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, 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 for incorporating 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:

$$F(q,x;\theta) = 0$$  \hspace{1cm} (3.1)

where \( q \) and \( x \) are vectors of outputs and inputs, respectively, and \( \theta \) 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 \( \theta \) 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 \( \bar{C} \) when the distinction between short-run and long-run cost functions 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 \( x_n \) becomes another argument of the resulting cost function, which we may write as \( C(q, w; \theta, x_n) \). By definition,

\[
\tilde{C}(q, w; \theta) = \min_{x_n} C(q, w; \theta, x_n).
\] (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 (\( w_n x_n \) in the case of fixed factor price); 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 any one output, \( q_i \), as \( mc = \partial C / \partial q_i \). One can show from (3.2) that the long-run marginal cost is equal to the short-run marginal cost with \( x_n \) 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 that capital stock is varied.

**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,

\[
s = \frac{ac}{mc} = \frac{C}{q \cdot (\partial C / \partial q)}. \] (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

$$\frac{sC}{\partial C / \partial q} = C / s.$$  (3.4)

Hence revenue will exactly cover total cost if there are neutral scale economies and $s=1$; 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,

¹ When the input price vector $w$ is constant, scale economies and diseconomies are equivalent to *increasing returns to scale* and *decreasing returns to scale* in the production function (3.1), respectively, which 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 example, if urban land becomes scarcer and thus more expensive as a highway network is expanded, this will produce long-run diseconomies of scale for highway travel even if the production function shows constant returns (Small, 1999a). Perhaps this is an underlying cause of the political factors shown by Altshuler and Luberoff (2003) to be behind the extraordinary increase in costs of large urban infrastructure projects during the last quarter century.
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.

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 included 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.\(^2\) 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 firm’s 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 a single conventional output but include descriptors of the operating environment, such as traffic speeds or population density, as parameters affecting the relationship between that output and cost. It is especially important to retain the distinction between expanding the density of output, for example by adding more vehicles or attracting more patrons on a given route, 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 such scale

\(^2\)Fielding (1987, pp. 60-63), in developing performance measures for transit operators, makes the same distinction. He refers to the firm’s efficiency at producing intermediate outputs as “cost efficiency,” and to its ability to translate them into final outputs as “service effectiveness.” The two combined determine the firm’s “cost effectiveness,” i.e., its ability to produce final outputs at low cost.
economies; if it does they would be called \textit{economies of size}. Many transportation industries have been found to have economies of density but not of size (Braeutigam, 1999).

\textbf{Methods of Measurement}

There are at least three general approaches to empirically measuring cost functions. The \textit{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 \textit{engineering} approach builds a production function from technical descriptions of the production process and adds information about input prices in order to determine costs. The \textit{statistical} approach infers how cost varies with levels of output and other variables by observing cost in many different situations: for example, a single firm over many time periods, or a cross section of many firms.\footnote{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 approaches, and any given study may make use of more than one.

\textbf{External, Social, and Full Costs}

Recent years have seen increased attention to costs that are borne not by the providing agency or the users of a given service, but by other parties. Examples abound: air pollution, noise, groundwater contamination, and wildlife disruption, to name a few. Such costs are called “external” because they fall on parties who are not part of the specific 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.

\textit{Social or full costs} are the total costs including any external costs. We define the \textit{marginal social cost} \((\text{msc})\) 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; see also Section 3.3.

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 third parties. 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 et al. (1997). Section 3.4 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.
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 $VM$. This rather strong assumption leads to the cost function:

$$C = c_1 \cdot RM + c_2 \cdot PV + c_3 \cdot VH + c_4 \cdot VM.$$  \hspace{1cm} (3.5)

This approach is one of “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 inextricably joint costs. However $PV$ could be considered a fixed input in the short run, so the second term serves as a way of distinguishing between short-run and long-run costs; to fully specify the latter would require an additional model describing how the transit provider adjusts $PV$ when other outputs are changed.

We can also use the outputs in this model to distinguish between 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.

