maintained directly competing service even though it was unprofitable.

The models described above assume constant returns to scale in producing intermediate outputs. Furthermore, most implicitly assume that any economies of scale due to user-supplied time is at a system rather than a firm level: that is, the traveler cares only about total bus frequency on the route, not about the frequency provided by a given firm. This, however, raises troubling questions about the viability of a non-integrated system of urban transit. What if it is

not feasible for each firm to use the same stops, for example because they use vehicles of different sizes or because major terminals are owned by one firm? What if consumers care about reputations of firms because otherwise they cannot judge the comfort or reliability of the vehicle they are about to enter? What if the unregulated equilibrium entails differentiated products, e.g., high-fare express and low-fare local service, so that travelers with a strong preference for one cannot benefit from the extra service frequency offered by the other? These questions, largely neglected by advocates of deregulation, are explored by Gwilliam et al. (1985) and Nash (1988). These authors emphasize the practical difficulties of creating, from an equilibrium of competing firms, the kind of integrated system that takes advantage of the returns to scale inherent in scheduled service. They also remind us of several sources of economies of scale and scope, such as through ticketing of passengers and scheduling of drivers.

Given the possiblity that monopoly service is best after all, we can then ask whether potential rather than actual competition will regulate it adequately. At the extreme, we can ask whether the transit market is contestable: Is the prospect of hit-and-run entry sufficient to restrain a monopoly provider in choosing its fare and service policies? Contestability requires that the entrant have low barriers to entry and exit (the latter requiring an absence of sunk costs), yet that the incumbent be unable to change fares and service levels too quickly (Baumol et al., 1982; Bailey, 1981).

Button (1988) argues that there is substantial though not perfect contestability in urban transit. Certain features favor low barriers to entry and exit: lack of significant economies of scale in providing vehicle-hours of service, low setup costs, and a good market in used bus equipment. On the other hand, substantial investments may be required to establish a reputation, build terminal facilities, or achieve efficiency through learning-by-doing. These investments cannot be retrieved if the monopolist responds to entry by lowering fare or increasing service. If

the monopolist can credibly threaten to do so temporarily, in order to drive out the entrant, it is said to be capable of predation, which discourages entry. Dodgson and Katsoulacos (1988a) analyze when a rational monopolist would respond in this way, showing that informational asymmetries can lead to successful predation. Contestability in transit markets could perhaps be tested empirically using the method developed for airlines by Morrison and Winston (1987).

There is also some limited empirical evidence suggesting that transit markets are not fully contestable. Evans (1988) describes the experience in Hereford, England, the site of an early experimental deregulation of transit service beginning in 1981. Following a brief period of intense competition, it appears that the dominant firm did in fact drive out all its rivals except in one small segment of its market. Fares ultimately returned nearly to the levels that prevailed prior to the experiment, but service levels remained substantially higher. Evans suggests that potential entry constrains the monopolist's service levels, which cannot be quickly increased in response to entry, but not its fares.

In comparing welfare under different regimes, it is important to remember that although artificially high wages cause distortions and create financial problems for public operators, the wages themselves represent a transfer and not a welfare loss. To the extent that workers are simply paid above their market wage, reducing their wages does not add directly to social welfare as normally defined. This point is made by Gwilliam et al. (1985, p. 110); it is taken into account in the formal welfare analysis of White (1990), but is often ignored in the more polemic literature on privatization. Its significance is muted by the possibility that high wages lead to inefficient rent-seeking activities, as in the general analysis of Tullock (1967): that is, to activities using real resources and serving no purpose other than deciding who gets the above-market wages, For example, a firm might allocate jobs by requiring higher skills than are really needed to perform the work.

Contracting Out and Deregulation. Based on the kinds of analysis discussed so far, most transportation researchers support increased private-sector participation in providing urban transit services. There is debate, however, over which of two directions is more promising: the letting of contracts by a governmental agency which maintains organizational responsibility for the service, or the outright privatization of publicly owned firms coupled with repeal of regulatory restrictions on entry. The British Transport Act of 1985 established the former policy for London and the latter for the rest of Britain, thereby creating a useful experiment for study. The intensive debate within the professional transportation community prior to the act is nicely captured by the papers by Beesley and Glaister (1985) and Foster (1985) supporting full deregulation, and by Gwilliam et al. (1985) opposing it. Results are described in the next subsection.

If contracts are let through competitive bids, the process is known as *competitive tendering*. However, competitive tendering can also mean awarding an exclusive franchise for operating a given set of routes, without centralized control of the service offered (Hensher, 1988b). Furthermore, competitive awarding of franchises can be done with or without subsidy.

Rooney and Teal (1986) and Tally and Anderson (1986) provide estimates of cost functions for contracted bus service, further adding to the evidence of substantial cost reductions, in part through lower wages.

Evidence from Cases of Deregulation and Privatization. Much has been learned from the first few years of experience following the British Transport Act of 1985. The main results outside London were higher fares, lower patronage, and substantial cost savings, less than half of which is attributable to wage reductions. Gómez-Ibáñez and Meyer (1990) provide a readable

account. They believe the experience is promising, with the main negative effect (higher fares) caused by the simultaneous cut in government subsidies rather than by the change in organizational form. Gwilliam (1987) and White (1990) offer views that, while less optimistic, are by no means negative. White shows that the service reductions immediately following the act were temporary, and that serivce expanded substantially over the first three years. Gwilliam's and White's main reservations seem to be the loss of advantages of integration. For example, White concurs with Gómez-Ibáñez and Meyer that, in several metropolitan areas at least, patronage fell by more than would be predicted from fare rises and service reductions alone, even after accounting for long-term downward trends in transit ridership; both papers attribute this to the instability and lack of information surrounding deregulation. Gómez-Ibáñez and Meyer attribute the problems of instability and information to transitional difficulties and mistakes on the part of local transport authorities, whereas White seems to think they are more inherent to an unregulated environment.

The nature of competition varied substantially among metropolitan areas. This observation leads Gómez-Ibáñez and Meyer to conclude that public authorities play an important role in setting the ground rules on competition. Contestability was not really tested because the legislation requires 42-day notice for new entry or exit, creating a substantial sunk cost. Predation does seem to have occurred, and probably more would have in the absense of legal strictures (Dodgson and Katsoulacos, 1991). Duopolies sharing a single market with differentiated products, as envisioned by Glaister (1986) and by Dodgson and Katsoulacos (1988b), have not been observed; perhaps taxi service is serving the high-quality market, thereby precluding a third competitor such as minibus (Dodgson and Katsoulacos, 1988b, p. 280).

White carefully tabulates the welfare effects of the 1985 Act over a three-year period. He concludes (his Table 9) that the six large English metropolitan areas showed a small net welfare

gain, with substantial losses to consumers and workers more than offset by very large operating-cost reductions. The shires, consisting of smaller cities and rural areas, suffered welfare declines, with little operating-cost savings to offset the considerable loss of consumer surplus. Scotland showed a slight welfare decline.

The cost reductions are more thoroughly analyzed by Heseltine and Silcock (1990), who find 15-20 percent reductions in unit costs by the newly privatized former divisions of the National Bus Company, and 30 percent reductions by the municipally owned operators outside Scotland. Productivity improved dramatically. In Scotland, where the major operator was not privatized, these gains were largely absent.

There seems more unanimous agreement on the experiment with competitive tendering in London: namely, that it has been beneficial. Cost per bus-kilometer was reduced by around 14 percent. This reduction is smaller than that achieved in other areas (White, 1990, Table 4); but because service levels were maintained, virtually all of it is a net welfare gain, shared between taxpayers (through lower subsidies) and users (through lower fares).

New Zealand adopted legislation in 1989 similar to Britain's (Fielding and Johnston, 1990). The United States has experimented with contracting of services and with carefully circumscribed relaxation of regulation. These and other innovations are discussed by Cervero (1988), Talley (1988), and various papers in Lave (1985). Experience in Southeast Asia is discussed by Rimmer (1988), and prospects for continental western Europe by Gwilliam and Van de Velde (1990).

6.4 Paratransit Untouched

One of the arguments for deregulating and privatizing urban transit is to promote the spontaneous development of innovative "paratransit" services filling particular niches (Cervero, 1988). Possibilities include rental cars, shared-ride taxi, other ridesharing arrangements such as commuter vanpools or casual carpools formed at designated pickup points, demand-responsive services activated by telephone or hailing, semi-scheduled jitney services by small vans, and subscription commuter buses. These services not only are discouraged by competition from highly subsidized conventional transit in areas where they might thrive, but are prohibited or inhibited by numerous regulations.

Both experience and analysis suggest that many such services are economical. Jitneys and owner-operated taxis, for example, have low overhead costs, may employ part-time drivers, and can adjust easily to meet changing demand. Carpooling provides flexible service with far less use of highway infrastructure and parking facilities than solo driver. (Carpooling also offers one of the few possibilities for quickly adapting to fuel shortages or natural disasters.) Dispatched vans or taxis offer convenient ways to meet the needs of special groups such as the elderly or handicapped, at a far lower cost than making conventional transit universally accessible.

Taxicab service is possibly the most seriously neglected sector of urban transportation in both research and policy analyses. Taxis handle large volumes of passenger trips 40 percent that of conventional transit in the United States in 1970, and 12 percent even in transit-oriented London in 1969 (Kirby et al., 1975, p. 61; Beesley, 1979, p. 103). Perhaps because they provide an inexpensive substitute for owning a car, taxis are heavily used by poor people (Kirby et al., 1975, chap. 7) as well as by people in higher income classes for specialized trips. When allowed

to offer shared-ride service, they do so more cheaply than publicly operated demand-responsive bus or van service (Cervero, 1988).

As noted earlier, taxicab service exhibits scale economies analogous to those on scheduled transit service, especially in the cruising sector where cabs are hailed by sight. This is because the average waiting time for finding a cab declines with the density of cruising cabs, which in turn rises with demand. So long as there is little product differentiation, these scale economies are clearly at the industry rather than the individual firm level, suggesting that the unregulated equilibrium will entail too little service. Countering this effect, perhaps, are congestion externalities imposed by taxis both while driving and while picking up traffic; but if buses or private cars are the main substitute, those externalities are to some degree present anyway. These basic tenets of analyzing the taxicab industry have been developed formally by Douglas (1972), Manski and Wright (1976), Beesley and Glaister (1983), and Frankena and Pautler (1986). It should be entirely feasible to incorporate such models into broader models of urban equilibria in the same way as conventional transit service.

Taxicab service is heavily regulated, with both fare and entry strictly controlled in many cities. There is some evidence that such regulation serves the interests mainly of existing operators who earn supernormal profits (Taylor, 1989). Several advantages of deregulating taxicab service have been suggested, and limited experience in the U.S. suggests that they can be realized in practice (Kirby et al., 1975, chap. 7; Frankena and Pautler, 1986; Cervero, 1988). First, eliminating restrictions on entry and fares has led to significant increases in the number of cabs (hence in level of service), and to decreases in fares. (This may also provide a needed source of casual employment for low-skilled inner-city residents.) Second, eliminating restrictions on shared-ride services has resulted in significant use of this mode. It should be noted that permitting free entry while maintaining high controlled fares would lead to

inefficiently underutilized service, because the high fares would attract operators to the point where extra cruising time depressed net earnings to those of low-paid alternative occupations.

The only significant problems with deregulation of taxis seem to be occasional incidents of overcharging or otherwise cheating riders,[3] and oversupply at taxi stands of limited capacity, especially at airports. These probably can best be handled through physical design of taxi stands, regulations on posting fares, or surcharges at congested stands. There have been virtually no serious problems with deregulation of radio-dispatched service, which accounts for about three-fourth of taxicab rides in the United States (Frankena and Pautler, 1986, p. 157).

It has been claimed that jitney service can benefit competing transit systems by helping them shed expensive peak demands. This view, however, may be mistaken because it does not take into account the adverse effect on service quality for other transit patrons. Conventional transit and jitney are both increasing-returns industries, so there is a legitimate role for public policy to consider whether allowing them to compete is efficient. An ideal situation might be where jitneys or taxis can use the same stops as conventional buses, thereby allowing those patrons who are willing to use either service to choose whichever comes first; such a situation exists, for example, between regular city buses and special "sherut" taxis on one of the major shopping street in Tel Aviv.

6.5 Summary Untouched – this pertains to current Chapters 4 – 5

It is clear that pricing and investment policies for urban transportation are far from

---

[3]See Glazer and Hassin (1983) for a theoretical analysis of cheating by taxi drivers.

optimal. It is now possible to say which deficiencies have the greatest adverse effects. The most important is the underpricing of congestion. By failing to charge for scarce peak-period highway capacity, transportation planners have allowed highway performance to degenerate. Meanwhile policy-makers have become paralyzed by the combination of inadequate financial resources and rapidly rising costs. Attempts to compensate for this failure have distorted policy in other areas, resulting in biases toward inflexible and capital-intensive forms of public transit, dramatic transit deficits, costly expansion of transit service to markets it cannot serve well, and bureaucratically complex incentive programs to discourage driving. The failure is exacerbated by regulations and tax policies that encourage free parking, by subsidies to automobile use by state and local taxpayers, and by insurance mechanisms that convert variable costs to fixed expenses. The result is an artificially low price of travel, which drains resources and distorts urban development patterns.

From an analytical viewpoint, understanding congestion is hampered by failing to incorporate queueing dynamics and trip scheduling into the standard analysis. This chapter has presented some models that alleviate this shortcoming. They suggest that time-varying congestion fees can provide benefits considerably beyond those arising from simple pricing schemes like the one in Singapore. These models also provide a better way to analyze situations where latent demand, especially in the form of potential shifts in travel schedules, threatens to undo most of the currently popular policies aimed at reducing congestion.

Optimal highway investment can be analyzed using a cost-benefit framework, either through marginal conditions leading to first- or second-best optimality or through a comparison of costs and benefits from discrete changes. Although the analytical difficulties are formidable, the technique is well understood. It is limited mainly by the accuracy of the predictive models used to forecast changes, and by the politicization of the environment in which the analyses are

done.

Dissatisfaction with the outcome of transportation services provided by the public sector has sparked renewed interest in applying classical free-market principles to transportation. Current research suggests that although the transportation sector is far from meeting the conditions under which unregulated markets are fully efficient, selected use of private enterprise can improve incentives and bring about significant cost savings. The study of regulation and privatization should mature as empirical evidence is gathered from the many experiments in progress around the world.

REFERENCES

NOTE: notation to be made consistent

Abbas, Khaled A., and Mona H. Abd-Allah (1999) "Estimation and Assessment of Cost Allocation Models for Main Transit Systems Operating in Cairo." *Transport Reviews,* **9**: 353-375.

Adler, Moshe (1985) "Street Parking: The Case for Communal Property." *Logistics and Transportation Review,* **21**:375-387.

Agnew, Carson E. (1977) "The theory of congestion tolls" *Journal of Regional Science* **17** 381-393.

Akçelik, Rahmi (1991) "Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time function" *Australian Road Research* **21** (3) 49-59.

Alberini, Anna, Maureen Cropper, Alan Krupnick, and Nathalie B. Simon (2004) "Does the Value of a Statistical Life Vary with Age and Health Status? Evidence from the US and Canada," *Journal of Environmental Economics and Management*, **48**: 769-792.

Alfa, Attahiru Sule (1986) "A Review of Models for the Temporal Distribution of Peak Traffic Demand." *Transportation Research Part B,* **20**:491-499.

Allport, R.J. (1981) "The Costing of Bus, Light Rail Transit and Metro Public Transport Systems." *Traffic Engineering and Control,* **22**:633-639.

Amemiya, Takeshi (1978) "On a Two-Step Estimation of a Multivariate Logit Model." *Journal of Econometrics,* **8**:3-21.

Amemiya, Takeshi (1981) "Qualitative Response Models: A Survey." *Journal of Economic Literature,* **9**:1483-1536.

American Association of State Highway and Transportation Officials (2003) *A Manual of User Benefit Analysis for Highways*. AASHTO, Washington, August.

Anas, Alex (1983) "Discrete Choice Theory, Information Theory and the Multinomial Logit and Gravity Models." *Transportation Research Part B,* **7**:3-23.

Anderson, Shirley C (1983) "The Effect of Government Ownership and Subsidy on Performance: Evidence from the Bus Transit Industry." *Transportation Research Part A,* **7**:191-200.

Anderson, David, and Herbert Mohring (1997) "Congestion Costs and Congestion Pricing." In: Greene, Jones, and Delucchi (1997), pp. 315-336.

Anderson, Simon P., André de Palma, and Jacques F. Thisse (1988) "A Representative Consumer Theory of the Logit Model." *International Economic Review,* **29**:461-466.

Anderson, Simon P., and André de Palma (2004) "The Economics of Pricing Parking." *Journal of Urban Economics*, **55**:1-20.

Angel, S., and G.M. Hyman (1976) *Urban Fields: A Geometry of Movement for Regional Science*. London: Pion.

Appleyard, Donald (1981) *Livable Streets*. Berkeley, California: University of California Press.

Ardekani, Siamak, and Robert Herman (1987) "Urban Network-Wide Traffic Variables and Their Relations." *Transportation Science,* **21**:1-16.

Armour, Rodney F. (1980) "An Economic Analysis of Transit Bus Replacement." *Transit Journal,* **6**:41-54.

Arnott, Richard (2004) *Some Downtown Parking Arithmetic* Unpublished manuscript, Department of Economics, Boston College.

Arnott, Richard, André de Palma, and Robin Lindsey (1988) "Schedule Delay and Departure Time Decisions with Heterogeneous Commuters." *Transportation Research Record*, **1197**:56-67.

Arnott, Richard, André de Palma, and Robin Lindsey (1990a) "Departure Time and Route Choice for the Morning Commute." *Transportation Research Part B,* **24**:209-228.

Arnott, Richard, André de Palma, and Robin Lindsey (1990b) "Economics of a Bottleneck." *Journal of Urban Economics,* **27**:111-130.

Arnott, Richard, André de Palma, and Robin Lindsey (1991a) "Does providing information to drivers reduce traffic congestion?" *Transportation Research* **25A** (3) 309-318.

Arnott, Richard, André de Palma, and Robin Lindsey (1991b) "A temporal and spatial equilibrium analysis of commuter parking" *Journal of Public Economics* **45** (3) 301-335.

Arnott, Richard, André de Palma, and Robin Lindsey (1992) "Route choice with heterogeneous drivers and group-specific congestion costs" *Regional Science and Urban Economics* **22** (1) 71-102.

Arnott, Richard, André de Palma, and Robin Lindsey (1993) "A structural model of peak-period congestion: a traffic bottleneck with elastic demand" *American Economic Review* **83** (1) 161-179.

Arnott, Richard, André de Palma, and Robin Lindsey (1998) "Recent developments in the bottleneck model". In: Button and Verhoef (1998), pp. 79-110.

Arnott, Richard, André de Palma, and Robin Lindsey (1999) "Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand" *European Economic Review* **43** (3) 525-548.

Arnott, Richard, and Eren Inci (2005) "An Integrated Model of Downtown Parking and Traffic Congestion." Working paper, Boston University, Chestnut Hill, Mass.

Arnott, Richard and Marvin Kraus (1998a) "When are anonymous congestion charges consistent with marginal cost pricing?" *Journal of Public Economics* **67** (1) 45-64.

Arnott, Richard and Marvin Kraus (1998b) "Self-financing of Congestible Facilities in a Growing Economy". In: David Pines, Efraim Sadka and Itzhak Zilcha (eds.) (1998) *Topics in Public Economics: Theoretical and Applied Analysis* Cambridge: Cambridge University Press, pp. 161-184.

Arnott, Richard, and John Rowse (1999) "Modeling Parking." *Journal of Urban Economics*, **45**:97-124.

Arnott, Richard and Kenneth A. Small (1994) "The economics of traffic congestion" *American Scientist* **82** 446-455.

Arnott, Richard and An Yan (2000) "The two-mode problem: second-best pricing and capacity" *Review of Urban and Regional Development Studies* **12** (3) 170-199.

Arsenio, Elisabete, Abigail L. Bristow, and Mark Wardman (2006) "Stated Choice Valuations of Traffic Related Noise," *Transportation Research Part D*, **11**:15-31.

Austroads (1996) *Benefit Cost Analysis Manual*. Publication No. AP-42/96, Austroads, Sidney.

Bailey, Elizabeth E. (1981) "Contestability and the Design of Regulatory and Antitrust Policy." *American Economic Review, Papers and Proceedings,* **71**:178-183.

Bailey, Elizabeth E., and Ann F. Friedlaender (1982) "Market Structure and Multiproduct Industries." *Journal of Economic Literature,* **20**:1024-1048.

Banister, David (1992) "Energy Use, Transport and Settlement Patterns". in Breheny, M. J. (eds.), *Sustainable Development and Urban Form*. 160-181, London: Pion.

Banister, David, Joseph Berechman, and Gines de Rus (1992) "Competitive Regimes within the European Bus Industry: Theory and Practice." *Transportation Research Part A,* **26**:167-178.

Banks, James H. (1989) "Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations." *Transportation Research Record,* **225**:53-60.

Barnes, Gary, and Peter Langworthy (2003) "The Per-Mile Costs of Operating Automobiles and Trucks," Report MN/RC 2003-19, Minnesota Department of Transportation, St. Paul, June. http://www.lrrb.gen.mn.us/PDF/200319.pdf, accessed 13 April 2005.

Basso, L.J. and S.R. Jara-Díaz (2006), "Is returns to scale with variable network size adequate for transport industry structure analysis?" *Transportation Science*, forthcoming.

Bates, John (1997) "Forecasting Travel Demand and Response". in De Rus, G., and C. Nash (eds.), *Recent Developments in Transport Econonomics*. 8-32

Bates, John, Denvil Coombe, Martin Dale, Mike Maher, Sally Cairns, Phil Goodwin, Carmen Hass-Klau, Ryuichi Kitamura, Toshiyuki Yamamoto, and Satoshi Fujii (1998) *Traffic Impact of Highway Capacity Reductions*. London: Landor Publishing Ltd.

Bates, John, John Polak, Peter Jones, and Andrew Cook (2001) "The Valuation of Reliability for Personal Travel." *Transportation Research Part E: Logistics and Transportation Review,* **37**:191-229.

Baumol, William J., and David F. Bradford (1970) "Optimal Departures from Marginal Cost Pricing." *American Economic Review,* **60**:265-283.

Baumol, William J., T John G. Panzer, and Robert D. Willig (1982) *Contestable Markets and the Theory of Industry Structure.*, New York: Harcourt Brace Jovanovich.

Baum-Snow, Nathaniel, and Matthew E. Kahn (2000.) "The Effects of New Public Projects to Expand Urban Rail Transit." *Journal of Public Economics,* **77**:241-263.

Becker, Gary S (1965) "A Theory of the Allocation of Time." *Economic Journal,* **75**:493-517.

Beckmann, Martin, C. Bartlett McGuire, and Christopher B. Winsten (1956) *Studies in the Economics of Transportation* Yale University Press, New Haven.

Beesley, M.E. (1979) "Competition and Supply in London Taxis." *Journal of Transport Economics and Policy,* **3**:102-131.

Beesley, M.E., and S. Glaister (1983) "Information for Regulating: The Case of Taxis." *The Economic Journal,* **93**:594-615.

Beesley, M.E., and S. Glaister (1985) "Deregulating the Bus Industry in Britain - (C) A Response." *Transport Reviews,* **5**:133-142.

Beesley, Michael E., and David A. Hensher (1990) "Private Tollroads in Urban Areas: Some Thoughts on the Economic and Financial Issues." *Transportation,* **6**:329-341.

Beggs, S., S. Cardell, and J. Hausman (1981) "Assessing the Potential Demand for Electric Cars." *Journal of Econometrics,* **6**:1-19.

Bell, Michelle L., Aidan McDermott, Scott L. Zeger, Johathan M. Samet, and Francesca Dominici (2004) "Ozone and Short-Term Mortality in 95 US Urban Communities, 1987-2000." *Journal of the American Medical Association*, **292**:2372-2378.

Ben-Akiva, Moshe (1974) "Structure of Passenger Travel Demand Models." *Transportation Research Record,* **526**:26-42.

Ben-Akiva, Moshe (1985) "Dynamic Network Equilibrium Research." *Transportation Research Part A,* **9**:429-431.

Ben-Akiva, Moshe, and Michel Bierlaire (2003) "Discrete Choice Methods and their Applications to Short Term Travel Decisions." In: Hall (2003), ch. 2.

Ben-Akiva, Moshe, and John L. Bowman (1998) "Activity Based Travel Demand Model Systems," in: Patrice Marcotte and Sang Nguyen (eds.), *Equilibrium and Advanced Transportation Modelling*, Kluwer, Boston: 27-46.

Ben-Akiva, Moshe, Michele Cyna, and André de Palma (1984) "Dynamic Model of Peak Period Congestion." *Transportation Research Part B*, **18**:339-355.

Ben-Akiva, Moshe, André de Palma, and Pavlos Kanaroglou (1986) "Dynamic Model of Peak Period Traffic Congestion with Elastic Arrival Rates." *Transportation Science,* **20**:164-181.

Ben-Akiva, Moshe, André de Palma and Isam Kaysi (1991) "Dynamic Network Models and Driver Information Systems" *Transportation Research* **25A** (5) 251-266.

Ben-Akiva, Moshe, and Steven R. Lerman (1979) "Disaggregate Travel and Mobility-Choice Models and Measures of Accessibility". in Hensher, D. A., and P. R. Stopher. (eds.), *Behavioural Travel Modelling*. 654-679, London: Croom Helm.

Ben-Akiva, Moshe, and Steven R. Lerman (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, Mass.: MIT Press.

Ben-Akiva, Moshe, and Takayuki Morikawa (1990) "Estimation of Travel Demand Models from Multiple Data Sources," in: M. Koshi (ed.), *Transportation and Traffic Theory: Proceedings of the Eleventh International Symposium on Transportation and Traffic Theory*, Elsevier, Amsterdam: 461-476.

Bento, Antonio M., Maureen L. Cropper, Ahmed Mushfiq Mobarak, and Katja Vinha (2004) "The Impact of Urban Spatial Structure on Travel Demand in the United States," Working paper, Dept. of Economics, Univ. of Maryland.

Berechman, Joseph (1983) "Costs, Economies of Scale and Factor Demand in Road Transport." *Journal of Transport Economics and Policy,* **8**:7-24.

Berechman, Joseph (1993) *Public Transit Economics and Deregulation Policy*. Amsterdam: North-Holland.

Berechman, Joseph, and Genevieve Giuliano (1985) "Economies of Scale in Bus Transit: A Review of Concepts and Evidence." *Transportation,* **2**:313-332.

Berechman, Joseph and David Pines (1991) "Financing road capacity and returns to scale under marginal cost pricing" *Journal of Transport Economics and Policy* **25** 177–181.

Berechman, Joseph, and Kenneth A. Small (1988) "Modeling Land Use and Transportation: An Interpretive Review for Growth Areas." *Environment and Planning Part A,* **20**:1285-1309. Research Policy and Review No. 25.

Berger, J.O., T and L.R. Pericchi (2001) "Objective Bayesian Methods for Model Selection: Introduction and Comparison," working paper, Duke University, Durham, North Carolina

Berry, Steven, James Levinsohn, and Ariel Pakes (1995) "Automobile Prices in Market Equilibrium," *Econometrica*, 63: 841-890.

Bertini, Robert L., and Anthony M. Rufolo (2004) "Technology Considerations for the Implementation of a Statewide Road User Fee System." In: Evangelos Bekiaris and Yuko J. Nakanishi (eds.), *Economic Impacts of Intelligent Transportation Systems: Innovations and Case Studies*, in series, *Research in Transportation Economics*, **8**:337-361.

Bhat, Chandra (1995) "A Heteroscedastic Extreme Value Model of Intercity Travel Mode Choice." *Transportation Research Part B,* **29**: 471-483.

Bhat, Chandra R., and Jessica Guo (2004) "A Mixed Spatially Correlated Logit Model: Formulation and Application to Residential Choice Modeling," *Transportation Research B*, **38**: 147-168.

Bhat, Chandra R., and Frank S. Koppelman (2003) "Activity-Based Modeling of Travel Demand," in: Hall (2003), ch. 3.

Black, Alan (1990) "Analysis of Census Data on Walking to Work and Working at Home." *Transportation Quarterly,* **44**:107-120.

Black, I.G., and J.G. Towriss (1993) "Demand Effects of Travel Time Reliability," Centre for Logistics and Transportation, Cranfield Institute of Technology

Blincoe, Lawrence J., Angela G. Seay, Eduard Zaloshnja, Ted R. Miller, Eduardo O. Romando, Stephen Luchter, and Rebecca Spicer (2002) *The Economic Impact of Motor Vehicle Crashes 2000*, Washington, D.C.: US National Highway Traffic Safety Administration (NHTSA) http://www.nhtsa.dot.gov.

Bly, P.H., F.V. Webster, and S. Pounds (1980) "Effects of Subsidies on Urban Public Transport." *Transportation,* **9**:311-331.

Boardman, Anthony E., and Lester B. Lave (1977) "Highway Congestion and Congestion Tolls." *Journal of Urban Economics,* **4**:340-359.

Boardman, Anthony E., David H. Greenberg, Aidan R. Vining, and David L. Weimer (2006) *Cost-Benefit Analysis: Concepts and Practice*, 3rd edition, Prentice Hall.

Boarnet, Marlon G., and Randall Crane (2000) *Travel by Design: The Influence of Urban Form on Travel*, Oxford Univ. Press, New York.

Boarnet, Marlon G., and Sharon Sarmiento (1998) "Can Land-use Policy Really Affect Travel Behaviour? A Study of the Link between Non-work Travel and Land-use Characteristics." *Urban Studies,* **35**:1155-1169.

Boiteux, M. (1949) "La Tarification des Demandes en Pointe: Application de la Theorie de la Vente au Cout Marginal." *Revue Generale de l'Electricite*. Reprinted in English translation as "Peak-Load Pricing," *Journal of Business*, 33, (1960), 157-179.

Bowman, John L., and Moshe E. Ben-Akiva (2001) "Activity-Based Disaggregate Travel Demand Model System with Daily Activity Schedules," *Transportation Research A*, 35: 1-28.

Boyce, David E. (1984) "Urban Transportation Network-Equilibrium and Design Models: Recent Achievements and Future Prospects." *Environment and Planning Part A,* **6**:1445-1474.

Boyce David E. (2002) "Is the Sequential Travel Forecasting Procedure Counterproductive?" *Journal of Urban Planning and Development*, **128**: 169-183.

Boyce, David E., and Hillel Bar-Gera (2004) "Multiclass combined models for urban traffic forecasting." *Networks and Spatial Economics*, **4**:115-124.

Boyce, David E., Hani S. Mahmassani, and Anna Nagurney (2005) "A Retrospective on Beckmann, McGuire and Winsten's *Studies in the Economics of Transportation*." *Papers in Regional Science*, **84**:85-103.

Boyd, J. Hayden (1976) "Benefits and Costs of Urban Transportation: He Who Is Inelastic Receiveth and Other Parables." *Transportation Research Forum Proceedings,* **7**:290-297.

Boyd, J. Hayden, Norman J. Asher, and Elliot S. Wetzler (1973) *Evaluation of Rail Rapid Transit and Express Bus Service in the Urban Commuter Market.* Institute for Defense Analyses. Prepared for US Dept. of Transportation, Report DOT-P-6520.1. US Government Printing Office, Washington.

Boyd, J. Hayden, Norman J. Asher, and Elliot S. Wetzler (1978) "Nontechnological Innovation in Urban Transit: A Comparison of Some Alternative." *Journal of Urban Economics,* **5**:1-20.

Boyd, J. Hayden, and Robert E. Mellman (1980) "The Effect of Fuel Economy Standards on the U.S. Automotive Market: An Hedonic Demand Analysis," *Transportation Research A*, 14: 367-78.

Boyer, Marcel, and Georges Dionne (1987) "The Economics of Road Safety." *Transportation Research Part B,* **21**:413-431.

Braess, Dietrich (1968) "Über ein Paradoxon aus der Verkehrsplanung" *Unternehmenforschung*, **12**, 258-268.

Braeutigam, Ronald R. (1999) "Learning about Transport Costs." In: Gómez-Ibáñez, Tye, and Winston (1999), pp. 57-97.

Braid, Ralph M. (1989) "Uniform versus Peak-Load Pricing of a Bottleneck with Elastic Demand." *Journal of Urban Economics,* **26**:320-327.

Braid, Ralph M. (1996) "Peak-Load Pricing of a Transportation Route with an Unpriced Substitute." *Journal of Urban Economics* **40** (2) 179-197.

Brajer, Victor, Jane V. Hall, and Robert Rowe (1991) "An Integrated Approach to Benefits Assessment: Attaining Ozone and Particulate Standards." *Contemporary Policy Issues*, **9**:81-91.

Branston, David (1976) "Link Capacity Functions: A Review." *Transportation Research*, **10**:223-236.

Brice, Stéphane (1989) "Derivation of Nested Transport Models Within a Mathematical Programming Framework." *Transportation Research Part B,* **23**:19-28.

Brownstone, David, and Xuehao Chu (1997) "Multiply-Imputed Sampling Weights for Consistent Inference with Panel Attrition." In: Golob, T. F., R. Kitamura, and L. Long (eds.), *Panels for Transportation Planning: Methods and Applications*, Kluwer, Amsterdam: 259-273.

Brownstone, David, Arindam Ghosh, Thomas F. Golob, Camilla Kazimi, and Dirk Van Amelsfort (2003) "Drivers' Willingness-to-Pay to Reduce Travel Time: Evidence from the San Diego I-15 Congestion Pricing Project." *Transportation Research Part A*, **37**:373-387.

Brownstone, David, and Kenneth A. Small (1989) "Efficient Estimation of Nested Logit Models." *Journal of Business and Economic Statistics,* **7**:67-74.

Brownstone, David, and Kenneth A. Small (2005) "Valuing Time and Reliability: Assessing the Evidence from Road Pricing Demonstrations," *Transportation Research Part A*, **39**:279-293.

Brownstone, David, and Kenneth Train (1999) "Forecasting new product penetration with flexible substitution patterns." *Journal of Econometrics,* **89**:109-129.

Bruzelius, Nils (1979) *The Value of Travel Time.*, London: Croom Helm.

Bunch, David S (1991) "Estimability in the Multinomial Probit Model." *Transportation Research Part B,* **25**:1-12.

Bureau of Transportation Statistics (BTS) (2001) *National Transportation Statistics 2000*. US Department of Transportation, Washington: US Government Printing Office.

Button, Kenneth J. (1988) "Contestability in the UK Bus Industry, Experience Goods and Economies of Experience". in Dodgson, J. S., and N. Topham. (eds.), *Bus Deregulation and Privatisation*. 69-96, Aldershot, U.K.: Gower.

Button, Kenneth J. (1993) *Transportation Economics*, Edward Elgar, Cheltenham, UK.

Button, Kenneth J., and K.J. O'Donnell (1985) "An Examination of the Cost Structures Associated with Providing Urban Bus Services in Britain." *Scottish Journal of Political Economy,* **32**:67-81.

Button, Kenneth J., and Erik T. Verhoef, eds. (1998) *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*, Edward Elgar, Cheltenham, UK.

Bye, Raymond Taylor (1926) "The Nature and Fundamental Elements of Costs." *Quarterly Journal of Economics,* **41**:30-62.

Calfee, John, and Clifford Winston (1998) "The Value of Automobile Travel Time: Implications for Congestion Policy." *Journal of Public Economics,* **69**: 83-102.

Calthrop, Edward and Stef Proost (2004) "Regulating On-Street Parking." Working Paper 2004-10, Catholic University Leuven, Belgium.

Calthrop, Edward, Stef Proost, and Kurt van Dender (2000) "Parking Policies and Road Pricing." *Urban Studies*, **37**:63-76.

Cambridge Systematics, Inc., in association with Barton-Aschman Associates (1977) *The Development of a Disaggregate Behavioral Work Mode Choice Model*. Cambridge, Mass.: Cambridge Systematics.Prepared for California Department of Transportation and the Southern California Associaton of Governments.

Cambridge Systematics, Inc., with Robert Cervero and David Aschauer (1998) *Economic Impact Analysis of Transit Investments: Guidebook for Practitioners*. Transit Cooperative Research Program Report 35 (Washington, D.C.: National Academy Press)

Cardell, N. Scott, and Frederick C. Dunbar (1980) "Measuring the Societal Impacts of Automobile Downsizing," *Transportation Research Part A*, 14: 423-34.

Cassidy, Michael J. and Robert L. Bertini (1999) "Some Traffic Features at Freeway Bottlenecks." *Transportation Research Part B*, **33** (1): 25-42.

Caudill, Steven B (1988) " An Advantage of the Linear Probability Model over Probit or Logit." *Oxford Bulletin of Economics and Statistics,* **50**:425-427.

Cervero, Robert (1982) "Multistage Approach for Estimating Transit Costs." *Transportation Research Record,* **877**:67-75.

Cervero, Robert (1986) "Time-of-Day Transit Pricing: Comparative US and International Experiences." *Transport Reviews,* **6**:347-364.

Cervero, Robert (1988) "Revitalizing Urban Transit". In: J.C. Weicher, (ed.), *Private Innovations in Public Transit,*. 71-81, Washington, D.C.: American Enterprise Institute for public Policy Research.

Cervero, Robert (1990) "Profiling Profitable Bus Routes." *Transportation Quarterly,* **44**:183-201.

Cervero, Robert (1996) "Jobs-Housing Balancing Revisited: Trends and Impacts in the San Francisco Bay Area," *Journal of the American Planning Association*, **62**:492-511.

Cervero, Robert, and Roger Gorhan (1995) "Commuting in Transit Versus Automobile Neighborhoods." *Journal of the American Planning Association,* **61**:210-225.

Cervero, Robert, and Mark Hansen (2002) "Induced Travel Demand and Induced Road Investment: A Simultaneous Equation Analysis." *Journal of Transport Economics and Policy*, **36**:469-490.

Cervero, Robert, and Kang-Li Wu (1996) "Subcentering and Commuting: Evidence from the San Francisco Bay Area, 1980-1990," working paper, Dept. of City & Regional Planning, Univ. of California, Berkeley, November.

Chamberlain, Gary (1984) "Panel Data". in Griliches, Z., and M. D. Intriligator. (eds.), *Handbook of Econometrics, Volume II,*. 1247-1318, Amsterdam: North-Holland.