Table 3.1 compares the results, adjusted to 2003 U.S. prices, of two studies which compute equation (3.5) by examining the cost accounts of transit agencies. The studies include both operating and capital costs and 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. It is apparent that they make somewhat different assumptions about how certain costs vary, especially whether rapid rail operating costs vary with time or distance. 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 over heavy rail 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, a positive coefficient of $PV$ indicates higher peak costs because a change in peak service, but not in off-peak service, is likely to call for a change in $PV$. Second, to the extent that congestion slows vehicles during peak periods, peak operations require a larger number of vehicle-hours $VH$ to provide the same frequency and geographical coverage of service. However, one would expect driver costs per vehicle-hour to differ between peak and offpeak service as well, 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.) It is also common to assume that drivers can be assigned to base (all-day) service at one level of unit cost, but that a premium must be paid in order to add additional peak-only service. 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. Empirical estimates range from 1.1 to 2.5. Of course, other coefficients could also be distinguished by time period but the rationale for doing so is less compelling.

More sophisticated analyses simulate the decisions of work schedulers faced with various combinations of work rules and overtime pay rates. This method was pioneered in a study of

---


5 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.
Bradford, UK, and more fully developed for Adelaide, Australia, as described by Savage (1988, 1989). Results of Chomitz and Lave (1984) using this technique suggest an approximate value of $c_p/c_b = 1.5$ for typical conditions. 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) provide a useful history and typology of accounting cost practices. They also describe in detail 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 that are fixed in the short run; this result is attributed by the authors to “overstaffing” of the transit system, the extent 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.)

As noted by Savage (1988, 1989) and Abbas and Abd-Allah (1999), use of private firms to provide transit service greatly increased the motivation to develop cost models that can finely distinguish among times of day or routes. For this reason, many such models were developed.

---

6 It also suggests a typical value for $c_3/c_2$ 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. Our computations are as follows. We assume that each bus in peak service operates 4 hours of peak service, and each bus in base service operates 10 hours of offpeak service. Then $(10B + 4P)c_3 = (10B)c_b + (4P)c_p$, where $B$ and $P$ are the number of buses in service during base and peak periods, respectively. Each of Chomitz and Lave’s cases provides observations on $P/B$ (first column of their Table 1) and on $(10B + 4P)/c_0$ (fourth column), where $c_0$ is the hourly pay rate. The two cases mentioned above thus enable us to solve this equation for the ratio $(c_p/c_b)$. The ratio $(c_3/c_2)$ is then calculated for $P/B = 2.0$, a value that is intermediate in the cases they analyze.
during the 1970s and 1980s, when the United States was experimenting with contracted service and most British urban bus transit was deregulated. As a result of these changes, private firms needed to analyze the many possibilities for introducing or expanding service, while public agencies needed a sound basis to plan subsidies or franchises.

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 use of such 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 3.5.1, 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 $s=2.04$) to strong diseconomies for some large ones ($s=0.30$ for New York).\footnote{From Viton (1980b, Table 2, p. 251). We have transformed his definition of economies of density ($ED$) to ours according to $s=1/(1-ED)$.} 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.\footnote{From Viton (1980b, Table 3, p. 252), updated using the transportation component of the consumer price index.} Viton then uses the fitted equation to predict the cost of operating the BART system (which is not part of the sample) for 1975-76; he finds that BART’s actual costs are 69 percent above prediction, suggesting that BART may be an unusually high-cost system.

Viton (1981b) turns to bus providers. He 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.\footnote{From Viton (1981b, Table V, p. 300), transforming his $SCE$ to our $s$ by $s=1/(1-SCE)$.} Viton also 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 approximately 0.89) for large firms.\footnote{From Button and O’Donnell (1985, Fig. 1, p. 75), using the same definitional transformation as with Viton.}

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 economies of scale 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:

$$LC^j = [c_{ia}^j \cdot TM^j + c_{ib}^j \cdot Stations^j + c_3 \cdot VH^j + (c_{4a}^j + c_{4b}^j \cdot n^j) \cdot VM^j] \cdot (w/\overline{w})$$

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, $LC = \sum_j S^j LC^j$ (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). In order to avoid heteroscedasticity (differing variances across observations), equation (3.6) is divided by $VM^j$ before aggregation and estimation.