Chan, Luke K.P (1975) *Nonpecuniary Return to Work: Theory and Empirical Evidence Based on the Value of Commutation Time*. Ph.D. Dissertation, University of California, Berkeley

Chan, Y., and F.L. Ou (1978) "Tabulating Demand Elasticities for Urban Travel Forecasting." *Transportation Research Record,* **673**:40-46.

Chang, Gang-Len, Hani S. Mahmassani, and Robert Herman (1985) "Macroparticle Traffic Simulation Model to Investigate Peak-Period Commuter Decision Dynamics." *Transportation Research Record,* **1005**:107-121.

Chattopadhyay, Sudip (2001) "Welfare Measurement in the Discrete-Choice Random Utility Model under General Preference Structure,*"* unpublished manuscript, Dept. of Economics, San Francisco State University

Chen, Huey-Kuo (1999) *Dynamic Travel Choice Models: A Variational Inequality Approach*, Springer, Berlin.

Chen, Mei, and David H. Bernstein (2004) "Solving the Toll Design Problem with Multiple User Groups." *Transportation Research Part B*, **38**:61-79.

Choi, Ki-Hong, and Choon-Geol Moon (1997) "Generalized Extreme Value Model and Additively Separable Generator Function." *Journal of Econometrics,* **76**:129-140.

Chomitz, Kenneth M., Charles A. Lave, and Urban Transit Costs (1984) "Part-Time Labour, Work Rules." *Journal of Transport Economics and Policy,* **8**:63-73.

Choo, Sangho, Patricia L. Mokhtarian, and Ilan Salomon (2005) "Does Telecommuting Reduce Vehicle-Miles Traveled? An Aggregate Time Series Analysis for the U.S." *Transportation*, **32**:37-64.

Chow, Gregory C (1983) *Econometrics*. New York: McGraw-Hill.

Chu, Chaushie (1981) *Structural Issues and Sources of Bias in Residential Location and Travel Mode Choice Models*. Ph.D. Dissertation,, Northwestern University. Ann Arbor, Michigan: University Microfilms

Chu, Xuehao (1995) "Endogenous Trip Scheduling: The Henderson Approach Reformulated and Compared with the Vickrey Approach." *Journal of Urban Economics,* **37**:324-343.

Chu, Xuehao (1999) "Alternative congestion pricing schedules." *Regional Science and Urban Economics* **29** (6): 697-722.

Coleman, Robert R. (1961) "A Study of Urban Travel Times in Pennsylvania Cities." *Highway Research Board Bulletin,* **303**:62-75.

Commissariat Général du Plan (2001) *Transports: Choix des Investissements et coût des nuisances* (*Transportation: Choice of Investments and the Cost of Nuisances*) Paris, June.

Coombe, R.D. (1989) "Review of Computer Software for Traffic Engineers." *Transport Reviews,* **9**:217-234.

Crane, Randall (2000) "The Influence of Urban Form on Travel: An Interpretive Review," *Journal of Planning Literature*, 15: 3-23.

D'Este, Glen (2000) "Urban Freight Movement Modeling." In: Hensher and Button (2000), pp. 539-552.

Dafermos, Stella C. (1972) "The Traffic Assignment Problem for Multi-User Transportation Network." *Transportation Science*, **6**: 73-87.

Dafermos, Stella C. (1980) "Traffic Equilibrium and Variational Inequalities." *Transportation Science*, **14:** 42-54.

Daganzo, Carlos F. (1997) *Fundamentals of Transportation and Traffic Operations*. Elsevier Science, New York.

Daganzo, Carlos F., Michael J. Cassidy and Robert L. Bertini (1999) "Possible Explanations of Phase Transitions in Highway Traffic." *Transportation Research Part A*, **33:**365-379.

Daganzo, Carlos, and Michael Kusnic (1993) "Two Properties of the Nested Logit Model." *Transportation Science*, **27**: 395-400.

Daganzo, Carlos F. and Yosef Sheffi (1977) "On Stochastic Models of Traffic Assignment." *Transportation Science*, **11:**253-274.

Dagenais, Marcel G., and Marc J.I. Gaudry (1986) "Can Aggregate Direct Travel Demand Models Work?" in *Research for Tomorrow's Transport Requirements: Proceedings of the World Conference on Transport Research, Vol. 2*, Vancouver: Centre for Transportation Studies, University of British Columbia, pp. 1669-1676.

Dahlgren, Joy (1998) "High occupancy vehicle lanes: not always more effective than general purpose lanes." *Transportation Research Part B,* **32**:99-114.

Daly, Andrew J. (1987) "Estimating 'Tree' Logit Models." *Transportation Research Part B,* **21**:251-267.

Darbera, Richard (1993) "Deregulation of Urban Transport in Chile: What Have We Learned in the Decade 1979-1989?" *Transport Reviews,* **3**:45-59.

Davies, R.B., A.R. Pickles, and R. Crouchley (1983) "Some Methods for the Testing and Estimation of Dynamic Models Which Use Panel Data." *Environment and Planning Part A,* **5**:1475-1488.

Davis, Stacy C., and Susan W. Diegel (2004) *Transportation Energy Data Book: Edition 24*, Oak Ridge National Laboratory, Oak Ridge, Tennessee. http://cta.ornl.gov/data, accessed 6 April 2005.

Dawson, J.A.L., and Fred N. Brown (1985) "Electronic Road Pricing in Hong Kong: A Fair Way to Go?" *Traffic Engineering and Control,* **26**:522-529.

Day, Brett (1999) "A Meta-Analysis of Wage-Risk Estimates of the Value of a Statistical Life," in European Commission, *Benefits Transfer and the Economic Valuation of Environmental Damage in the European Union: with Special Reference to Health*, DG-XII, contract ENV4-CT96-0234. http://www.cserge.ucl.ac.uk/VOSL.pdf, accessed 27 April 2005.

De Borger, Bruno, and Kristiaan Kerstens (2000) "The Performance of Bus-Transit Operators," in Hensher and Button (2000), pp. 577-595.

De Borger, Bruno and Stef Proost (2001) *Reforming Transport Pricing in the European Union: A Modelling Approach* Edward Elgar, Cheltenham, UK.

De Borger, Bruno, Stef Proost and Kurt van Dender (2005) "Congestion and tax competition in a parallel network" *European Economic Review* **49** (8) 2013-2040.

De Borger, Bruno, and Kurt Van Dender (2003) "Transport Tax Reform, Commuting and Endogenous Values of Times." *Journal of Urban Economics*, **53**:510-530.

De Ceuster, Griet, Laurent Franckx, Bart Van Herbruggen, Steven Logghe, Bruno Van Zeebroeck, Stijn Tastenhoya, Stef Proost, Jasper Knockaert, Ian Williams, Gordon Deane, Angelo Martino, and Davide Fiorello (2005) *TREMOVE 2.30 Model and Baseline Description: Final Report*. Catholic University of Leuven, Belgium, for European Commission, DG ENV, Directorate C, 18 Feb. See http://europa.eu.int/comm/environment/air/tremove/tremove_model_dev.htm.

De Jong, Gerard (2000) "Value of Freight Travel-Time Savings." In: Hensher and Button (2000), pp. 553-564.

Delucchi, Mark (2000) “Should We Try To Get the Prices Right?” *Access*, **16**:10-14. University of California Transportation Center, Berkeley.

De Neufville, Richard, and Joseph H. Stafford (1971) *Systems Analysis for Engineers and Managers*. New York: McGraw-Hill.

De Palma, André, and Richard Arnott (1986) "Usage-Dependent Peak-Load Pricing." *Economics Letters,* **20**:101-105.

De Palma, André, and Philippe Jehiel (1995) “Queuing May Be First-Best Efficient.” Discussion Paper 95-20, Thema, Paris.

De Palma, André, Moez Kilani, and Robin Lindsey (2004) “A Comparison of Second-best and Third-best Tolling Schemes on a Road Network.” Unpublished manuscript, University of Alberta, Edmonton.

De Palma, André, C. Lefevre, and M. Ben-Akiva (1987) "A Dynamic Model of Peak Period Traffic Flows and Delays in a Corridor." *Computational Mathematics Applications,* **4**:201-223.

De Palma, André, and Robin Lindsey (2002) “Private Roads, Competition and Incentives to Adopt Time-Based Congestion Tolling.” *Journal of Urban Economics* **52**:217-241.

De Palma, André and Robin Lindsey (2005) “Relation between pricing, toll revenues and investment”. Task Report 2.1, Project REVENUE, DG-TREN Fifth Framework Programme.

De Palma, André, and Robin Lindsey (1998) “Information and Usage of Congestible Facilities Under Different Pricing Regimes.” *Canadian Journal of Economics*, **31** (3): 666-692.

De Palma, André, and Robin Lindsey (2000) “Private Roads: Competition Under Various Ownership Regimes.” *Annals of Regional Science*, **34** (1): 13-35.

De Palma, André, Robin Lindsey, and Stef Proost (eds.) (2006) *Modelling of Urban Road Pricing and its Implementation*. Special issue of *Transport Policy*, **13**(2).

De Palma, André, and Fabrice Marchal (2002) “Real Cases Applications of the Fully Dynamic METROPOLIS Tool-Box: An Advocacy for Large-Scale Mesoscopic Transportation Systems.” *Networks and Spatial Economics*, **2**:347-369.

De Rus, Gines, and Chris Nash (1997) *Recent Developments in Transport Econonomics*. Aldershot, UK: Ashgate.

Deaton, Angus (1985) "The Demand for Personal Travel in Developing Countries: An Empirical Analysis." *Transportation Research Record,* **1037**:59-66.

DeCorla-Souza, Patrick (2004) “Recent U.S. Experience: Pilot Projects.” In: Santos (2004), pp. 283-308.

DeSerpa, A.C (1971) "A Theory of the Economics of Time." *Economic Journal,* **81**:828-846.

DeVany, Arthur S., and Thomas R. Saving (1980) "Competition and Highway Pricing for Stochastic Traffic." *Journal of Business,* **53**:45-60.

Dewees, Donald N (1976) "Urban Express Bus and Railroad Performance: Some Toronto Simulations." *Journal of Transport Economics and Policy,* **10**:16-25.

Dewees, Donald N (1978) "Simulations of Traffic Congestion in Toronto." *Transportation Research,* **12**:153-165.

Dewees, Donald N (1979) "Estimating the Time Costs of Highway Congestion." *Econometrica,* **47**:1499-1512.

Dodgson, John S (1986) "Benefits of Changes in Urban Public Transport Subsidies in the Major Australian Cities." *The Economic Record,* **62**:224-235.

Dodgson, John S., and Neville Topham (1987) "Benefit-Cost Rules for Urban Transit Subsidies." *Journal of Transport Economics and Policy,* **21**:57-71.

Dodgson, John S., and Yannis Katsoulacos (1988a) "Models of Competition and the Effect of Bus Service Deregulation". in Dodgson, J. S., and N. Topham (eds.), *Bus Deregulation and Privatisation,*. 45-68, Aldershot, U.K.: Gower.

Dodgson, John S., and Yannis Katsoulacos (1988b) "Quality Competition in Bus Services." *Journal of Transport Economics and Policy,* **22**:263-281.

Dodgson, John S., and Yannis Katsoulacos (1991) "Competition, Contestability and Predation: The Economics of Competition in Deregulated Bus Markets." *Transportation Planning and Technology,* **5**:263-275.

Domencich, Thomas A., and Gerald Kraft (1970) *Free Transit*. Lexington, Mass.: D.C. Heath.

Douglas, George W. (1972) "Price Regulation and Optimal Service Standards." *Journal of Transport Economics and Policy,* **4**:116-127.

Douglas, George W., James C. Miller III, and Efficiency in the Price-Constrained Airline Market (1974) "Quality Competition, Industry Equilibrium." *American Economic Review,* **64**:657-669.

D'Ouville, Edmond L., and John F. McDonald (1990a) "Effects of Demand Uncertainty on Optimal Capacity and Congestion Tolls for Urban Highways." *Journal of Urban Economics,* **28**:63-70.

D'Ouville, Edmond L., and John F. McDonald (1990b) "Optimal Road Capacity with a Suboptimal Congestion Toll." *Journal of Urban Economics,* **28**:34-49.

Dowling, R.G., R. Singh, and W.W.K. Cheng (1998) "The Accuracy and Performance of Improved Speed-Flow Functions." *Transportation Research Record*, **1646**:9-17.

Downes, J.D., and P. Emmerson (1983) *Urban Transport Modelling with Fixed Travel Budgets (An Evaluation of the UMOT Process)* Supplementary Report 799, Crowthorne, England: Transport and Road Research Laboratory,

Downs, Anthony (1962) "The Law of Peak-Hour Expressway Congestion." *Traffic Quarterly,* **6**:393-409.

Downs, Anthony (2004) *Still Stuck in Traffic: Coping with Peak-Hour Traffic Congesiton*, Brookings Institution, Washington, D.C.

Dubin, Jeffrey A., and Daniel L. McFadden (1984) "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption." *Econometrica,* **52**:345-362.

Dupuit, Jules (1844) "De l'Influence des Peages sur l'Utilite des Voies de Communication." *Annales des Ponts et Chaussees*. Translated by Elizabeth Henderson as "On Tolls and Transport Charges," *International Economic Papers, 11 (1962): 7-31*.

Dupuit, Jules (1849) "De la Mesure de l'Utilite des Travaux Publics." *Annales des Ponts et Chaussees,* **8**. Translated by R.H. Barback as "On the Measurement of the Utility of Public Works," in *International Economic Papers, 2* (1952): 83-110.

Economides, Nicholas and Steven C. Salop (1992) "Competition and integration among complements, and Network market structure" *The Journal of Industrial Economics* **40** (1) 105-123.

Edelson, Noel M. (1971) "Congestion Tolls Under Monopoly." *American Economic Review,* **61**:873-882.

Edlin, Aaron S., and Pinar Karaca-Mandic (2003) "The Accident Externality from Driving." Working paper, University of California at Berkeley, July. http://repositories.cdlib.org/iber/econ/E03-332/, accessed 11 May 2005.

El Sanhouri, I., and David Bernstein (1994) "Integrating Driver Information and Congestion Pricing Systems." *Transportation Research Record*, **1450**:44-50.

Else, P.K (1981) "A Reformulation of the Theory of Optimal Congestion Taxes." *Journal of Transport Economics and Policy,* **5**:217-232.

Eliasson, Jonas, and Lars-Goran Mattsson (2006). "Equity Effects of Congestion Pricing: Quantitative Methodology and a Case Study for Stockholm," *Transportation Research Part A*, **40**:602-620.

Emmerink, Richard H.M. (1998) *Information and Pricing in Road Transportation*. Springer, Berlin.

Emmerink, Richard H.M., and Peter Nijkamp (eds.) (1999) *Behavioural and Network Impacts of Driver Information Systems*. Ashgate, Aldershot, UK.

Engel, Eduardo, Ronald Fisher and Alexander Galetovic (1997) "Highway franchising: pitfalls and opportunities" *American Economic Review, Papers and Proceedings* **87** (2) 68-72.

Estache, Antonio (2001) "Privatization and regulation of transport infrastructure in the 1990's" *The World Bank Research Observer* **16** (1) 85-107.

Ettema, Dick, and Harry Timmermans (1997) "Theories and Models of Activity Patterns," in: Dick Ettema and Harry Timmermans (eds.), *Activity-Based Approaches to Travel Analysis*, Pergamon, Amsterdam: 1-36.

European Commission DG Regional Policy (2002) *Guide to Cost-Benefit Analysis of Investment Projects*. European Commission, Brussels. http://europa.eu.int/comm/regional_policy/sources/docgener/guides/cost/guide02_en.pdf (accessed July 26, 2006).

EXTRA (2001) *Getting Prices Right: Results from the Transport Research Programme*. European Commission, DG Energy and Transport, Consortium for EXploitation of TRAnsport Research, Brussels. http://europa.eu.int/comm/transport/extra/web/downloadfunction.cfm?docname=200406/20040617_110400_88575_pricing.pdf&apptype=application/pdf, accessed July 14, 2006.

European Conference of Ministers of Transport (ECMT, 1998) *Efficient Transport for Europe: Policies for Internalisation of External Costs*. Paris: OECD Publications Service.

Evans, Andrew (1987) "A Theoretical Comparison of Competition with Other Economic Regimes for Bus Services." *Journal of Transport Economics and Policy,* **21**:7-36.

Evans, Andrew (1988) "Hereford: A Case Study of Bus Deregulation." *Journal of Transport Economics and Policy*, **22**:283-306.

Fargier, Paul-Henri (1983) "Effects of the Choice of Departure Time on Road Traffic Congestion". in Hurdle, V. F., E. Hauer, and G. N. Steuart (eds.), *Proceedings of the Eighth International Symposium on Transportation and Traffic Theory*. 223-263, Toronto: University of Toronto Press.

Fernald, John G. (1999) "Roads to Prosperity? Assessing the Link Between Public Capital and Productivity." *American Economic Review*, **89:**619-638.

Fielding, Gordon J., and Douglas C. Johnston (1992) "Restructuring Land Transport in New Zealand." *Transport Reviews,* **12**:271-289.

Florian, Michael, and Marc Gaudry (1980) "A Conceptual Framework for the Supply Side in Transportation Systems." *Transportation Research Part B,* **4**:1-8.

Florian, Michael, and Donald Hearn (2003) "Network Equilibrium and Pricing," in Hall (2003), ch. 11.

Flyvbjerg, Bent, Matte K. Skamris Holm, and Søren L. Buhl (2003) "How Common and How Large Are Cost Overruns in Transport Infrastructure Projects?" *Transport Reviews*, **23**:71-88.

Flyvbjerg, Bent, Matte K. Skamris Holm, and Søren L. Buhl (2004) "What Causes Cost Overrun in Transport Infrastructure Projects?" *Transport Reviews*, **24**:3-18.

Flyvbjerg, Bent, Mette K. Skamris Holm, and Søren L. Buhl (2006) "Inaccuracy in Traffic Forecasts," *Transport Reviews*, **26**, pp. 1–24.

Forsyth, P.J. (1980) "The Value of Time in an Economy with Taxation." *Journal of Transportation Economics and Policy,* **14**:337-362.

Foster, Christopher D. (1974) "The Regressiveness of Road Pricing." *International Journal of Transport Economics,* **1**:133-141.

Foster, Christopher D. (1985) "The Economics of Bus Deregulation in Britain." *Transport Reviews,* **5**:207-214.

Foster, Christopher, and Jeanne Golay (1986) "Some Curious Old Practices and Their Relevance to Equilibrium in Bus Competition." *Journal of Transport Economics and Policy,* **20**:191-216.

Fowkes, A.S., P.E. Firmin, G. Tweddle, and A.E. Whiteing (2004) "How Highly Does the Freight Transport Industry Value Journey Time Reliability – and for What Reasons?" *International Journal of Logistics: Research and Applications*, **7**:33-43.

Frank, M. and P. Wolfe (1956) "An algorithm for quadratic programming" *Naval Research Logistics Quarterly* **3** (1-2) 95-110.

Frankena, Mark W. (1981) "The Effects of Alternative Urban Transit Subsidy Formulas." *Journal of Public Economics,* **15**:337-348.

Frankena, Mark W., and Paul A. Pautler (1986) "Taxicab Regulation: An Economic Analysis." *Research in Law and Economics,* **9**:129-165.

Frankena, Mark W., and Bus Scrapping (1987) "Capital-Biased Subsidies, Bureaucratic Monitoring." *Journal of Urban Economics,* **21**:180-193.

Fridstrøm, Lasse, Jan Ifver, Siv Ingebrigtsen, Risto Kulmala, and Lars Krogsgård Thomsen (1995) "Measuring the Contribution of Randomness, Exposure, Weather, and Daylight to the Variation in Road Accident Counts," *Traffic Analysis and Prevention*, **27**: 1-20.

Friesz, Terry L. (1980) "Transportation Network Equilibrium, Design and Aggregation: Key Developments and Research Opportunities." *Transportation Research Part B,* **4**:413-427.

Fulton, Lewis M., Robert B. Noland, Daniel J. Meszler, and John V. Thomas (2000) "A Statistical Analysis of Induced Travel Effects in the U.S. Mid-Atlantic Region." *Journal of Transportation and Statistics,* **3**:1-14.

Garling, Tommy, Thomas Laitila, and eds Kerstin Westin (1998) *Theoretical Foundations of Travel Choice Modeling,*. Amsterdam: Elsevier.

Gaudry, Marc J.I (1975) "An Aggregate Time-Series Analysis of Urban Transit Demand: The Montreal Case." *Transportation Research,* **9**:249-258.

Gaudry, Marc J.I., and Michael J. Wills (1978) "Estimating the Functional Form of Travel Demand Models." *Transportation Research,* **12**:257-289.

Geltner, David, and Fred Moavenzadeh (1987) "An Economic Argument for Privatization of Highway Ownership." *Transportation Research Record,* **1107**:14-20.

Giuliano, Genevieve (1986) "Land Use Impacts of Transportation Investments: Highway and Transit". in Hanson., S. (eds.), *The Geography of Urban Transportation*. 247-279, New York: Guilford Press.

Giuliano, Genevieve (1989) "New Directions for Understanding Transportation and Land Use." *Environment and Planning Part A,* **21**:145-159. Research Policy and Review No. 27.

Giuliano, Genevieve (1991) "Is Jobs-Housing Balance a Transportation Issue?" *Transportation Research Record*, 1305: 305-312.

Giuliano, Genevieve (1994) "Equity and Fairness Considerations of Congestion Pricing". In: National Research Council (1994), *Volulme 2: Commissioned Papers*, pp. 250-279.

Giuliano, Genevieve (2004) "Land Use Impacts of Transportation Investments: Highway and Transit." In: Susan Hanson and Genevieve Giuliano (eds.), *The Geography of Urban Transportation*, Guilford Press, New York: 237-273.

Giuliano, Genevieve, and Kenneth A. Small (1993) "Is the Journey to Work Explained by Urban Structure?" *Urban Studies*, 30: 1485-1500.

Glaister, Stephen (1974) "Generalised Consumer Surplus and Public Transport Pricing." *Economic Journal*, **84**:849-867.

Glaister, Stephen (1986) "Bus Deregulation, Competition, and Vehicle Size." *Journal of Transport Economics and Policy,* **20**:217-244.

Glaister, Stephen (1997) "Deregulation and Privatisation: British Experience". in De Rus, G., and C. Nash (eds.), *Recent Developments in Transport Econonomics*. 135-197

Glaister, Stephen, and David Lewis (1978) "An Integrated Fares Policy for Transport in London." *Journal of Public Economics,* **9**:341-355.

Glazer, Amihai (1981) "Congestion Tolls and Consumer Welfare." *Public Finance*, **36**:77-83.

Glazer, Amihai, and Rafael Hassin (1983) "The Economics of Cheating in the Taxi Market." *Transportation Research Part A,* **7**:25-31.

Glazer, Amihai, and Esko Niskanen (1992) "Parking Fees and Congestion." *Regional Science and Urban Economics,* **22**:123-132.

Goh, Mark (2002) "Congestion Management and Electronic Road Pricing in Singapore." *Journal of Transport Geography*, **10**:29-38.

Golob, Thomas F. (2003), "Structural Equation Modeling for Travel Behavior Research," *Transportation Research B*, 37: 1-25.

Golob, Thomas F., Martin J. Beckmann, and Yacov Zahavi (1981) "A Utility-Theory Travel Demand Model Incorporating Travel Budgets." *Transportation Research Part B,* **5**:375-389.

Golob, Thomas F., Ryuishi Kitamura, and Lyn Long, (eds.) (1997) *Panels for Transportation Planning: Methods and Applications*. Kluwer Academic Press.

Golob, Thomas F., and Amelia C. Regan (2001) "Impacts of Highway Congestion on Freight Operations: Perceptions of Trucking Industry Managers," *Transportation Research A*, **35**:577-599.

Gómez-Ibáñez, José A. (1985) "Transportation Policy as a Tool for Shaping Metropolitan Development". in Keeler, T. E. (eds.), *Research in Transportation Economics, Vol.2*. 55-81, Greenwich, Connecticut: JAI Press.

Gómez-Ibáñez, José A. (1996) "Big-City Transit Ridership, Deficits, and Politics: Avoiding Reality in Boston." *Journal of the American Planning Association,* **62**:30-50.

Gómez-Ibáñez, José A., and Gary R. Fauth (1980) "Downtown Auto Restraint Policies: The Costs and Benefits for Boston." *Journal of Transport Economics and Policy,* **14**:133-153.

Gómez-Ibáñez, José A., and John R. Meyer (1990) "Privatizing and Deregulating Local Public Services: Lessons from Britain's Buses." *Journal of the American Planning Association,* **56**:9-21.

Gómez-Ibáñez, José A., and John R. Meyer (1993) *Going Private: The International Experience with Transport Privatization* The Brookings Institution, Washington.

Gómez-Ibáñez, José A., William B. Tye, and Clifford Winston, eds. (1999) *Essays in Transportation Economics and Policy: A Handbook in Honor of John R. Meyer*. Brookings Institution, Washington.

Goodwin, Phil B. (1989) "The Rule of Three: A Possible Solution to the Political Problem of Competing Objectives for Road Pricing." *Traffic Engineering and Control*, **30**:495-497.

Goodwin, P.B. (1992) "A Review of New Demand Elasticities with Special Reference to Short and Long run Effects of Price Changes." *Journal of Transport Economics and Policy,* **26**:155-169.

Goodwin, Phil B. (1996) "Empirical Evidence on Induced Traffic." *Transportation,* **23**:35-54.

Gordon, Peter, Ajay Kumar, and Harry W. Richardson (1989) "The Influence of Metropolitan Spatial Structure on Commuting Time," *Journal of Urban Economics*, 26: 138-151.

Gordon, Peter, and Harry W. Richardson (1994) "Congestion Trends in Metropolitan Areas." In: National Research Council (1994), *Volulme 2: Commissioned Papers*, pp. 1-31.

Gordon, Peter, and Richard Willson (1984) "The Determinants of Light-Rail Transit Demand -- An International Cross-Sectional Comparison." *Transportation Research Part A,* **18**:135-140.

Graham, Daniel J., and Stephen Glaister (2002) "The Demand for Automobile Fuel: A Survey of Elasticities." *Journal of Transport Economics and Policy,* **36**:1-26.

Greenberg, H. (1959) "An Analysis of Traffic Flow." *Operations Research*, **7**:78-85.

Greene, David L. (1992) "Vehicle Use and Fuel Economy: How Big is the Rebound Effect?" *Energy Journal,* **13**:117-143.

Greene, David L., Donald W. Jones, and Mark A. Delucchi (eds.) (1997) *The Full Costs and Benefits of Transportation: Contributions to Theory, Method and Measurement*. Berlin: Springer-Verlag.

Greening, Lorna A., David L. Greene, and Carmen Difiglio (2000) "Energy Efficiency and Consumption ? The Rebound Effect ? A Survey." *Energy Policy* **28**:389-401.

Greenshields, B.D. (1935) "A Study of Traffic Capacity." *Highway Research Board Proceedings*, **14**:448-477.

Gunn, Hugh F. (1981) "Travel Budgets -- A Review of Evidence and Modelling Implications." *Transportation Research Part A,* **5**:7-24.

Gunn, Hugh (2001) "Spatial and Temporal Transferability of Relationships between Travel Demand, Trip Cost and Travel Time," *Transportation Research Part E: Logistics and Transportation Review*, **37**:163-189.

Guttman, Joel (1975) "Avoiding Specification Errors in Estimating the Value of Time." *Transportation,* **1**:19-42.

Gwilliam, Ken M. (1987) "Deregulation , Commercialisation, and Privatisation: Transport Under the Conservatives, 1979-1987". in Harrison, A., and J. Gretton (eds.), *Transport UK 1987: An Economic, Social and Policy Audit*. 7-19, Newbury, England: Policy Journals.

Gwilliam, Ken M., Peter J. Mackie, and Christopher A. Nash (1985) "Deregulating the Bus Industry in Britain - (B) The Case Against." *Transport Reviews,* **5**:105-132.

Gwilliam, Ken M., and D. M. Van de Velde (1990) "The Potential for Regulatory Change in European Bus Markets." *Journal of Transport Economics and Policy,* **24**:333-350.

Haight, Frank (1963) *Mathematical Theories of Traffic Flow*. New York: Academic Press.

Hall, Fred L. (2002) "Traffic Stream Characteristics." In: Turner-Fairbank Highway Research Center, *Traffic Flow Theory: A State-of-the-Art Report*, http://www.tfhrc.gov/its/tft/tft.htm, accessed May 2004.

Hall, Fred L., Brian L. Allen, and Margot A. Gunter (1986) "Empirical Analysis of Freeway Flow-Density Relationships." *Transportation Research Part A,* **20**:197-210.

Hall, Fred L., and Lisa M. Hall (1990) "Capacity and Speed Flow Analysis of the QEW in Ontario." *Transportation Research Record,* **1287**:108-118.

Hall, Fred L., V.F. Hurdle and J.H. Banks (1992) "Synthesis of Recent Work on the Nature of Speed-Flow and Flow-Occupancy (or Density) Relationships for Freeways." *Transportation Research Record*, **365**:12-18.

Hall, Peter, and Carmen Hass-Klau (1985) *Can Rail Save the City? The Impacts of Rail Rapid Transit and Pedestrianisation on British and German Cities*. Aldershot, U.K.: Gower.

Hall, Randolph W., ed. (2003), *Handbook of Transportation Science* , Boston: Kluwer.

Hall, Randolph and Cenk Caliskan (1999) "Design and evaluation of an automated highway system with optimized lane assignment" *Transportation Research* **7C** (1) 1-15.

Hansen, Mark, and Yuanlin Huang (1997) "Road Supply and Traffic in California Urban Areas." *Transportation Research Part A,* **31**:205-218.

Hardin, Garrett (1968) "The Tragedy of the Commons." *Science,* **62**:1243-1248.

Harker, Patrick T. (1988) "Private Market Participation in Urban Mass Transportation: Application of Computable Equilibrium Models of Network Competition." *Transportation Science,* **22**:96-111.

Hart, Stanley (1985) "An Assessment of the Municipal Costs of Automobile Use," Department of Civil Engineering, University of California, Irvine. Unpublished paper.

Hastings, N.A.J., and J.B. Peacock (1975) *Statistical Distributions: A Handbook for Students and Practitioners*. London: Butterworth.

Hausman, Jerry A., and Paul A. Ruud (1987) "Specifying and testing econometric models for rank-ordered data." *Journal of Econometrics,* **34**:83-104.

Hausman, Jerry A., and David A. Wise (1978) "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences." *Econometrica,* **46**:403-426.

Hearn, Donald W., and Motakuri V. Ramana (1998) "Solving Congestion Toll Pricing Models." In: Marcotte, Patrice, and Sang Nguyen (eds.), *Equilibrium and Advanced Transportation Modelling*, Kluwer, Dordrecht.

Heckman, James J. (1979) "Sample Selection Bias as a Specification Error." *Econometrica* **47**: 153-162.

Heckman, James J. (1981) "Statistical Models for Discrete Panel Data". in Manski, C. F., and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. 114-178, Cambridge, Mass.: MIT Press.

Henderson, J. Vernon (1974) "Road Congestion: a Reconsideration of Pricing Theory." *Journal of Urban Economics*, **1**:346-365.

Henderson, J. Vernon (1977) *Economic Theory and the Cities*. New York: Academic Press.

Henderson, J. Vernon (1992) "Peak shifting and cost-benefit miscalculations" *Regional Science and Urban Economics* **22** (1) 103-121.

Hendrickson, Chris, and George Kocur (1981) "Schedule Delay and Departure Time Decisions in a Deterministic Model." *Transportation Science,* **15**:62-77.

Hendrickson, Chris, Daniel Nagin, and Edward Plank (1983) *Characteristics of Travel Time and Dynamic User Equilibrium for Travel-to-Work.* Proceedings of the Eighth International Symposium on Transportation and Traffic Theory, Toronto: University of Toronto Press. DEL??

Hensher, David A. (1978) "Valuation in Journey Attributes: Some Existing Empirical Evidence". in Hensher, D. A., and Q. Dalvi (eds.), *Determinants of Travel Choice,*. 203-265, New York: Praeger.

Hensher, David A. (1986) "Sequential and Full Information Maximum Likelihood Estimation of a Nested Logit Model." *Review of Economics and Statistics,* **56**:657-667.

Hensher, David A. (1988a) "Productivity in Privately Owned and Operated Bus Firms in Australia". in Dodgson, J. S., and N. Topham (eds.), *Bus Deregulation and Privatisation*. 141-170, Aldershot, U.K.: Gower.

Hensher, David A. (1988b) "Some Thoughts on Competitive Tendering in Local Bus Operations." *Transport Reviews,* **8**:363-372.

Hensher, David A. (1989) "Behavioural and Resource Values of Travel Time Savings: a Bicentennial Update." *Australian Road Research,* **19**:223-229.

Hensher, David A. (1994) "Stated preference analysis of travel choices: the state of practice." *Transportation,* **21**:107-133. .

Hensher, David A. (1997) "Behavioral Value of Travel Time Savings in Personal and Commercial Automobile Travel," in Greene et al. (1997): 245-279.

Hensher, David A. (2001) "The Valuation of Commuter Travel Time Savings for Car Drivers: Evaluating Alternative Model Specifications," *Transportation*, 28: 101-118.

Hensher, David A., and Kenneth J. Button, (eds.) (2000) *Handbook of Transport Modelling*. *Handbooks in Transport*, Vol. 1, Pergamon, Amsterdam.

Hensher, David A., and Kenneth J. Button, (eds.) (2001) *Handbook of Transport Systems and Traffic Control*. *Handbooks in Transport*, Vol. 3, Pergamon, Amsterdam.

Herman, R., E.W. Montroll, R.B. Potts and R.W. Rothery (1958) "Traffic dynamics: analysis of stability in car following" *Operations Research* **17E** 86-106.

Heseltine, P.M., and D.T. Silcock (1990) "The Effects of Bus Deregulation on Costs." *Journal of Transport Economics and Policy,* **24**:239-254.

Hess, Daniel B., Brian D. Taylor, and Allison C. Yoh (2005) "Light Rail Lite or Cost-Effective Improvements to Bus Service?" *Transportation Research Record*, **1927**:22-30.

Highway Research Board (1965) *Highway Capacity Manual*. Washington, D.C.: Highway Research Board.Special Report 87.

Hills, Peter (1993) "Road Congestion Pricing: When Is It a Good Policy? A Comment." *Journal of Transport Economics and Policy*, **27**:91-99.

Holden, David J. (1989) "Wardrop's Third Principle: Urban Traffic Congestion and Traffic Policy." *Journal of Transport Economics and Policy,* **23**:239-262.

Hoogendoorn, Serge P., and Piet H.L. Bovy (1998) "Modeling Multiple User-Class Traffic." *Transportation Research Record*, **1644**:57-69.

Horowitz, Joel L. (1980) "The Accuracy of the Multinomial Logit Model as an Approximation to the Multinomial Probit Model of Travel Demand." *Transportation Research Part B,* **14**:331-341.

Hotelling, Harold (1929) "Stability in Competition." *Economic Journal,* **39**:41-57.

Hotelling, Harold (1938) "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates." *Econometrica,* **6**:242-269.

Hu, Pat S., and Timothy R. Reuscher (2004) *Summary of Travel Trends: 2001 National Household Travel Survey*, US Federal Highway Administration, Washington, D.C.

Imada, T., and A.D. May (1985) *FREQ8PE - A Freeway Corridor Simulation and Ramp Metering Optimization Model.* Report UCB-ITS-RR85-10, Univ. of California, Berkeley,

Imbens, Guido W., and Tony Lancaster (1994) "Combining Micro and Macro Data in Microeconometric Models." *Review of Economics Studies,* **61**:655-680.

Inman, Robert P (1978) "A Generalized Congestion Function for Highway Travel." *Journal of Urban Economics,* **5**:21-34.

International Energy Agency (2002) *Bus Systems for the Future: Achieving Sustainable Transport Worldwide*. IEA and Organisation for Economic Co-operation and Development, Paris.

Jansson, Jan Owen (1980) "A Simple Bus Line Model for Optimisation of Service Frequency and Bus Size." *Journal of Transport Economics and Policy,* **14**:53-80.

Jansson, Jan Owen (1984) *Transport System Optimization and Pricing*. Chichester, UK: John Wiley & Sons.

Japan Ministry of Health, Labour and Welfare (1999) *Final Report of Monthly Labour Survey: July 1999*. http://www.mhlw.go.jp/english/database/db-l/, accessed Dec. 30, 2004.

Japan Research Institute Study Group on Road Investment Evaluation (2000) *Guidelines for the Evaluation of Road Investment Projects*. Japan Research Institute, Tokyo, May.

Jara-Díaz, Sergio R. (1982) "The Estimation of Transport Cost Functions: A Methodological Review." *Transport Reviews,* **2**:257-278.

Jara-Díaz, Sergio R. (1986) "On the Relation Between Users' Benefits and the Economic Effects of Transportation Activities." *Journal of Regional Science*, **26**:379-391.

Jara-Díaz, Sergio R. (2000) "Allocation and Valuation of Travel-time Savings," in Hensher, D. A., and K. J. Button (eds.), *Handbook of Transport Modelling*, 303-319.

Jara-Díaz, Sergio R. (2003) "On the Goods-Activities Technical Relations in the Time Allocation Theory." *Transportation*, **30**:245-260.

Jenkins, Glenn P. (1997) "Project Analysis and the World Bank." *American Economic Review Papers and Proceedings*, **87**(2):38-42.

Johansson, Olof, Lee Schipper, and Mean Annual Driving Distance (1997) "Measuring the Long-Run Fuel Demand of Cars: Separate Estimations of Vehicle Stock, Mean Fuel Intensity." *Journal of Transport Economics and Policy,* **31**:277-292.

Johnson, M. Bruce (1966) "Travel Time and the Price of Leisure." *Western Economic Journal,* **4**:135-145.

Jones-Lee, Michael (1990) "The Value of Transport Safety." *Oxford Review of Economic Policy,* **6**:39-60.

Jones-Lee, Michael W., M. Hammerton, and P.R. Philips (1985) "The Value of Safety: Results of a National Sample Survey." *Economic Journal,* **95**:49-72.