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 franks) 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) determine 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

---

11 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 addition term $c_{ia}^j \cdot VM^j$, not multiplied by wage, is added to the equation to represent capital cost, with $c_{ia}^j$ estimated.
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 \( c_{ab} \) 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 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.

A few studies have also considered final outputs as the arguments for cost functions. Berechman (1983) and Berechman and Giuliano (1984) use time-series data to estimate translog long-run cost functions for bus service in Israel and the eastern San Francisco Bay Area, respectively. These studies find economies of scale using measures of final output, but diseconomies using a measure of intermediate output.\(^\text{12}\) This difference indicates that increased passenger demand is accommodated with a less than proportional increase in vehicle-miles. We

\(^{12}\) There is a danger in such studies: final output may not be exogenous because usage depends on fare, which could in turn depend on costs (Nelson 1970). However, subsidy policies intervene between costs and fares so the problem is probably not too serious.
investigate the relationship between final and intermediate outputs more closely in the next subsection.

Berechman (1993) and De Borger and Kerstens (2000) review many other statistical studies, reaching conclusions similar to the results just noted. We summarize them as follows. First, for bus providers, statistical evidence suggests that intermediate outputs such as vehicle-miles are produced with a mildly U-shaped relationship between average cost and output, so that small firms exhibit modest scale economies and large firms modest diseconomies. 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 seems to be a bias toward operating with a larger than optimal capital stock, possibly due to incentives built into capital subsidy programs.

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.\textsuperscript{13} 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

\textsuperscript{13}See Mohring (1976), pp. 145-146.
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 $\alpha^W$. 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:

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

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

$$q \leq nV.$$

Letting $\lambda$ be the Lagrangian multiplier of the constraint, the first-order condition is:

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

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

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

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

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

$$C_w^* = \alpha^W \cdot W^* = \sqrt{\frac{\alpha^W \cdot c_p}{2}} \cdot \sqrt{q}.$$

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 the optimal fare 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, $C_w/q$. 

3-17
The intuition here is simple: buses are not full when \( \lambda = 0 \), so it costs nothing to take another passenger. The aggregate subsidy is equal to the total cost of waiting time, \( C_{\text{W}}^* \).

If \( \lambda > 0 \), indicating the capacity constraint is binding, the solution is \( V^* = q/n, W^* = q/(2V^*) = n/2, C_{\text{B}}^* = c_p q/n, \) and \( C_{\text{W}}^* = \alpha W^* n/2 \). Over the range of output for which this solution holds, the total cost function is linear in output and has fixed cost \( C_{\text{W}}^* \). It thus again exhibits density economies \( s = 1 + (C_{\text{B}}^* / C_{\text{W}}^*) \), which are greater, the greater are waiting-time costs compared to operating costs.\(^{14}\)

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.

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.\(^{15}\) 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 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

---

\(^{14}\) Derivation: \( s = \frac{ac}{mc} = [\frac{(C_{\text{B}}^* + C_{\text{W}}^*) / q}{c_p / n}] = \frac{(C_{\text{B}}^* + C_{\text{W}}^*)}{C_{\text{B}}^*} \). Note however that \( s \leq 2 \) in an optimum, because if \( C_{\text{W}}^* \) were to rise above \( C_{\text{B}}^* \), a marginal increase in frequency \( V \) would reduce waiting costs by more (namely by \( C_{\text{W}}^* / V \)) than it would raise the cost to the bus agency (namely by \( c_p = C_{\text{B}}^* / V \)), so an optimizing agency would prevent this inequality from occurring.

\(^{15}\) See Jansson (1980) and Nash (1988).
smaller than the actual size in typical situations in Sweden and Britain. There is some evidence from the partial bus deregulation in Britain that small firms using small buses do, in fact, find a niche when allowed to do so.