Jun, Myung-Jin (2004) "The Effects of Portland's Urban Growth Boundary on Urban Development Patterns

and Commuting," *Urban Studies*, 41: 1333-1348.

Kain, John F. (1992) *Increasing the Productivity of the Nation's Urban Transportation Infrastructure: Measures to Increase Transit Use and Carpooling.* Report DOT-T-92-17, U.S. Federal Transit Administration, Washington, D.C,

Kain, John F. (1999) "The Urban Transportation Problem: A Reexamination and Update". In: Gómez-Ibáñez, Tye, and Winston (1999), pp. 359-401.

Kain, John F., and Zhi Liu (2002) "Efficiency and Locational Consequences of Government Transport Policies and Spending in Chile". in Glaeser, E. L., and J. R. Meyer (eds.), *Chile: Political Economy of urban Development*. ch. 5**:** 105-195, Cambridge, Mass.: Harvard University Press.

Kanemoto, Yoshitsugu (1987) "Externalities in Space". in Miyao, T., and Y. Kanemoto (eds.), *Urban Dynamics and Urban Externalities,*. 1-42, Chur, Switzerland: Harwood Academic Publishers.

Kaplow, Louis (1996) "The Optimal Supply of Public Goods and the Distortionary Cost of Taxation." *National Tax Journal*, **49**:513-533.

Kawamura, Kazuya (2000) "Perceived Value of Time for Truck Operators." *Transportation Research Record*, **1725**:31-36.

Keeler, Theodore E., and Kenneth A. Small (1977) "Optimal Peak-Load Pricing, Investment, and Service Levels on Urban Expressways." *Journal of Political Economy,* **85**:1-25.

Keeler, Theodore E., Kenneth A. Small, and Associates (1975) "The Full Costs of Urban Transport, Part III: Automobile Costs and Final Intermodal Cost Comparisons.," Monograph No. 21, Institute of Urban and Regional Development, University of California, Berkeley

Keon, Chin Kian (2002) "Road Pricing: Singapore's Experience." Paper presented to the 3rd IMPRINT-EUROPE seminar, Brussels.

Kerner, B.S., and H. Rehborn (1997) "Experimental Properties of Phase Transitions in Traffic Flow." *Physical Review Letters,* **79**:4030-4033.

Keyes, Dale L. (1982) "Energy for Travel: The Influence of Urban Development Patterns," *Transportation Research A*, 16: 65-70.

Kirby, Ronald F., Kiran U. Bhatt, Michael A. Kemp, Robert G. McGillivray, and Martin Wohl (1975) *Para-Transit: Neglected Options for Urban Mobility*. Washington, D.C.: The Urban Institute.

Kitamura, Ryuichi (2000) "Longitudinal Methods". in Hensher, D. A., and K. J. Button (eds.), *Handbook of Transport Modelling*. 113-129

Klein, Daniel B., Adrian T. Moore, and Binyam Reja (1997) *Curb Rights: A Foundation for Free Enterprise in Urban Transit*. Washington, D.C.: Brookings Institution Press.

Knight, Frank (1924) "Some Fallicies in the Interpretation of Social Costs." *Quarterly Journal of Economics,* **38**:582-606.

Knight, Robert L., and Lisa L. Trygg (1977) *Land Use Impacts of Rapid Transit: Implications of Recent Experience*. Springfield, Virginia: National Technical Information Service.U.S. Dept. of Transportation Report No. DOT-TPI-10-77-29.

Kockelman, Kara M. (2001) "A Model for Time- and Budget-Constrained Activity Demand Analysis." *Transportation Research Part B,* **35**:255-269.

Koppelman, Frank S., and Geoffrey Rose (1985) "Geographic Transfer of Travel Choice Models: Evaluation and Procedures". in Hutchinson, B. G., P. Nijkamp, and M. Batty (eds.), *Optimization and*

Discrete Choice in Urban Systems: Proceedings of the International Symposium on New Directions in Urban Systems Modelling. 272-309, Berlin: Springer-Verlag.

Koppelman, Frank S., and Vaneet Sethi (2000) "Closed-Form Discrete-Choice Models." In: Hensher and Button (2000), pp. 211-227.

Koppelman, Frank S., and Chieh-Hua Wen (2000) "The Paired Combinatorial Logit Model: Properties, Estimation and Application." *Transportation Research Part B,* **34**:75-89.

Koppelman, Frank S., and Chester G. Wilmot (1982) "Transferability Analysis of Disaggregate Choice Models." *Transportation Research Record,* **895**:18-24.

Krammes, Raymond A., and Kenneth W. Crowley (1986) "Passenger Car Equivalents for Trucks on Level Freeway Segments." *Transportation Research Record,* **1091**:10-17.

Kraus, Marvin (1981) "Scale Economies Analysis for Urban Highway Networks." *Journal of Urban Economics,* **9**:1-22.

Kraus, Marvin (1982) "Highway Pricing and Capacity Choice under Uncertain Demand." *Journal of Urban Economics,* **2**:122-128.

Kraus, Marvin (1991) "Discomfort Externalities and Marginal Cost Transit Fares." *Journal of Urban Economics,* **29**:249-259.

Kraus, Marvin, Herbert Mohring, and Thomas Pinfold (1976) "The Welfare Costs of Nonoptimum Pricing and Investment Policies for Freeway Transportation." *American Economic Review,* **66**:532-547.

Kraus, Marvin, and Yuichiro Yoshida (1999) "A Model of a Welfare-Maximizing Urban Mass Transit Authority,″ working paper, Boston College

Kroes, Eric P, Robert W. Antonisse, and Sten Bexelius (1987) "Return to the Peak?" in, *Transportation Planning Methods*. Proceedings of Seminar C held at the PTRC Summer Annual Meeting**:** 233-245, University of Bath, England. Bath: PTRC Education and Research Services, Ltd., on behalf of Planning and Transport Research and Computation (International) Co. Ltd.

Krupnick, Alan J. (2004) *Valuing Health Outcomes: Policy Choices and Technical Issues*. RFF Report, Resources for the Future, Washington, D.C., March. http://www.rff.org/rff/Publications/Reports.cfm, accessed 29 April 2005.

La Croix, Sumner J., James Mak, and Walter Miklius (1992) "Evaluation of Alternative Arrangements for the Provision of Airport Taxi Service." *Logistics and Transportation Review,* **28**:147-166.

Lago, Armando M., Patrick D. Mayworm, and J. Matthew McEnroe (1981) "Further Evidence on Aggregate and Disaggregate Transit Fare Elasticities." *Transportation Research Record,* **799**:42-47.

Lam, Terence C., and Kenneth A. Small (2001) "The Value of Time and Reliability: Measurement from a Value Pricing Experiment." *Transportation Research Part E: Logistics and Transportation Review,* **37**:231-251.

Larsen, Odd I. (1993) "Road Investment with Road Pricing – Investment Criteria and the Revenue/Cost Issue". In: A. Talvitie, D. Hensher, and M. E. Beesley. (eds.), *Privatization and Deregulation in Passenger Transportation*. Second International Conference on Privatization and Deregulation in Passenger Transportation c/o Viatek Ltd**:** 273-281. Espoo, Finland.

Lave, Charles A. (1969) "A Behavioral Approach to Modal Split Forecasting." *Transportation Research,* **3**:463-480.

Lave, Charles A., (eds.) (1985) *Urban Transit: The Private Challenge to Public Transportation*. Pacific Studies in Public Policy, Cambridge, Mass.: Ballinger.

Layard, Richard (1977) "The Distributional Effects of Congestion Taxes." *Economica*, **44**:297-304.

Layard, Richard, and Stephen Glaister, eds. (1994) *Cost-Benefit Analysis*. Cambridge Univ. Press, Cambridge, UK.

Lee, Douglass (1972) *The Costs of Private Automobile Usage to the City of San Francisco*. Report No. UMTA-CAL-11-0006-72-4, Springfield, Virginia: National Technical Information Service,

Lee, Douglass (1989) "Transit Cost and Performance Measurement." *Transport Reviews*, **9**:147-170.

Lee, Douglass (1993) "A Market-Oriented Transport and Land Use System: How Different Would it Be?" in Talvitie, A., D. Hensher, and M. Beesley (eds.), *Privatization and Deregulation in Passenger Transportation*. 219-246, Espoo, Finland: Viatek Ltd.

Lee, Douglass (1999) "Induced Traffic and Induced Demand." In: U.S. Federal Highway Administration, *Highway Economic Requirements System Technical Report*, March 1999, Washington, Appendix B.

Lerman, Steven R., and Charles F. Manski (1979) "Sample Design for Discrete Choice Analysis of Travel Behavior: The State of the Art." *Transportation Research Part A*, **13**:29 44.

Levinson, David M. (1998) "Road Pricing in Practice." In: Button and Verhoef (1998), pp. 14-38.

Levinson, David M. (2001) "Why states toll: an empirical model of finance choice" *Journal of Transport Economics and Policy* **35** (2) 223-238.

Levinson, Herbert, Samuel Zimmerman, Jennifer Clinger, Scott Rutherford, Eric Bruhn, James Gast, Rodney L. Smith, John Cracknell, and Richard Soberman, *Bus Rapid Transit*. Transit Cooperative Research Program Report 90 (Volumes 1 and 2). Transportation Research Board, Washington.

Levinson, Herbert S, Edward J. Regan III, and Eugene J. Lessieu (1980) "Estimating Behavioral Response to Peak-Period Pricing." *Transportation Research Record*, **767**:21-26.

Levitt, Steven D., and Jack Porter (2001) "How Dangerous Are Drinking Drivers?" *Journal of Political Economy*, **109**: 1198-1237.

Li, Michael Z.F. (2002) "The Role of Speed-Flow Relationship in Congestion Pricing Implementation with an Application to Singapore." *Transportation Research Part B*, **36**:731-754.

Lighthill, M.H., and G.B. Whitham (1955) *On Kinematic Waves, II: A Theory of Traffic Flow on Long Crowded Roads*. Proceedings of the Royal Society, London.

Lindberg, Gunnar (2001) "Traffic Insurance and Accident Externality Charges," *Journal of Transport Economics and Policy*, **35**: 399-416.

Lindberg, P.O., E.A. Eriksson, and L.-G. Mattsson (1995) "Invariance of Achieved Utility in Random Utility Models." *Environment & Planning Part A*, **27**:121-142.

Linder, Staffan B. (1970) *The Harried Leisure Class*. New York: Columbia Univ. Press.

Lindsey, Robin (2006) "Do Economists Reach a Conclusion on Road Pricing? The Intellectual History of an Idea," *Econ Journal Watch*, **3**:292-379. www.econjournalwatch.org.

Lindsey, Robin and Erik T. Verhoef (2000) "Congestion Modelling." In: Hensher and Button (2000), pp. 353-373.

Lindsey, Robin, and Erik T. Verhoef (2001) "Traffic Congestion and Congestion Pricing." In Hensher and Button (2001), pp. 77-105.

Lipsey, Richard G., and Kelvin J. Lancaster (1956) "The General Theory of Second Best." *Review of Economic Studies*, **24**:11-32.

Litman, Todd (2005) *Transportation Cost and Benefit Analysis.* Victoria Transport Policy Institute, Victoria, British Columbia. http://www.vtpi.org/documents/transportation.php, accessed 2 June 2005.

Little, I.M.D., and J.A. Mirrlees (1968) *Manual of Industrial Project Analysis for Developing Countries, II.*, Paris: Organization for Economic Co-operation and Development.

Liu, Louie Nan, and John F. McDonald (1998) "Efficient Congestion Tolls in the Presence of Unpriced Congestion: A Peak and Off-Peak Simulation Model." *Journal of Urban Economics,* **44**:352-366.

Louviere, Jordan J., D.H. Henley, G. Woodworth, R.J. Meyer, I.P. Levin, J.W. Stoner, D. Curry, and D.A. Anderson (1981) "Laboratory-Simulation versus Revealed-Preference Methods for Estimating Travel Demand Models." *Transportation Research Record,* **794**:42-51.

Louviere, Jordan J., and David A. Hensher (2001) "Combining Sources of Preference Data," in: David A. Hensher (ed.), *Travel Behaviour Research: The Leading Edge*, Pergamon, Oxford: 125-144.

Louviere, Jordan J., David A. Hensher, and Joffre D. Swait (2000) *Stated Choice Methods: Analysis and Application*, Cambridge University Press, Cambridge.

Luk, James, and Stephen Hepburn (1993) "New Review of Australian Travel Demand Elasticities." *Research Report ARR 249.*

Mackett, Roger L. (1985a) "Integrated Land Use – Transport Models." *Transport Reviews,* **5**:325-343.

Mackett, Roger L. (1985b) "Modelling the Impact of Rail Fare Increases." *Transportation,* **12**:293-312.

Mackie, P.J., and P.W. Bonsall (1989) "Traveller Response to Road Improvements: Implications for User Benefits." *Traffic Engineering and Control,* **30**:411-416.

Mackie, P.J., S. Jara-Díaz, and A.S. Fowkes (2001) "The Value of Travel Time Savings in Evaluation." *Transportation Research Part E: Logistics and Transportation Review*, **37**:91-106.

Mackie, P.J., M. Wardman, A.S. Fowkes, G. Whelan, J. Nellthorp, and J. Bates (2003) "Values of Travel time Savings in the UK: Summary Report." Report to the UK Department for Transport. Institute for Transport Studies, University of Leeds, UK, January. http://www.dft.gov.uk/stellent/groups/dft_econappr/documents/page/dft_econappr_022708-01.hcsp, accessed Dec. 30, 2004.

Mahmassani, Hani S. (2000) "Trip Timing." In Hensher and Button (2000), pp. 393-407.

Mahmassani, Hani S. (ed.) (2002) *In Perpetual Motion: Travel Behavior Research Opportunities and Application Challenges*. Amsterdam: Elsevier.

Mahmassani, Hani S. and Robert Herman (1984) "Dynamic User Equilibrium Departure Time and Route Choice on Idealized Traffic Arterials." *Transportation Science*, **18**:362-384.

Mannering, Fred L., and David A. Hensher (1987) "Discrete/Continuous Econometric Models and Their Application to Transport Analysis." *Transport Reviews,* **7**:227-244.

Manski, Charles F., and Steven R. Lerman (1977) "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica,* **45**:1977-1988.

Marcotte, Patrice, and Sang Nguyen (eds.) (1998) *Equilibrium and Advanced Transportation Modelling*. Kluwer, Dordrecht.

May, Adolf D. (1990) *Traffic Flow Fundamentals*, Prentice-Hall, Upper Saddle River, New Jersey.

May, Adolf D., and Hartmut E.M. Keller (1966) "A Deterministic Queueing Model." Paper presented at the Operations Research Society of America, Twenty-Ninth National Meeting, Santa Monica, California.

May, Anthony D. (1992) "Road Pricing: An International Perspective." *Transportation,* **19**:313-333.

May, Anthony D., R. Liu, S.P. Shepherd, and A. Sumalee (2002) "The Impact of Cordon Design on the Performance of Road Pricing Schemes." *Transport Policy*, **9**:209-220.

May, Anthony D., and Dave S. Milne (2000) "Effects of Alternative Road Pricing Systems on Network Performance." *Transportation Research Part A*, **34**:407-436.

May, Anthony D., Simon P. Shepherd, and John J. Bates (2000) "Supply Curves for Urban Road Networks." *Journal of Transport Economics and Policy*, **34**:261-290.

Mayeres, Inge, and Stef Proost (2001) "Marginal Tax Reform, Externalities and Income Distribution." *Journal of Public Economics*, **79**:343-363.

McCarthy, Patrick S. (2001) *Transportation Economics: Theory and practice: A Case Study Approach,*. Malden, Mass.: Blackwell Publishers.

McCarthy, Patrick S., and Richard Tay (1993) "Economic Efficiency *vs* Traffic Restraint: A Note on Singapore′s Area License Scheme." *Journal of Urban Economics*, **34**:96-100.

McClenahan, J.W., M. Elms D. Nichols, and P.H. Bly (1978) *Two Methods for Estimating the Crew Costs of Bus Service*. Special Report 364, Corwthorne, England: Transport and Road Research Laboratory,

McCubbin, Donald R., and Mark A. Delucchi (1999) "The Health Costs of Motor-Vehicle-Related Air Pollution." *Journal of Transport Economics and Policy,* **33**:253-286.

McDonald, John F., Edmond L. d'Ouville, and Louie Nan Liu (1999) *Economics of Urban Highway Congestion and Pricing*, Kluwer, Boston.

McFadden, Daniel (1973) "Conditional Logit Analysis of Qualitative Choice Behavior". in Zarembka., P. (eds.), *Frontiers in Econometrics*. 105-142., New York: Academic Press.

McFadden, Daniel (1978) "Modelling the Choice of Residential Location". in Karlqvist, A., L. Lundqvist, F. Snickars, and J. W. Weibull (eds.), *Spatial Interaction Theory and Planning Models,*. 75-96, Amsterdam: North-Holland.

McFadden, Daniel (1981) "Econometric Models of Probabilistic Choice". in Manski, C. F., and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. 198-272, Cambridge, Mass.: MIT Press.

McFadden, Daniel (2001) "Economic Choices." *American Economic Review,* **91**:351-378.

McFadden, Daniel, and Fred Reid (1976) "Aggregate Travel Demand Forecasting from Disaggregated Behavioral Models." *Transportation Research Record,* **534**:24-37.

McFadden, Daniel, Antti P. Talvitie, and Associates (1977) *Demand Model Estimation and Validation. Urban Travel Demand Forecasting Project*. Phase I Final Report Series, Vol. V, Berkeley: University of California Institute of Transportation Studies.Special Report UCB-ITS-SR-77-9.

McFadden, Daniel, and Kenneth Train (2000) "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, 15: 447-470.

Meland, Solveig (1995) "Generalised and Advanced Urban Debiting Innovations: The GAUDI Project 3 – the Trodheim Toll Ring." *Traffic Engineering and Control*, **36**:150-155.

Meyer, John R., and José A. Gómez-Ibáñez (1981) *Autos, Transit, and Cities*. A Twentieth Century Fund report, Cambridge, Mass.: Harvard University Press,

Meyer, J. R., J.F. Kain, and M. Wohl (1965) *The Urban Transportation Problem*. Cambridge, Mass.: Harvard University Press.

Miller, Ted R. (1993) "Costs and Functional Consequences of U.S. Roadway Crashes," *Accident Analysis and Prevention*, **25**: 593-607.

Miller, Ted R., Rebeccas S. Spicer, and David T. Levy (1999) "How Intoxicated Are Drivers in the United States? Estimating the Extent, Risks and Costs per Kilometer of Driving by Blood Alcohol Level." *Accident Analysis and Prevention*, **31**: 515-523.

Mills, David E. (1981) "Ownership Arrangements and Congestion-Prone Facilities." *American Economic Review*, **71**:493-502.

Milne, David S., Esko Niskanen, and Erik T. Verhoef (2000) *Operationalisation of Marginal Cost Pricing within Urban Transport*, AFFORD Deliverable 1, European Commission 4th Framework – Transport RTD) Government Institute for Economic Research (VATT), Research Report 63, Helsinki., http://data.vatt.fi/afford/reports-del1.html, accessed August 8, 2005.

Mishan, E.J. (1988) *Cost-Benefit Analysis: An Informal Introduction.*, London: Allen and Unwin.