In reality, of course, there is no precise maximum number of people who can ride a bus. When all seats are full, people stand and crowd, their boarding and alighting times increase, and people wishing to board may be deferred to a later bus. Turvey (1975) shows how each of these forms of crowding imposes costs or disutility on passengers, thereby adding to marginal social cost. Kraus (1991) provides a more formal analysis.

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 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 precommit 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 et al. (1955). 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:

$$V \equiv D \cdot S,$$  \hspace{1cm} (3.7)
which is consistent with its units of measure: vehicles/hour \equiv (\text{vehicles/mile}) \cdot (\text{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.\(^{16}\)

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 relation \(S(V)\) 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 the upper branch as congested or normally congested and the lower branch as hypercongested.\(^{17}\) 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 \(S_f\). But zero flow also occurs 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 \(c\).

---

\(^{16}\) 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.

\(^{17}\) 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 \(V \to 0\) on the upper branch. We use the terms “congested” and “normally congested” interchangeably, depending on whether an explicit distinction with “hypercongested” is needed.
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
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 from the congested to the hypercongested regime or vice versa (Hall et al., 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, Branston (1976) and Hall and Hall (1990) show that the observed speed-flow relationship can depend crucially on the presence of a nearby bottleneck.

Despite these difficulties, many empirical estimates are reported in the literature, using various functional forms and data sets. Detailed overviews can be found in May (1990) and Hall (2002). Some illustrative examples are described here.

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

\[ S = S_m \cdot [1 - (D / D)] \]

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:\(^18\)

\[ V = 2490. - 0.523 \cdot (S - 35.34)^2 \]

where \(V\) is in vehicles per hour per lane. The two solutions to this quadratic equation represent congestion and hypercongestion. Inman (1978) obtains:\(^19\)

\(^18\) This is their preferred equation (20), p. 350, converted to units of vehicles per lane per hour by setting \(V=5q\).

\(^19\) 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 \(N=V/500, G=S/10, \text{ and } X=36.488\). The Inman curve is undefined for \(S < 7.2 \text{ mi./hr.}\), 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.
\[ V^{(2.95)} = 3.351 \cdot 10^9 - 231.4 \cdot (S - 7.2)^{(4.06)} \]

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 tendency, obvious from the raw data, for expressway speed in the congested region to remain nearly constant throughout most of the range of volume-capacity ratios.  

FIGURE 3.3: Original Fig. 3.4

Two more recent “official” speed-flow relations, which do account for this phenomenon, are the COBA11 and Highway Capacity Manual (HCM2000) formulations in the UK and US, respectively.  

Figure 3.4 shows that both portray normal congestion, not hypercongestion, and do so with two segments joined at a “breakpoint” \( V_B \). HCM2000 joins a completely flat segment with a segment defined by a power function, such that the derivative is continuous. COBA11, by contrast, joins two linear segments but with a discontinuous slope; such a discontinuity could arise, for example, from speed limits. The parameters depend on highway geometry (hilliness, bendiness), 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: New (see end of document)

---

20 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.

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.

Keeler and Small (1977) use observations on three expressways in the San Francisco Bay Area to estimate functions relating speed and volume-capacity ratio $V/V_k$, each averaged over a long stretch of highway. (Capacity $V_k$ at each point is calculated using an earlier version of the Highway Capacity Manual, and so does not necessarily equal the maximum flow as in the instantaneous speed-flow curve.) For each expressway, Keeler and Small estimate a quadratic relationship like Boardman and Lave’s, exhibiting a hypercongested portion. The resulting equation for the Eastshore Freeway is:22

$$V = 0.8603 - 0.001923 \cdot (S - 45.68)^2.$$  

In contrast to the instantaneous speed-flow curves, these 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,

---

22This is Keeler and Small’s equation (12), p. 11, rewritten to make transparent the maximum $V/V_k$ ratio and the maximum speed.
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 simple 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 volume approaches capacity.

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:\(^{23}\)

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

\[^{23}\text{Their equation (8), with values } n=1.58 \text{ and } v_{ma}=60/1.95 \text{ as indicated just above their equation (6) on p. 6.}\]
\[ T = T_f \cdot \left[ 1 + a \cdot \left( \frac{V}{V_K} \right)^b \right] \]  

(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.\(^{24}\) With values \( a=0.2 \) (freeways) or 0.05 (arterials), and \( b=10 \), it is known as the “updated BPR function,” derived by Skabardonis and Dowling (1996) to approximate the speed-flow functions in the 1994 Highway Capacity Manual, a predecessor to HCM2000 discussed above.

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 \( V_K \):\(^{25}\)

\[
T = \begin{cases} 
T_f & \text{if } V \leq V_K \\
T_f + \frac{1}{2} \cdot P \cdot \left( \frac{V}{V_K} - 1 \right) & \text{if } V > V_K.
\end{cases} \]  

(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.\(^{26}\) Similarly, Small (1983a, pp. 32-33) finds that (3.10) approximates surprising well the pattern of travel times during the afternoon peak period on an eleven-mile stretch of freeway in the San Francisco Bay area.

\(^{24}\) 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.

\(^{25}\) 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.

\(^{26}\) We used nonlinear least squares with the nine data points representing simulation runs 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.
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 intended to represent the behavior of expected waiting times in stochastic queuing models with random arrivals. The function is:

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

(3.11)

which has (3.10) as a special case when $J_a=0$.27 Akçelik’s function seems to produce reasonable equilibrium network speeds and flows when used in network models (Dowling, Singh and Cheng, 1998; Singh, 1999).

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, $V_K=2000$ 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 and BPR-U curves represents BPR and the updated BPR functions, respectively.

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.

27 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 $\{600,25,1.6\}$.
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 bottleneck queuing, an approach developed by Small and Chu (2003).

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, calculations by Newell (1971, p. 125) suggest that stochastic queuing accounts for only a small fraction of travel delays in most circumstances, so we can safely restrict attention to deterministic queuing, which is much simpler. For these reasons, deterministic queuing is used extensively later in this text. We therefore begin with its basic equations (May and Keller, 1966).

Let $V_a(t)$ be the (non-stochastic) 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. 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

3-29
“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:

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

(3.12)

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

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

(3.13)

where the dot denotes a time derivative (or, if $V_a(t)$ is discontinuous, a one-sided derivative).

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 $t \leq t_q$ 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:

$$N(t_q') \equiv \int_{t_q}^{t_q'} (V_a(t) - V_K) dt = 0.$$  

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:

$$T_D(t) = \frac{N(t)}{V_K} = \int_{t_q}^{t} \left( \frac{V_a(t')}{V_K} - 1 \right) dt', \quad t_q \leq t \leq t_q'.$$

(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 a function 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.
Consider the special case of a fixed peak period covering the time interval \([t_p, t_p']\), with incoming traffic constant at \(V_a\) during this interval 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 \(t_p \leq t \leq t_p'\):

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

The average travel delay is:

\[
\bar{T}_D(t) = \frac{1}{t_p' - t_p} \int_{t_p}^{t_p'} T_D(t) \, dt
\]

\[
= \begin{cases} 
0 & \text{if } V_a \leq V_K \\
\frac{1}{2} (t - t_p) \left(\frac{V_a}{V_K} - 1\right) & \text{if } V_a > V_K.
\end{cases}
\]

Adding a free-flow travel time \(T_f\) per unit distance for the journey yields equation (3.10) for \(P = (t_p' - t_p)\).

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:\(^{28}\)

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

(3.16)

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 enter another. For example, consider an upstream steady state with $V_u = S_u D_u$, and a downstream steady state with $V_d = S_d D_d$. 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 $D_u (S_u - S_w)$. Similarly, downstream traffic leaves the boundary at relative speed $S_d - S_w$, and hence exits the wave at rate $D_d (S_d - S_w)$. Conservation of vehicles requires these rates to be equal, implying:

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

(3.17)

It can be shown that a shock wave can never travel faster than the traffic that carries it provided the function $V(D)$ is concave and crosses the origin.\(^{29}\)

---

\(^{28}\) Equation (3.16) can be verified by considering a discrete example. Let $\Delta N$ 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 $\Delta t$ is small enough, the flow rates at the two locations can be treated as constant over time; letting $\Delta V$ 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 $-\Delta V$: that is, $\Delta N = -\Delta t \Delta V$. If $\Delta x$ is small enough, density at the two locations can be treated as equal; letting $\Delta D$ denote the change in density over time, we see that the number of vehicles in this space of length $\Delta x$ changes in proportion to the change in density: $\Delta N = \Delta x \Delta D$. Because vehicles are conserved, these two expressions for $\Delta N$ 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.