Mogridge, M.J.H., J. Bird, D.J. Holden, and G.C. Terzis (1987) "The Downs/Thomson Paradox and the Transportation Planning Process." *International Journal of Transport Economics (Revista Internazionale di Economia dei Trasporti),* **14**:283-311.

Mohktarian, Patricia L., Francisco J. Samaniego, Robert H. Shumway, and Neil H. Willits (2002) "Revisiting the Notion of Induced Traffic through a Matched-Pairs Study." *Transportation*, **29**:193-220.

Mohring, Herbert (1961) "Land Values and the Measurement of Highway Benefits." *Journal of Political Economy,* **69**:236-249.

Mohring, Herbert (1965) "Urban Highway Investments". in Dorfman, R. (eds.), *Measuring Benefits of Government Investment*. 231-275, Washington, D.C.: The Brookings Institution.

Mohring, Herbert (1970) "The Peak Load Problem with Increasing Returns and Pricing Constraints." *American Economic Review,* **60**:693-705.

Mohring, Herbert (1972) "Optimization and Scale Economies in Urban Bus Transportation." *American Economic Review,* **62**:591-604.

Mohring, Herbert (1976) *Transportation Economics*. Cambridge, Mass.: Ballinger

Mohring, Herbert (1985) "Profit Maximization, Cost Minimization and Pricing for Congestion-Prone Facilities." *Logistics and Transportation Review,* **21**:27-36.

Mohring, Herbert, and Mitchell Harwitz (1962) *Highway Benefits: An Analytical Framework.*, Evanston, Illinois: Northwestern University Press.

Mohring, Herbert, and Harold F. Williamson, Jr. (1969) "Scale and 'Industrial Reorganisation' Ecocomies of Transport Improvements." *Journal of Transport Economics and Policy*, **3**:251-271.

Moore, Terry, and Randy Pozdena (2004) "Framework for an Economic Evaluation of Transportation Investments." In: Evangelos Bekiaris and Yuko J. Nakanishi, eds., *Economic Impacts of Intelligent Transportation Systems: Innovations and Case Studies*. In series *Research in Transportation Economics*, **8**:17-45.

Morlok, Edward K., and Philip A. Viton (1980) "Self-Sustaining Public Transportation Services." *Transportation Policy and Decision Making,* **1**:169-194.

Morlok, Edward K., and Philip A. Viton (1985a) "The Comparative Costs of Public and Private Providers of Mass Transit". in Lave, C. A. (eds.), *Urban Transit: The Private Challenge to Public Transportation*. 233-253, Pacific Studies in Public Policy. Cambridge, Mass.: Ballinger.

Morlok, Edward K., and Philip A. Viton (1985b) "Recent Experience with Successful Private Transit in Large U.S. Cities". in Lave., C. A. (eds.), *Urban Transit: The Private Challenge to Public Transportation*. 121-149, Pacific Studies in Public Policy. Cambridge, Mass.: Ballinger.

Morrison, Steven A., and Clifford Winston (1987) "Empirical Implications and Tests of the Contestability Hypothesis." *Journal of Law and Economics,* **30**:53-66.

Mrozek, Janusz R., and Laura O. Taylor (2002) "What Determines the Value of Life? A Meta-Analysis," *Journal of Policy Analysis and Management*, **21**: 253-270.

Mun, Se-il (1999) "Peak-Load Pricing of a Bottleneck with Traffic Jam." *Journal of Urban Economics*, **46**:323-349.

Mun, Se-il (2002) "Bottleneck Congestion with Traffic Jam: A Reformulation and Correction of Earlier Result." Working paper, Graduate School of Economics, Kyoto University.

Mun, Se-il, Ko-ji Konishi, and Kazuhiro Yoshikawa (2003) "Optimal Cordon Pricing." *Journal of Urban Economics*, **54**:21-28.

Mun, Se-il, Ko-ji Konishi, and Kazuhiro Yoshikawa (2005) "Optimal Cordon Pricing in a Non-Monocentric City." *Transportation Research Part A*, **39**:723-736.

Munizaga, Marcela A., Rodrigo Correia, Sergio R. Jara-Díaz, and Juan de Dios Ortúzar (2004) "Valuing Time with a Joint Mode Choice Activity Model." Working paper, Dept. of Civil Engineering, University of Chile, Santiago.

Murchland, J.D. (1970) "Braess's Paradox of Traffic Flow." *Transportation Research,* **4**:391-394.

Murphy, James J., and Mark A. Delucchi (1998) "A Review of the Literature on the Social Cost of Motor Vehicle Use in the United States," *Journal of Transportation and Statistics*, **1**:15-42.

MVA Consultancy, Institute for Transport Studies of the University of Leeds, and Transport Studies Unit of the University of Oxford (1987) *The Value of Travel Time Savings*. A Report of Research Undertaken for the Department of Transport., Newbury, England: Policy Journals,

Nagurney, Anna (1999) *Network Economics: A Variational Inequality Approach*, second edition. Kluwer, Dordrecht.

Nash, Christopher A. (1974) "The Treatment of Capital Costs of Vehicles in Evaluating Road Schemes." *Transportation,* **3**:225-242.

Nash, Christopher A. (1988) "Integration of Public Transport: an Economic Assessment". in Dodgson, J. S., and N. Topham (eds.), *Bus Deregulation and Privatisation,*. 97-118, Aldershot, U.K.: Gower.

Nash, Christopher A., with partners (2003) *UNITE (UNIfication of accounts and marginal costs for Transport Efficiency) Final Report for Publication*. Report for European Commission – DG TREN, Fifth Framework Programme, Brussels, November.
http://www.its.leeds.ac.uk/projects/unite/downloads/FinalReport.doc, accessed 21 May 2005.

Nash, Christopher A., and Bryan Matthews (2005) *Measuring the Marginal Social Costs of Transport*. In series: *Research in Transportation Economics*, vol. 5, Elsevier JAI.

National Research Council (1994) *Curbing gridlock: peak-period fees to relieve traffic congestion. Volulme 1: Committee Report and Recommendations*. Transportation Research Board Special Report 242. Washington, D.C.: National Academy Press.

National Research Council (1994) *Curbing gridlock: peak-period fees to relieve traffic congestion, Volume 2: Commissioned Papers*. Transportation Research Board Special Report 242. Washington, D.C.: National Academy Press.

National Research Council (1994) *Curbing Gridlock: Peak-Period Fees to Relieve Traffic Congestion*, Transportation Research Board Special Report 242. National Academy Press, Washington, D.C.

Navrud, Ståle (2003) "State-of-the-Art on Economic Valuation of Noise." Paper presented at the

WCE/WHO Pan-European Program on Transport, Health, and Environment, Stockholm, June. http://www.fhi.se/pdf/navrud.pdf, accessed 19 May 2004.

Nelson, Gary R. (1972) "An Econometric Model of Urban Bus Transit Operations". in al., J. D. W. e. (eds.), *Economic Characteristics of the Urban Public Transportation Industry*. prepared for U.S. Department of Transportation.**:** chapter IV, Washington, D.C.: U.S. Government Printing Office, Institute for Defense Analyses.

Nelson, Jon P. (1978) *Economic Analysis of Transportation Noise Abatement*. Cambridge, Mass.: Ballinger.

Newbery, David M. (1988a) "Road User Charges in Britain." *Economic Journal,* **98**:161-176.

Newbery, David M. (1988b) "Road damage externalities and road user charges" *Econometrica* **56** 295-316.

Newbery, David M. (1989) "Cost recovery from optimally designed roads" *Economica* **56** 165-185.

Newbery, David M. (2005) "Road User and Congestion Charges." In Sijbren Cnossen (ed.), *Theory and Practice of Excise Taxation: Smoking, Drinking, Gambling, Polluting, and Driving*. Oxford: Oxford University Press, pp. 193-229.

Newell, Gordon F. (1971) *Applications of Queueing Theory.*, London: Chapman and Hall.

Newell, Gordon F. (1987) "The Morning Commute for Nonidentical Travelers." *Transportation Science,* **21**:74-88.

Newell, Gordon F. (1988) "Traffic Flow for the Morning Commute." *Transportation Science*, **22**:47-58.

Newman, Peter W.G., and Jeffrey R. Kenworthy (1989) *Cities and Automobile Dependence: An International Sourcebook*, Gower, Brookfield, Vermont.

Newman, Peter W.G., and Jeffrey R. Kenworthy (1991) "Transport and Urban Form in Thirty-two of the World's Principal Cities." *Transport Reviews,* **11**:249-272.

Nguyen, Kim Phi (1999) "Demand, Supply, and Pricing in Urban Road Transport: The Case of Ho Chi Minh City, Vietnam," *Research in Transportation Economics*, **5**:107-154.

Nielsen, Otto A. (2000) "A Stochastic Transit Assignment Model Considering Differences in Passengers' Utility Functions." *Transportation Research Part B*, **34**:377-402.

Niskanen, Esko, and Chris Nash (2004) "MC-ICAM (Implementation of Marginal Cost Pricing in Transport – Integrated Conceptual and Applied Model Analysis): Final Report," Institute for Transport Studies, University of Leeds. http://www.strafica.fi/mcicam/reports.html, accessed August 8, 2005.

Noland, Robert B. (2001) "Relationships between highway capacity and induced vehicle travel." *Transportation Research Part A*, **35**:47-72.

Noland, Robert B., and Kenneth A. Small (1995) "Travel-Time Uncertainty, Departure Time Choice, and the Cost of Morning Commutes," *Transportation Research Record*, **1493**:150-158.

Ohta, H. (2001) "Probing a Traffic Congestion Controversy: Density and Flow Scrutinized," *Journal of Regional Science*, **41**:659-680.

Olszewski, P., and W. Suchorzewski (1987) "Traffic Capacity of the City Centre." *Traffic Engineering and Control,* **28**:336-343, 348.

Oort, C.J. (1969) "The Evaluation of Travelling Time." *Journal of Transport Economics and Policy,* **3**:279-286.

Organization for Economic Co-operation and Development (OECD) (1983) *Impacts of Heavy Freight Vehicles.*, Paris: OECD.A Report Prepared by an OECD Road Research Group.

Organisation for Economic Co-operation and Development (OECD) (1987) *Toll Financing and Private Sector Involvement in Road Infrastructure Development.*, Paris: OECD.A Report Prepared by an OECD Scientific Expert Group.

Parkany, Emily (1999) *Traveler Responses to New Choices: Toll vs. Free Alternatives in a Congested Corridor*. Ph.D. dissertation, Univ. of California, Irvine.

Parry, Ian W.H. (2004) "Comparing Alternative Policies to Reduce Traffic Accidents," *Journal of Urban Economics*, **56**: 346-358.

Parry, Ian W.H., and Antonio M. Bento (2001) "Revenue Recycling and the Welfare Effects of Congestion Pricing." *Scandinavian Journal of Economics* **103**:645-671.

Parry, Ian W.H., and Kenneth A. Small (2005) "Does Britain or The United States Have the Right Gasoline Tax?" *American Economic Review*, **95**:1276-1289.

Patriksson, Michael (2004) "Algorithms for Computing Traffic Equilibria." *Networks and Spatial Economics*, **4**:23-38.

Payne, H.J. (1971) "Models of Freeway Traffic and Control. In: G. A. Bekey (ed.), *Mathematical Models of Public Systems*, Simulation Council Proceedings, Vol. 1, pp. 51-61.

Payne, Harold J. (1984) "Discontinuity in Equilibrium Freeway Traffic Flow." *Transportation Research Record,* **971**:140-146.

Peltzman, Sam (1975) "The Effects of Automobile Safety Regulation." *Journal of Political Economy*, **83**:677-725.

Pels, Eric, and Piet Rietveld (2000) "Cost Functions in Transport." In: Hensher and Button (2000), pp. 321-333.

Pendyala, Ram M., and Ryuichi Kitamura (1997) "Weighting Methods for Attrition in Choice-Based Panels". in Golob, T. F., R. Kitamura, and L. Long (eds.), *Panels for Transportation Planning: Methods and Applications*. 233-257

Perry, James L., Timlynn Babitsky, and Hal Gregersen (1988) "Organizational Form and Performance in Urban Mass Transit." *Transport Reviews,* **8**:125-143.

Petitte, Ryan A. (2001) "Fare Variable Construction and Rail Transit Ridership Elasticities." *Transportation Research Record,* **753**:102-110.

Pickrell, Don H. (1983) "Sources of Rising Operating Deficits in Urban Bus Transit." *Transportation Research Record,* **915**:18-24.

Pickrell, Don H. (1989) *Urban Rail Transit Projects: Forecast versus Actual Ridership and Costs*. U.S. Department of Transportation, Transport Systems Center, Cambridge, Mass.

Pickrell, Don H. (1992) "A Desire Named Streetcar: Fantasy and Fact in Rail Transit Planning." *Journal of the American Planning Association,* **58**:158-176.

Pigou, Arthur C (1920) *The Economics of Welfare*. London: Macmillan.

Pipes, L.A. (1953) "An operational analysis of traffic dynamics" *Journal of Applied Physics* **24** 271-281.

Plaut, Pnina O. (1997) "Transportation-Communications Relationships in Industry," *Transportation Research Part A*, 31: 419-429.

Plaut, Pnina O. (2004) "Non-Commuters: The People Who Walk to Work or Work at Home." *Transportation,* **31**:229-255.

Polinsky, A. Mitchell (1972) "Probabilistic Compensation Criteria." *Quarterly Journal of Economics*, **86**:407-425.

Poole, Robert W., Jr. (1988) "Resolving Gridlock in Southern California." *Transportation Quarterly,* **42**:499-527.

Prashker, Joseph N (1979) "Direct Analysis of the Perceived Importance of Attributes of Reliability of Travel Modes in Urban Travel." *Transportation,* **8**:329-346.

Pratt, Richard H., Texas Transportation Institute, Cambridge Systematics, Parsons Brinckerhoff Quade & Douglas, SG Associates, and McCollom Management Consulting (2000) "Transit Pricing and Fares". in, *Traveler Response to Transportation System Changes: Interim Handbook*. Online. http://www4.nationalacademies.org/trb/crp.nsf/All+Projects/TCRP+B-12 (accessed 27 May 2005)

PRoGRESS (2004) *PRoGRESS Main Project Report*. PRoGRESS (Pricing ROad use for Greater Responsibility, Efficiency and Sustainability in citieS) Project 2000-CM.10390 (July), European Commission, DG TREN. http://www.progress-project.org/Progress/report.html (accessed 6 July 2006).

Proost, Stef and Kurt van Dender (1998) "Variabilization of Car Taxes and Externalities." In: Button and Verhoef (1998), pp. 136-149.

Proost, Stef and Kurt van Dender (2001) "Methodology and Structure of the Urban Model." In: Bruno de Borger and Stef Proost, *Reforming Transport Pricing in the European Union: A Modelling Approach* Edward Elgar, Cheltenham, UK, pp. 65-92.

Pucher, John (1984) "Allocating Federal Transit Subsidies: A Critical Analysis of Alternatives." *Transportation Research Record,* **967**:14-23.

Pucher, John (1988) "Urban Travel Behavior as the Outcome of Public Policy: The Example of Modal-Split in Western Europe and North America." *Journal of the American Planning Association,* **54**:509-520.

Pucher, John, and Anders Markstedt (1983) "Consequences of Public Ownership and Subsidies for Mass Transit: Evidence from Case Studies and Regression Analysis." *Transportation,* **11**:323-345.

Pucher, John, and John L. Renne (2003) "Socioeconomics of Urban Travel: Evidence form the 2001 NHTS." *Transportation Quarterly,* **57**:49-77.

Putman, Stephen H. (1983) *Integrated Urban Models: Policy Analysis of Transportation and Land Use*. London: Pion.

Quandt, Richard E., and William J. Baumol (1966) "The Demand for Abstract Transport Modes: Theory and Measurement." *Journal of Regional Science,* **6**:13-26.

Quinet, Emile (2004) "A Meta-Analysis of Western European External Costs Estimates." *Transportation Research Part D*, **9**:465-476.

Ramjerdi, Farideh, Harald Minken, and Knut Østmoe (2004) "Norwegian Urban Tolls." In: Santos (2006), pp. 237-249.

Ramsey, Frank P. (1927) "A Contribution to the Theory of Taxation." *Economic Journal,* **37**:47-61.

Ran, Bin and David Boyce (1996) *Modeling Dynamic Transportation Networks: An Intelligent Transportation System Oriented Approach* Second edition. Springer, Berlin.

Recker, Wilfred W., Thomas F. Golob, Chang-Wei Hsueh, and Paula Nohalty (1988) *An Analysis of the Characteristics and Congestion Impacts of Truck-Involved Freeway Accidents.* Final Report to the California Department of Transportation, No. RTA 13945-55D281., Institute of Transportation Studies, Univ. of California at Irvine,

Reilly, John M. (1977) "Transit Costs During Peak and Off-Peak Hours." *Transportation Research Record,* **625**:22-26.

Richards, Martin G. (2006). *Congestion Charging in London: The Policy and the Politics.* Basingstoke, Hampshire: Palgrave MacMillan.

Richards, Paul I. (1956) "Shock Waves on the Highway." *Operations Research* **4**:42-51.

Rimmer, Peter J. (1988) "Buses in Southeast Asian Cities: Privatisation without Deregulation". in Dodgson, J. S., and N. Topham. (eds.), *Bus Deregulation and Privatisation,*. 185-208, Aldershot, U.K.: Gower.

Rooney, S., and Roger F. Teal (1986) "Developing a Cost Model for Privately Contracted Commuter Bus Service." *Transportation Research Record,* **1051**:48-56.

Ross, Paul (1988) "Traffic Dynamics." *Transportation Research Part B,* **22**:421-435.

Rotemberg, Julio J. (1985) "The Efficiency of Equilibrium Traffic Flows." *Journal of Public Economics,* **26**:191-205.

Roth, Gabriel J. (1996) *Roads in a Market Economy* Avebury Technical, Aldershot.

Rothengatter, Werner (2000) "External Effects of Transport." In Jacob B. Polak and Arnold Heertje, eds., *Analytical Transport Economics: An International Perspective*, Edward Elgar, Cheltenham, UK, pp. 79-116.

Rust, John (1988) "Statistical Models of Discrete Choice Processes." *Transportation Research Part B,* **22**:125-158.

SACTRA (1994)

Safirova, Elena, Kenneth Gillingham, Ian Parry, Peter Nelson, Winston Harrington, and David Mason (2004) "Welfare and Distributional Effects of Road Pricing Schemes for Metropolitan Washington DC." In: Santos (2004), pp. 179-206.

Samuel, Peter (2005) "Technologies Will Work in Parallel." *World Highways*, 19 April. http://www.worldhighways.com/features/article.cfm?recordID=1748 (accessed 6 July 2006).

Santos, Georgina (ed.) (2004) *Road Pricing: Theory and Evidence*. In series *Research in Transportation Economics* **9**, Elsevier.

Santos, Georgina, Wai Wing Li, and Winston T.H. Koh (2004) "Transport Policies in Singapore." In: Santos (2004), pp. 209-235.

Santos, Georgina, David Newbery, and Laurent Rojey (2001) "Static Versus Demand-Sensitive Models and Estimation of Second-Best Cordon Tolls: An Exercise for Eight English Towns." *Transportation Research Record*, **1747**:44-50.

Savage, Ian (1988) "The Analysis of Bus Costs and Revenues by Time Period: I. Literature Review." *Transport Reviews,* **8**:283-299.

Savage, Ian (1989) "The Analysis of Bus Costs and Revenues by Time Period: II. Methodology Review." *Transport Reviews,* **9**:1-17.

Savage, Ian (1997) "Scale Economies in United States Rail Transit Systems." *Transportation Research Part A*, **31**:459-473.

Savage, Ian (2004) "Management Objectives and the Causes of Mass Transit Deficits." *Transportation Research Part A*, **38**, pp. 181-199.

Schafer, Andreas (2000) "Regularities in Travel Demand: An International Perspective." *Journal of Transportation Statistics,* **3**:1-31.

Schwanen, Tim, Martin Dijst, and Frans M. Dieleman (2004) "Policies for Urban Form and their Impact on Travel: The Netherlands Experience," *Urban Studies*, 41: 579-603.

Sheffi, Yosef (1985) *Urban Transportation Networks: Equilibrium Analysis with Mathematical Methods*, Prentice-Hall, Englewood Cliffs, New Jersey.

Sherret, Alistair (1975) *Immediate Travel Impacts of Transbay BART.* Report No. TM 15-3-75 for U.S. Department of Transportation., Burlingame, California: Peat, Marwick, Mitchell & Co. Distributed by National Technical Information Service, Springfield, Virginia.

Shiftan, Yoram, and John Suhrbier (2002) "The Analysis of Travel and Emission Impacts of Travel Demand Management Strategies Using Activity-Based Models," *Transportation*, 29: 145-168.

Shirley, Chad, and Clifford Winston (2004) "Firm Inventory Behavior and the Returns from Highway Infrastructure Investments." *Journal of Urban Economics*, **55**:398-415.

Shoup, Donald C. (1982) "Cashing Out Free Parking." *Transportation Quarterly,* **36**:351-364.

Shoup, Donald C. (1997) "Evaluating the Effects of Cashing Out Employer-Paid Parking: Eight Case Studies." *Transport Policy*, **4**:201-216.

Shoup, Donald C. (2005) *The High Cost of Free Parking*. Planners Press, American Planning Association, Chicago.

Shrank, David and Timothy Lomax (2002) *The 2002 Urban Mobility Report* Texas A&M University, Texas Transportation Institute.

Skabardonis, A., and R. Dowling (1996) "Improved Speed-Flow Relationships for Planning Application." *Transportation Research Record*, **1572**:18-23.

Skinner, Louise, with Kiran Bhat (1978) *Comparative Costs of Urban Transportation Systems.* Report of the U.S. Federal Highway Administration. Washington, D.C.: U.S. Government Printing Office.

Small, Kenneth A. (1982) "The Scheduling of Consumer Activities: Work Trips." *American Economic Review*, **72**:467-479.

Small, Kenneth A. (1983a) "Bus Priority and Congestion Pricing on Urban Expressways". in Keeler, T. E. (eds.), *Research in Transportation Economics, Vol. 1*. 27-74, JAI Press.

Small, Kenneth A. (1983b) "The Incidence of Congestion Tolls on Urban Highways." *Journal of Urban Economics,* **13**:90-111.

Small, Kenneth A. (1985) "Transportation and Urban Change". in Peterson., P. (eds.), *The New Urban Reality*. 197-223, Washington, D.C.: The Brookings Institution.

Small, Kenneth A. (1987) "A Discrete Choice Model for Ordered Alternatives." *Econometrica,* **55**:409-424.

Small, Kenneth A. (1992a) *Urban Transportation Economics*. Fundamentals of Pure and Applied Economics Vol. 51. Harwood, Chur.

Small, Kenneth A. (1992b) "Using the revenues from congestion pricing" *Transportation* **19** (4) 359-381.

Small, Kenneth A. (1994) "Approximate Generalized Extreme Value Models of Discrete Choice." *Journal of Econometrics,* **62**: 351-382.

Small, Kenneth A. (1999a) "Economies of Scale and Self-Financing Rules with Noncompetitive Factor Markets." *Journal of Public Economics*, **74**:431-450.

Small, Kenneth A. (1999b) "Project Evaluation." In: Gómez-Ibáñez, Tye, and Winston (1999), pp. 137-177.

Small, Kenneth A. (2004) "Road Pricing and Public Transport." In: Santos (2004), pp. 133-158.

Small, Kenneth A., and Xuehao Chu (2003) "Hypercongestion." *Journal of Transport Economics and Policy,* **37**:319-352.

Small, Kenneth A., and José A. Gómez-Ibáñez (1998) "Road Pricing for Congestion Management: The Transition from Theory to Policy". In: Button and Verhoef (1998), pp. 213-246.

Small, Kenneth A., and José A. Gómez-Ibáñez (1999) "Urban Transportation". In: P. Cheshire and E. S. Mills (eds.), *Handbook of Regional and Urban Economics, Vol. 3*. 1937-1999, Amsterdam: North-Holland.

Small, Kenneth A., and Terence C. Lam (2001) "The Value of Time and Reliability: Measurement from a Value Pricing Experiment." *Transportation Research Part E: Logistics and Transportation Review,* **37**:231-251.

Small, Kenneth A., Robert Noland, Xuehao Chu, and David Lewis (1999) *Valuation of Travel-Time Savings and Predictability in Congested Conditions for Highway User-Cost Estimation*, National Cooperative Highway Research Program Report 431, National Academy Press.

Small, Kenneth A., and Harvey S. Rosen (1981) "Applied Welfare Economics with Discrete Choice Models." *Econometrica,* **49**:105-130.

Small, Kenneth A., and Clifford Winston (1999) "The Demand for Transportation: Models and Applications." In: Gómez-Ibáñez, Tye, and Winston (1999), pp. 11-55.

Small, Kenneth A., Clifford Winston, and Carol A. Evans (1989) *Road Work: A New Highway Pricing and Investment Policy.*, Washington: Brookings Institution.

Small, Kenneth A., Clifford Winston, and Jia Yan (2005) "Uncovering the Distribution of Motorists' Preferences for Travel Time and Reliability." *Econometrica*, **73**: 1367-1382.

Small, Kenneth A., Clifford Winston, and Jia Yan (2006) ""Differentiated Road Pricing, Express Lanes, and Carpools: Exploiting Heterogeneous Preferences in Policy Design," *Brookings-Wharton Papers on Urban Affairs*, **7**: forthcoming.

Small, Kenneth A., and Jia Yan (2001) "The Value of 'Value Pricing' of Roads: Second-Best Pricing and Product Differentiation." *Journal of Urban Economics,* **49**:310-336.

Smeed, R.J. (1968) "Traffic Studies and Urban Congestion." *Journal of Transportation Economics and Policy,* **2**:33-70.

Smith, Edward (1973) "An Economic Comparison of Urban Railways and Express Bus Services." *Journal of Transport Economics and Policy,* **7**:20-299.

Smith, Michael J. (1979) "The Marginal Cost Pricing of a Transportation Network." *Transportation Research Part B*, **13**:237-242.

Smith, Tony E., Erik Anders Eriksson, and Per Olov Lindberg (1995) "Existence of Optimal Tolls under Conditions of Stochastic User-Equilibria." In: Börje Johansson and Lars-Göran Mattsson, *Road Pricing: Theory, Empirical Assessment and Policy*, Kluwer, Dordrecht, pp. 65-87.

Smith, W. Spencer, Fred L. Hall, and Frank O. Montgomery (1996) "Comparing Speed-Flow Relationships for Motorways with New Data from the M6." *Transportation Research Part A*, **30**:89-101.

Smith, Wilbur, and Associates (1965) *Parking in the City Center*. New Haven, Connecticut: Wilbur Smith and Associates.

Sorensen, Paul A., and Brian D. Taylor (2005) *Review and Synthesis of Road-Use Metering and Charging Systems*. Report Commissioned by the Committee for the Study of the Long-Term Viability of Fuel Taxes

for Transportation Finance, Transportation Research Board, Washington. http://www.trb.org/publications/news/university/SRFuelTaxRoad-MeterPaper.pdf (accessed 6 July 2006).

Southworth, Frank (2001) "On the Potential Impacts of Land Use Change Policies on Automobile Vehicle Miles of Travel." *Energy Policy*, 29: 1271-1283.

Spady, Richard H., and Ann F. Friedlaender (1978) "Hedonic Cost functions for the Regulated Trucking Industry." *Bell Journal of Economics,* **9**:154-179.

Ståle, Navrud (2002) *The State of the Art on Economic Valuation of Noise*. Report prepared for the European Commission, DG Environment, 14 April. http://europa.eu.int/comm/environment/noise/pdf/020414noisereport.pdf (accessed June 2006).

Starrs, Margaret M., and Christine Perrins (1989) "The Markets for Public Transport: The Poor and the Transport Disadvantaged." *Transport Reviews,* **9**:59-74.

Starrs, Margaret M., and David N.M. Starkie (1986) "An Integrated Road Pricing and Investment Model: A South Australian Application." *Australian Road Research,* **16**:1-9.

Steimetz, Seiji S.C. (2004) *New Methods for Modeling and Estimating the Social Costs of Motor Vehicle Use*. Ph.D. Dissertation, University of California at Irvine.

Steinberg, Richard, and Willard I. Zangwill (1983) "The Prevalence of Braess' Paradox." *Transportation Science,* **17**:301-318.

Steiner, Peter O (1957) "Peak Loads and Efficient Pricing." *Quarterly Journal of Economics,* **71**:585-610.

Storchmann, Karl (2004) "On the Depreciation of Automobiles: An International Comparison," *Transportation*, 31: 371-408.

Strotz, Robert H. (1965) "Urban Transportation Parables". in Margolis, J. (eds.), *The Public Economy of Urban Communities,*. 127-169, Washington, D.C.: Resources for the Future.

Sullivan, Edward, *et al.* (2000) *Continuation Study to Evaluate the Impacts of the SR91 Value-Priced Express Lanes: Final Report*, California Polytechnic State University, San Luis Obispo. http://ceenve.calpoly.edu/sullivan/sr91/sr91.htm, accessed August 8, 2005.

Sumalee, Agachai, Tony May, and Simon Shepherd (2005) "Comparison of Judgmental and Optimal Road Pricing Cordons." *Transport Policy*, **12**:384-390.

Supernak, Janusz, *et al.* (2001) *I-15 Congestion Pricing Project Monitoring and Evaluation Services: Phase II Year Three Overall Report*, San Diego State University Foundation. http://argo.sandag.org/fastrak/pdfs/yr3_overall.pdf, accessed August 8, 2005.

Tabuchi, Takatoshi (1993) "Bottleneck Congestion and Modal Split." *Journal of Urban Economics*, **34**:414-431.

Talley, Wayne K. (1988) "Competition in the Provision of US and UK Urban Bus Services: Privatisation vs Deregulation". in Dodgson, J. S., and N. Topham. (eds.), *Bus Deregulation and Privatisation,*. 171-184, Aldershot, U.K.: Gower.

Talley, Wayne K., and Eric E. Anderson (1986) "An Urban Transit Firm Providing Transit, Paratransit and Contracted-Out Services: A Cost Analysis." *Journal of Transport Economics and Policy,* **20**:353-368.

Tanner, J.C (1961) *Factors Affecting the Amount of Travel.*Road Research Technical Paper No. 51. London: Her Majesty's Stationery Office.

Taylor, D. Wayne (1989) "The Economic Effects of the Direct Regulation of the Taxicab Industry in Metropolitan Toronto." *Logistics and Transportation Review,* **25**:169-182.

Teal, Roger F., and Terry Nemer (1986) "Privatization of Urban Transit: The Los Angeles Jitney Experience." *Transportation,* **13**:5-22.

Thomas, Thomas C. (1968) "Value of Time for Commuting Motorists." *Highway Research Record,* **245**:17-35.

Thomson, J.M. (1977) *Great Cities and their Traffic*. London: Gollancz, Peregrine Edition.

Timmermans, Harry, Peter van der Waerden, Mario Alves, John Polak, Scott Ellis, Andrew S. Harvey, Shigeyuki Kurose, and Rianne Zandee (2002) "Time Allocation in Urban and Transport Settings: An International Inter-Urban Perspective," *Transport Policy*, 9: 79-93.

Toh, Rex S. (1992) "Experimental Measures to Curb Road Congestion in Singapore: Pricing and Quotas." *Logistics and Transportation Review,* **28**:289-317.

Train, Kenneth (1978) "A Validation Test of a Disaggregate Mode Choice Model." *Transportation Research,* **12**:167-174. DEL??

Train, Kenneth (1979) "A Comparison of the Predictive Ability of Mode Choice Models with Various Levels of Complexity." *Transportation Research Part A,* **13**:11-16.

Train, Kenneth (1980) "A Structured Logit Model of Auto Ownership and Mode Choice." *Review of Economic Studies,* **47**:357-370.

Train, Kenneth (1986) *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*, Cambridge, Mass.: MIT Press.

Train, Kenneth (2001) "A Comparison of Hierarchical Bayes and Maximum Simulated Likelihood for Mixed Logit," working paper, Dept. of Economics, University of California, Berkeley

Train, Kenneth (2003) *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge University Press.

Train, Kenneth, and Daniel McFadden (1978) "The Goods/Leisure Tradeoff and Disaggregate Work Trip Mode Choice Models." *Transportation Research,* **12**:349-353.

Transport Canada (1994) *Guide to Benefit-Cost Analysis in Transport Canada*. Ottawa (September). http://www.tc.gc.ca/finance/BCA/en/TOC_e.htm, accessed July 21, 2006.

Transport for London (2004) *Congestion Charging Impacts Monitoring: Second Annual Report*, London, April. http://www.tfl.gov.uk/tfl/cclondon/cc_monitoring-2nd-report.shtml, accessed April 13, 2005.

Transportation Research Board (1998) *National Automated Highway System Research Program: A Review.* Special Report 253, Committee for a Review of the National Automated Highway System Consortium Research Program, Transportation Research Board, National Research Council, Washington.

Transportation Research Board (2000) *Highway Capacity Manual 2000*. Transportation Research Board, National Research Council, Washington.

Traynor, Thomas L. (1994) "The Effects of Varying Safety Conditions on the External Costs of Driving," *Eastern Economic Journal*, **20**: 45-60.

Tretvik, Terje (2003) "Urban road pricing in Norway: public acceptability and travel behaviour". In Jens Schade and Bernhard Schlag (2003) *Acceptability of Transport Pricing Strategies* Elsevier / Pergamon, Amsterdam, pp. 77-92.

Tullock, Gordon, and Theft (1967) "The Welfare Cost of Tariffs, Monopolies." *Western Economic Journal,* **5**:224-232.

UK Department for Transport (2002) *COBA 11 User Manual, Part 5: Speed on Links*, Department for Transport, London.

UK Department for Transport (2004a) *Feasibility Study of Road Pricing in the UK*, Department for Transport, London, July. http://www.dft.gov.uk/stellent/groups/dft_roads/documents/page/dft_roads_029788-01.hcsp#P57_1651, accessed July 13, 2006.

UK Department for Transport (2004b) *Transport Statistics Bulletin: National Travel Survey: 2003 Final Results*, October. http://www.dft.gov.uk/stellent/groups/dft_transstats/documents/page/dft_transstats_031839.hcsp, accessed 4 April 2005.

UK Ministry of Transport (1964) *Road Pricing: The Economic and Technical Possibilities*, London: Her Majesty's Stationery Office.

UK National Statistics Online (2004) *Labour Force Survey (LFS) Historical Quarterly Supplement*. http://www.statistics.gov.uk/STATBASE/Expodata/Spreadsheets/D7938.xls, accessed Dec. 18, 2004.

UNITE (2003) *Pilot Account Results*, Deliverables 5, 8, 12. UNIfication of accounts and marginal costs for Transport Efficiency, European Commission – DG TREN, Firfth Framework Programme. http://www.its.leeds.ac.uk/projects/unite/deliverables, accessed July 14, 2006.

US BLS (2004) *National Compensation Survey: Occupational Wages in the United States, July 2003*, Bulletin 2568, (September). US Bureau of Labor Statistics, Washington. http://www.bls.gov/ncs/ocs/sp/ncbl0658.pdf, accessed 6 April 2005.

US Bureau of Public Roads (1964) *Traffic Assignment Manual.*, Washington, D.C.: U.S. Bureau of Public Roads.

US Census Bureau (various years) *Statistical Abstract of the United States*, Government Printing Office, Washington.

US CEA (2005) *Annual Report of the Council of Economic Advisers*. US Council of Economic Advisors. In: *Economic Report of the President*, US Government Printing Office, Washington. http://www.whitehouse.gov/cea/erpcover2005.pdf, accessed 21 April 2005.

US Department of Commerce (1998) "Fixed Reproducible Tangible Wealth in the United States: Revises Estimates for 1995-97 and Summary Estimates for 1925-97." *Survey of Current Business*, **78** (Sept.): 36-46. http://www.bea.gov/bea/pubs.htm, accessed 10 May 2005.

US Department of Transportation (1997) *The Value of Travel Time: Departmental Guidance for Conducting Economic Evaluations*. Washington.

US FHWA (1997) *Final Report on the Federal Highway Cost Allocation Study.* US Federal Highway Administration. US Government Printing Office, Washington. http://www.fhwa.dot.gov/policy/otps/costallocation.htm.

US FHWA (1998) *Highway Statistics 1997.* Washington, D.C.: US Government Printing Office. http://www.fhwa.dot.gov/policy/ohpi/hss/index.htm, accessed 16 May 2005.

US FHWA (2000) *Addendum to the 1997 Federal Highway Cost Allocation Study Final Report.* Washington, D.C.: US FHWA, May. http://www.fhwa.dot.gov/policy/hcas/addendum.htm, accessed 16 May 2005.

US FHWA (2002) *Highway Statistics 2001.* Washington, D.C.: US Government Printing Office. http://www.fhwa.dot.gov/policy/ohpi/hss/index.htm.

US FHWA (2004) *Highway Statistics 2003.* Washington, D.C.: U.S. Government Printing Office. http://www.fhwa.dot.gov/policy/ohpi/hss/index.htm.

US GAO (2001) *Mass Transit: Bus Rapid Transit Shows Promise*. US General Accountability Office, Washington.

US GAO (2005) *Highway and Transit Investments: Options for Improving Information on Projects' Benefits and Costs and Increasing Accountability for Results*. US General Accountability Office Report GAO-05-172, Washington, January.

US OMB (1992) *Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programs*, Circular No. A-94, Revised, Section 8. US Office of Management and Budget, Washington, October.

US OMB (2003) *Regulatory Analysis*, Circular No. A-4, Revised, US Office of Management and Budget, Washington, September.

Van den Bossche, M.A., C. Certan, Simme Veldman, Chris Nash, Daniel Johnson, Andrea Ricci, Riccardo Enei (2003) *Guidance on Adapting Marginal Cost Estimates*. UNITE (UNIfication of accounts and marginal costs for Transport Efficiency) Deliverable 15. Report for European Commission – DG TREN, Fifth Framework Programme, Brussels, April. http://www.its.leeds.ac.uk/projects/unite/downloads/, accessed 21 May 2005.

Van Dender, Kurt (2001) *Aspects of Congestion Pricing for Urban Transport*. Ph.D. Dissertation No. 149, Faculty of Economics, Katholieke Universiteit Leuven, Leuven, Belgium

Van Dender, Kurt (2003). "Transport taxes with multiple trip purposes," *Scandinavian Journal of Economics*, **105**, pp. 295-310.