\(^{29}\) After rotating the $V(D)$ curve of Figure 3.1c by $90^\circ$, $S_w$ between two states can be found geometrically as the slope of the straight line connecting these states. The traffic speeds in a state is given by the slope of the ray through that state (and the origin). Under the stated assumptions, $S_w$ is then always smaller than $S_u$ and $S_d$.  

3-32
Finding a solution for the LWR-model is tedious when traffic inflow varies continuously over time (see Newell, 1988). For this reason, 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 has infinite density $D_d$.

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:

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

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 schedules 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 $Q(t)$ 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 show that this total trip time is the sum of two components: the time that would be calculated from the Henderson model, representing normally congested flow in a steady state with traffic entering at rate $Q(t)$, plus the queuing delay calculated from the deterministic queuing model of (3.14).

*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):

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

(3.18)

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

Under stationary traffic conditions, car-following models imply a relationship between density (the inverse of vehicle spacing) and speed 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,

$$
S = a \cdot \log(C_i/D),
$$

(3.19)

where $D=a \cdot \log[C_0]$ 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:

$$\dot{x}_{n+1}(t) = S(D) \quad \text{and} \quad D = \left[ x_n(t) - x_{n+1}(t) \right]^{-1} .$$

(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 is found 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. 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. At this point there could be a considerable payoff from incorporating some of the more realistic engineering models, already well developed, into economic analysis. Examples will be encountered in the sequel as we discuss dynamic congestion tolls and hypercongestion.

### 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 (whereas certain other types of capital, such as vehicles and parking facilities, may be variable). Section 3.5 considers long-run cost functions.

To simplify the exposition, we assume that everyone has an identical value of time (denoted in this chapter by $\alpha$). The average cost on a defined length of road, $c$, consists of monetary expenses $c_{00}$ (like fuel consumption and maintenance), the cost of free-flow travel time $\alpha T_f$, the cost of travel time delays $\alpha (T - T_f)$, 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 $c_g$:

$$c = c_0 + c_g = [c_{00} + \alpha \cdot T_f] + [\alpha \cdot (T - T_f) + c_S].$$  \hspace{1cm} (3.21)

---

30 Other interesting phenomena include multiple-class traffic (e.g. 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). Zhang (1999) considers traffic hysteresis, i.e., path-dependency of observed speed-density combinations.
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 “full cost” or “generalized cost” because it indicates the monetary and other resources required for an individual to take a trip. The related concept of “full price” or “generalized price”, denoted \( p \), adds to \( c \) any tolls and taxes, if levied.

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, which justifies 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

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

So long as we stay on the normally congestion 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 – in particular their decision of whether and when to use the road. 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 this 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 about equilibria $y$ and $z$: how should one interpret a situation where an increase in traffic inflow produces faster travel and a resultant lower average cost?

**FIGURE 3.7 New (see end of the document)**

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.31 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 analyses is that the perturbations considered each assume that simultaneous changes occur in the flow rates into and along the entire road,

---

31 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.
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.\(^{32}\) 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 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; hypercongestion exists within the entrance queue (Verhoef, 2003), consistent with the terminology of Figure 3.4 and the assumption of Mun (2002), but the hypercongested speed and flow rate are irrelevant to total trip time.

FIGURE 3.8 New (see end of the document)

Thus the true supply curve for stationary-state traffic, $c_{stat}(V)$, is everywhere rising and can intersect a downward-sloping demand curve only once. Although it has two distinct regimes, one of them vertical, it may sometimes be approximated by $c(V)$ as derived from the power function (3.9), which we derived as a time-averaged congestion relationship:

\(^{32}\) 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, $V_o - V_f$ and $D_o - D_f$ 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.
\[ c(V) = c_{00} + \alpha T_j \cdot \left[ 1 + a \cdot \left( \frac{V}{V_k} \right)^{\gamma} \right]. \] (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. 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 (or number of travelers on the road), or the total number of trips *per se* (*i.e.*, 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 *per se* 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 (also note that a cost function cannot be defined over the number of trips *per se*). And traffic density is an aggregate measure of road space occupied, but not of the number of trips started and completed per unit of time. An aggregate demand function defined over density would therefore assume that the good demanded is the occupation of road space, not the completion of trips. Traffic flow therefore appears the appropriate output measure for stationary-state analyses, while the number of trips *per se* becomes useful only in time-averaged or dynamic models that specify (or determine endogenously) the duration of the period under consideration.

### 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 \( c_{\text{stat}}(V) \) is shown for comparison. Both of the time-averaged functions become steeper when \( P \) rises; 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 may allow for too much since, for high enough values of $P$, it will cross the $C_{\text{stat}}$ curve and exhibit arbitrarily high travel times even when $V<V_K$.

FIGURE 3.9 New (see end of document)

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 a well-defined period and zero flow at other times. Second, $P$ is set exogenously, but in reality will vary with traffic conditions and policies. Third, the assumed exogenous inflow 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)

\[33\text{ Note that a naïve but understandable choice of } P, \text{ defined by the instants that queuing begins and ends, would imply a time-averaged inflow } V \text{ equal to capacity } V_K. \text{ This choice would produce travel delay equal to zero according to the piecewise-linear function and } 4\left((PJ/V)\right)^{1/3} \text{ according to Akçelik function.}\]
with $\theta=0$. Schedule delay costs are piecewise-linear with per-minute costs of early and late arrival denoted by $\beta$ and $\gamma$, respectively. Then the travel cost for an individual departing from home at time $t$ is:

$$c(t) = c_{00} + \alpha \cdot T(t) + c_s(t); \quad c_s(t) = \begin{cases} \beta \cdot (t_d - t - T(t)) & \text{if } t + T(t) \leq t_d \\ \gamma \cdot (t + T(t) - t_d) & \text{if } t + T(t) > t_d, \end{cases}$$

(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.

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 subtract the constant $c_{00} + \alpha T_f$; hence we define average congestion cost:

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

Without loss of generality, we set free-flow travel time $T_f$ to zero so that $T(t)$ is equal to the travel delay $T_D(t)$ defined in (3.15) (except we replace the exogenous time $t_p$ when congestion begins with an endogenous time, denoted $t_q$).

Because desired schedules are defined in terms of arrival time at work, it is convenient to focus on the time $t'$ when a traveler departs from 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 $\max_{t'} V_d(t) \leq V_K$, 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 a the interval \([t_p, t_p']\) and zero outside that interval. Hence there are a total of \( Q \equiv V_d q \) 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).\(^{34}\)

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): \( \alpha T_D(t) + \beta [t_d - t - T_D(t)] \). This requires that \( T_D(t) \) changes 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, \( \alpha T_D(t) + \gamma [t + T_D(t) - t_d] \) must be minimized, so \( T_D(t) \) must change at rate \( -\gamma (\alpha + \gamma) \).\(^{35}\) 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 those implied by equation (3.15), we see that vehicles must be entering the queue at rates

\[
V^\text{early}_a = V_k \cdot \frac{\alpha}{\alpha - \beta}; \quad V^\text{late}_a = V_k \cdot \frac{\alpha}{\alpha + \gamma}
\]

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, \( \overline{t} \) is the entry time for the commuter incurring maximum queuing delay \( T_{Dm} \), and this commuter’s exit time is:

\[
t^* \equiv \overline{t} + T_D(\overline{t}) \equiv \overline{t} + T_{Dm}.
\]  

---

\(^{34}\) The ADL analysis treats only the special case in which the desired schedule \( t_d \) is identical for all commuters, i.e. \( Q \equiv V_d q \) is fixed but \( q = 0 \) and \( V_d \) is infinite. We achieve more realism at a fairly small cost in complexity by retaining Vickrey’s original assumption of a uniform distribution of \( t_d \) with nonzero \( q \) and finite \( V_d \).