Van Dender, Kurt, and Stef Proost (2004). "Optimal urban transport pricing in the presence of congestion, economies of density and costly public funds." Working paper, Department of Economics, University of California at Irvine.

Van Ommeren, Jos, Gerard J. Van den Berg, and Cees Gorter (2000) "Estimating the Marginal Willingness to Pay for Commuting." *Journal of Regional Science*, **40**:541-563.

Van Wissen, L.J.G., and H.J. Meurs (1989) "The Dutch Mobility Panel: Experiences and Evaluation." *Transportation,* **16**:99-119.

Vaughan, Rodney (1987) *Urban Spatial Traffic Patterns.*, London: Pion.

Verges, Joaquin (1989) "Medicion de la Diferencia en Cuanto a Costes Sociales por Congestion entre el Transporte Individual y Transporte Publico Urbano: Con una Aplicacion para el Area Urbana de Barcelona." *Investigaciones Economicas,* **13**:183-205.

Verhoef, Erik T. (2001) "An integrated dynamic model of road traffic congestion based on simple car-following theory: exploring hypercongestion" *Journal of Urban Economics* **49** 505-542.

Verhoef, Erik T. (2002a) "Second-best congestion pricing in general static transportation networks with elastic demands" *Regional Science and Urban Economics* **32** (3) 281-310.

Verhoef, Erik T. (2002b) "Second-best congestion pricing in general networks: heuristic algorithms for finding second-best optimal toll levels and toll points" *Transportation Research Part B*, **36**:707-729.

Verhoef, Erik T. (2003) "Inside the queue: hypercongestion and road pricing in a continuous time – continuous place model of traffic congestion" *Journal of Urban Economics* **54** 531-565.

Verhoef, Erik T. (2005) "Transport infrastructure charging and capacity choice". Paper presented to the 135th Round Table of the OECD/ECMT Transport Research Centre on *Transport Infrastructure Investment Charges and Capacity Choice*, ECMT, Paris.

Verhoef, Erik T. (2006) "Second-best road pricing through highway franchising" Paper presented at the Conference in Honor of Kenneth A. Small, 3-4 February 2006, UCI Irvine.

Verhoef, Erik T., Richard H.M. Emmerink, Peter Nijkamp and Piet Rietveld (1996) "Information provision, flat- and fine congestion tolling and the efficiency of road usage" *Regional Science and Urban Economics* **26** 505-529.

Verhoef, Erik T., Peter Nijkamp and Piet Rietveld (1995a) "Second-best regulation of road transport externalities" *Journal of Transport Economics and Policy* **29** (2) 147-167.

Verhoef, Erik T., Peter Nijkamp and Piet Rietveld (1995b) "The economics of regulatory parking policies: the (im-)possibilities of parking policies in traffic regulation" *Transportation Research* **29A** (2) 141-156.

Verhoef, Erik T., Peter Nijkamp and Piet Rietveld (1996) "Second-best congestion pricing: the case of an untolled alternative" *Journal of Urban Economics* **40** (3) 279-302.

Verhoef, Erik T., Peter Nijkamp and Piet Rietveld (1997) "The social feasibility of road pricing: a case study for the Randstad area" *Journal of Transport Economics and Policy* **31** (3) 255-267.

Verhoef, Erik T. and Jan Rouwendal (2004) "A Behavioural Model of Traffic Congestion: Endogenizing Speed Choice, Traffic Safety and Time Losses" *Journal of Urban Economics*, **56**:408-434.

Verhoef, Erik T., and Kenneth A. Small (2004) "Product Differentiation on Roads: Constrained Congestion Pricing with Heterogeneous Users." *Journal of Transport Economics and Policy,* **38**:127-156.

Vickrey, William S. (1963) "Pricing in Urban and Suburban Transport." *American Economic Review, Papers and Proceedings,* **53**:452-465.

Vickrey, William S. (1968) "Automobile Accidents, Tort Law, Externalities, and Insurance: An Economist's Critique." *Law and Contemporary Problems,* **33**:464-487.

Vickrey, William S. (1969) "Congestion Theory and Transport Investment." *American Economic Review, Papers and Proceedings,* **59**:251-260.

Vickrey, William S. (1973) "Pricing, Metering, and Efficiently Using Urban Transportation Facilities." *Highway Research Record,* **476**:36-48.

Viscusi, V. Kip, and Joseph E. Aldy (2003) "The Value of a Statistical Life: A Critical Review of Market Estimates Throughout the World," *Journal of Risk and Uncertainty*, **27**: 5-76.

Viton, Philip A. (1980b@) "On the Economics of Rapid-Transit Operations." *Transportation Research Part A,* **14**:247-253.

Viton, Philip A. (1980c@) "The Possibility of Profitable Bus Service." *Journal of Transport Economics and Policy,* **14**:295-314.

Viton, Philip A. (1981a) "On Competition and Product Differentiation in Urban Transportation: The San Francisco Bay Area." *Bell Journal of Economics,* **12**:362-379.

Viton, Philip A. (1981b) "A Translog Cost Function for Urban Bus Transit." *Journal of Industrial Economics,* **24**:287-304.

Viton, Philip A. (1982) "Privately-Provided Urban Transport Services: Entry Deterrence and Welfare." *Journal of Transport Economics and Policy,* **16**:85-94.

Viton, Philip A. (1983) "Pareto-Optimal Urban Transportation Equilibria". in Keeler, T. E. (eds.), *Research in Transportation Economics, Vol. 1*. 75-101, Greenwich, Connecticut: JAI Press.

Viton, Philip A. (1986) "Quasi-Optimal Pricing and the Structure of Urban Transportation." *Transportation Research Part A,* **20**:295-305.

Viton, Philip A. (1995) "Private roads" *Journal of Urban Economics* **37** (3) 260-289.

Voith, Richard (1997) "Fares, Service Levels, and Demographics: What Determines Commuter Rail ridership in the Long run?" *Journal of Urban Economics,* **41**:176-197.

Vovsha, Peter (1997) "The Cross-Nested Logit Model: Application to Mode Choice in the Tel-Aviv Metropolitan Area.," paper presented to the Transportation Research Board, No. 97-0387.

Wachs, Martin (1986) "Technique vs. Advocacy in Forecasting: A Study of Rail Rapid Transit." *Urban Resources,* **4**:23-30.

Wachs, Martin (1990) "Regulating Traffic by Controlling Land Use: The Southern California Experience." *Transportation,* **16**:241-256.

Walters, A.A. (1961) "The Theory and Measurement of Private and Social Cost of Highway Congestion." *Econometrica,* **29**:676-699.

Walters, A.A. (1968) *The Economics of Road User Charges*. World Bank Staff Occasional Papers No. 5. Baltimore: Johns Hopkins, International Bank for Reconstruction and Development.

Walters, A.A. (1987a) "Congestion". in, *The New Palgrave: A Dictionary of Economics*.**1:** 570-573, New York: Macmillan.

Walters, A.A. (1987b) "Ownership and Efficiency in Urban Buses". in Hanke, S. H. (eds.), *Prospects for Privatization*. 83-92, New York: Academy of Political Science.

Wardman, Mark (1998) "The Value of Travel Time: A Review of British Evidence." *Journal of Transport Economics and Policy*, **32**:285-316.

Wardman, Mark (2001) "A Review of British Evidence on Time and Service Quality Valuations." *Transportation Research Part E: Logistics and Transportation Review*, **37**:107-128.

Wardman, Mark (2004) "Public Transport Values of Time." Transport Policy, **11**:363-377.

Wardrop, John G. (1952) "Some Theoretical Aspects of Road Traffic Research." *Proceedings of the Institute of Civil Engineers*, **1**:325-378.

Washington, Simon P., Matthew G. Karlaftis, and Fred L. Mannering (2003) *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall, Boca Raton, Florida.

Waters, William G., II (1996) "Values of Travel Time Savings in Road Transport Project Evaluation." In: Hensher, David, Jenny King, and Tae Hoon Oum (eds.), *Proceedings of 7th World Conference on Transport Research*, Vol. 3: 213-223.

Webber, Melvin M. (1976) "The BART Experience: What Have We Learned?" *The Public Interest,* **45**:79-108.

Weisbrod, Glen, Donald Vary, and George Treyz (2001) *Economic Implications of Congestion*. National Cooperative Highway Research Program Report 463 Washington, D.C.: National Academy Press.

Weitzman, Martin L. (2001) "Gamma Discounting." *American Economic Review*, **91**:260-271.

West, Sarah E. (2004) "Distributional Effects of Alternative Vehicle Pollution Control Policies." *Journal of Public Economics*, **88**:735-757.

Wheaton, William C. (1978) "Price-induced distortions in urban highway investment" *Bell Journal of Economics* **9** 622–632.

White, Michelle J. (2004) "The 'Arms Race' on American Roads: The Effect of Sport Utility Vehicles and Pickup Trucks on Traffic Safety," *Journal of Law and Economics*, **47**: 333-355.

White, Peter R. (1990) "Bus Deregulation: A Welfare Balance Sheet." *Journal of Transport Economics and Policy,* **24**:311-332.

Wigan, Marcus, Nigel Rockliffe, Thorolf Thoresen, and Dimitris Tsolakis (2000) "Valuing Long-Haul and Metropolitan Freight Travel Time and Reliability." *Journal of Transportation and Statistics*, **3**:83-89.

Williams, Huw C.W.L. (1977) "On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit." *Environment and Planning Part A,* **9**:285-344.

Williams, Huw C.W.L., and Laurence A.R. Moore (1990) "Appraisal of Highway Investments under Fixed and Variable Demand." *Journal of Transport Economics and Policy,* **24**:61-81.

Willig, Robert (1976) "Consumer's Surplus Without Apology." *American Economic Review,* **66**:589-597.

Willson, Richard W. (1992) "Estimating the Travel and Parking Demand Effects of Employer-Paid Parking." *Regional Science and Urban Economics*, **22**:133-145.

Willson, Richard W. (1995) "Suburban Parking Requirements: A Tacit Policy for Automobile Use and Sprawl." *Journal of the American Planning Association*, **61**:29-42.

Willson, Richard W., and Donald C. Shoup (1990) "Parking Subsidies and Travel Choices: Assessing the Evidence." *Transportation,* **17**:141-157.

Wilson, John D. (1983) "Optimal road capacity in the presence of unpriced congestion" *Journal of Urban Economics* **13** 337-357.

Winston, Clifford, Vikram Maheshri, and Fred Mannering (2006) "An Exploration of the Offset Hypothesis Using Disaggregate Data: The Case of Airbags and Antilock Brakes," *Journal of Risk and Uncertainty*, **32**:83-99.

Winston, Clifford, and Chad Shirley (1998) *Alternate Route: Toward Efficient Urban Transportation.* Washington: Brookings Institution Press.

World Bank (1999) *Asian Toll Road Development Program: Review of Recent Toll Road Experience in Selected Countries and Preliminary Tool Kit for Toll Road Development*, World Bank, Washington.

World Bank (2006) *Toll roads and concessions* , World Bank, Washington. http://www.worldbank.org/transport/ roads/toll_rds.htm; accessed 18 April 2006.

Wunsch, Pierre (1996) "Cost and Productivity of Major Urban Transit Systems in Europe." *Journal of Transport Economics and Policy,* **30**:171-186.

Yang, Hai (1999a) "System Optimum, Stochastic User Equilibrium, and Optimal Link Tolls." *Transportation Science*, **33**:354-360.

Yang, Hai (1999b) "Evaluating the Benefits of a Combined Route Guidance and Road Pricing System in a Traffic Network with Recurrent Congestion." *Transportation*, **20**:299-321.

Yang, Hai and Hai-Jun Huang (1999) "Carpooling and congestion pricing in a multilane highway with high-occupancy-vehicle lanes" *Transportation Research* **33A** (2) 139-155.

Yang, Hai and Qiang Meng (2002) "A note on 'Highway pricing and capacity choice in a road network under a build-operate-transfer scheme'." *Transportation Research Part A*, **36**:659-663.

Yang, Hai, Qiang Meng, and Der-Horng Lee (2004) "Trial-and-Error Implementation of Marginal-Cost Pricing on Networks in the Absence of Demand Functions." *Transportation Research Part B*, **38**:477-493.

Young, William, Russell G. Thompson, and Michael A.P. Taylor (1991) "A Review of Urban Car Parking Models." *Transport Reviews,* **11**:63-84.

Zellner, Arnold, and Peter E. Rossi (1984) "Bayesian Analysis of Dichotomous Quantal Response Models." *Journal of Econometrics,* **25**:365-393.

Zerbe, Richard O., Jr (1983) "Seattle Taxis: Deregulation Hits a Pothole." *Regulation,* Nov./Dec.:43-48.

Zhang, H. Michael (1999) "A Mathematical Model of Traffic Hysteresis." *Transportation Research Part B*, **33**:1-24.

Zhang, H. Michael (2001) "New Perspectives on Continuum Traffic Flow Models." *Networks and Spatial Economics*, **1**:9-33.

Zhang, Xiaoning, and Hai Yang (2004) "The Optimal Cordon-Based Network Congestion Pricing Problem." *Transportation Research Part B*, **38**:517-537.

SELECTED SYMBOLS AND ABBREVIATIONS

$A(t)$	Cumulative queue-entries
a	Exponent in Cobb-Douglas utility function of goods and leisure; Parameter in BPR congestion function
ac	Average cost
$B(t)$	Cumulative queue-exits
B	Total benefits (willingness to pay including actual payments)
b	Parameter (exponent) in BPR congestion function
C	Total cost
C_B	Cost to bus agency
C_g	Congestion-related part of total cost, including capacity cost
C_W	Cost of waiting time to bus passengers
c	Short-run average variable cost (SRAVC)
c_0	SRAVC on an uncongested road
c_{00}	SRAVC on an uncongested road exclusive of the value of free-flow travel time
c_g	Congestion-related part of SRAVC
$\bar{c}_g$	Time-averaged equilibrium SRAVC in dynamic models
c_b,c_p	Cost per vehicle-mile of base and peak bus service, respectively
c_S,c_T	Parts of congestion-related SRAVC attributable to schedule delay and travel time (i.e., queuing delay), respectively
c_1-c_5	Other cost parameters (definition varies with context)
cdf	Cumulative distribution function
D	Density of vehicle traffic
D^i	Alternative-specific dummy variable for alternative i
D_j	Jam density
D_m	Density consistent with maximum flow
d	Inverse demand function (in highway cost analysis); Average passenger trip length (in transit cost analysis)
d_{jn}	Choice variable (=1 if decision-maker n chooses alternative j)
e, exp	Exponential function
$F(t)$	Cumulative desired queue-exits
G	Function for generating GEV models of discrete choice
GEV	Generalized extreme value
H	Number of time periods per weekday

h	Time period
I_r	Inclusive value for alternative group r
iid	Identically and independently distributed
J	Number of dependent variables (aggregate models) or alternatives (disaggregate models)
J_a	Delay parameter in Akçelik's travel time function
J_r	Number of alternatives in alternative group r
K	Capital cost (present value)
K_0	Fixed part of capital cost
K_1	Coefficient of capacity in variable part of capital cost
L	Leisure; Length of a road
log	Natural logarithm
MB_K	Marginal benefit of capacity expansion
mc	Marginal cost
mec	Marginal external cost
$mecc$	Marginal external congestion cost
N	Number of people; Number of vehicles in queue
n	Bus capacity
P	Choice probability; Inflow period (in duration-dependent congestion functions)
PV	Present value
p	Inclusive price of travel to user
Q	Total number of travelers ($=qV_a$ or qV_d when V_a or V_d are constant)
q	Output (in general cost analysis); Bus passenger volume (in bus-cost analysis); Duration of period of desired queue entries = $t_{p'}$-t_p (in highway queuing analysis)
$\boldsymbol{q}$	Vector of durations of different time periods
R	Revenue
r	Interest rate
S	Speed
S_D	Schedule delay ($=t'$-t_d)
S_f	Free-flow speed
S_m	Speed consistent with maximum flow
s	Returns to scale: ratio of average to marginal cost

s_K Returns to scale in producing highway capacity
s_n Vector of socioeconomic or other characteristics of decision-maker n

T Time (or vector of times) spent in activities; Travel Time (usually in-vehicle) if used as scalar without sub- or super-scripts
T^0 Out-of-vehicle travel time
T_0 Free-flow travel time
T_D Queuing-delay portion of travel time
t_i^k k-th travel-time component on i-th mode
T_w Time spent at work (in value-of-time analysis)
t Time of day (in queueing analysis)
t' Time of queue exit
t_d Desired time of queue exit
t_p , $t_{p'}$ Beginning and end of desired period of queue-exits
t_q , $t_{q'}$ Beginning and end of actual period of queue-exits
t^* Time for which actual and desired queue-exit times coincide
$\tilde{t}$ Time of maximum queuing delay

U Utility function

V Indirect utility function (in travel-demand analysis);
Volume (flow) of vehicle traffic (vehicles per hour) in highway-cost analysis;
Frequency of transit service (vehicles per hour) in transit-cost analysis
$\boldsymbol{V}$ Vector of vehicle flows in different time periods
V_a Rate (volume) of actual entries to queue (vehicles per hour)
V_b Rate (volume) of actual exits from bottleneck (vehicles per hour)
V_d Rate (volume) of desired exits from bottleneck (vehicles per hour)
V_h Volume (flow) of vehicle traffic during time period h
V_i Rate (volume) of actual entries to road
V_o Rate (volume) of actual exits from road
V_K Capacity of highway or bottleneck (vehicles per hour)
v_T Value of time (usually in-vehicle time)

W Welfare measure; Aggregate waiting time
w Wage rate (in travel-demand analysis);
Input-price vector (in general cost-function analysis)

X Generalized consumption good (numeraire)
x Consumption vector (in value-of-time analysis);
Input vector (in general cost-function analysis)
x_n Fixed input in short-run cost function

Y Unearned income

y Generalized argument for function G generating GEV models

z Independent variables for travel-demand models

α Value of travel time (in highway cost analysis)
α_i Alternative-specific constant for alternative i in discrete-choice indirect utility function
β Parameter vector in discrete-choice indirect utility function (in travel-demand analysis);
Shadow price of schedule-delay early (in highway cost analysis)
γ Shadow price of schedule-delay late
γ_i Coefficient of an independent variable interacted with an alternative-specific constant for alternative i in discrete-choice utility function
δ $\beta\gamma/(\beta+\gamma)$ (in highway cost analysis)
ε_i Stochastic term for alternative i in discrete-choice indirect utility function
θ Parameter vector in general cost-function analysis
Λ Lagrangian function
λ Lagrange multiplier; Marginal utility of income
μ Scale parameter for probability density function (in discrete-choice analysis);
Lagrange multiplier (in value-of-time analysis)
Π Profit
π Annual rate of inflation
ρ Parameter of GEV functions (in discrete-choice analysis);
Capital recovery factor divided by number of weekdays per year (converts capital cost to an ongoing daily cost) (in cost analysis)
σ Standard deviation (in statistical estimates)
Parameter of GEV function (=1-ρ) (in travel-demand analysis)
Fraction of travelers exiting queue prior to desired time (in highway-cost analysis)
τ Congestion fee ($ per vehicle-mile or per passage; depending on context)
φ Lagrange multiplier
ψ Lagrange multiplier