\(^{35}\) 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 \( \beta / \alpha \) and \( \gamma / \alpha \). He gets values much smaller than the direct behavioral estimates of Small (1982), possibly reflecting the limited realism of the model.
Commuters with \( t_d > t^* \) enter the queue before (or possibly at) \( \bar{t} \), and those with \( t_d \geq t^* \) enter after (or possibly at) \( \bar{t} \). (Due to linearity in the cost function, each commuter is in fact indifferent among departure times \([t_q, \bar{t}]\) or else among departure times \([\bar{t}, t_q]\); 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.)

It remains to determine \( \bar{t} \). 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:

\[
t_q' - t_q = Q / V_K = q \cdot V_d / V_K > q.
\] (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.\(^{36}\) In order to solve for the entire equilibrium configuration, let

\[
\sigma = \frac{t^* - t_p}{q} = \frac{t^* - t_p}{t_q' - t_q}
\] (3.28)

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

\[
\sigma \cdot Q = V_a^{early} \cdot (\bar{t} - t_q).
\] (3.29)

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

\[
(1 - \sigma) \cdot Q = V_a^{late} \cdot (t_q' - \bar{t}).
\] (3.30)

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

\(^{36}\) The dependence of this interval on \( V_K \) verifies Downs’s (1962) argument that expanding capacity narrows the peak period.
where
\[ \delta = \beta \gamma / (\beta + \gamma) = \beta \sigma. \] (3.35)

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

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 \( \beta(t_p - t_q) \) for the earliest traveler to zero (at \( \tilde{t} \)), then rises linearly to a maximum of \( \gamma(t_q' - t_p') \); computing these maxima from (3.31) - (3.33), we find they are both equal to \( (\delta Q/V_K)(1 - V_K/V_d) \). Note that their sum, \( c_q(t) \) need not be constant in equilibrium because each consumer has a different desired schedule \( t_d \). (This differs from the Fargier and ADL models.)

FIGURE 3.11: redraw old Figure 3.10

From the piecewise-linear cost patterns just described, we see easily that the time-averaged costs are just half the maximum costs. Thus time-averaged travel-delay cost per traveler is:

\[ \bar{c}_T = \frac{1}{2} \cdot \frac{\delta \cdot Q}{V_K} = \frac{\delta \cdot q}{2} \cdot \frac{V_d}{V_K}. \] (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:
\[
\bar{c}_S = \frac{1}{2} \cdot \frac{\delta \cdot Q}{V_K} \cdot \left(1 - \frac{V_K}{V_d}\right) \equiv \frac{\delta \cdot q}{2} \cdot \left(\frac{V_d}{V_K} - 1\right).
\]  

(3.37)

This does depend on the distribution of desired exit times; for a given number of travelers \(Q = V_d \cdot q\), 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 \(V_d = \infty\) (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 \(V_d \leq V_K\) with its lack of queuing, we can write the time-averaged congestion cost as:

\[
\bar{c}_g(V_d, q; V_K) = \begin{cases} 
0 & \text{if } V_d \leq V_K \\
\frac{\delta \cdot Q}{V_K} \cdot \left(1 - \frac{V_K}{2V_d}\right) \equiv \frac{\delta \cdot q}{2} \cdot \left(\frac{V_d}{V_K} - 1\right) & \text{otherwise.}
\end{cases}
\]  

(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 \(V_a(t)\) 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 \(\frac{1}{2} \cdot \delta q\). 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 \(V_d\).

Because \(\delta\) in (3.35) does not depend on \(\alpha\), these costs have the remarkable feature of not depending on value of travel time, \(\alpha\), so long as \(\alpha\) remains greater than \(\beta\) 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 — as can be determined by a careful examination of equations...
(3.31)-(3.34). 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:

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

(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 involve non-linearities, the analytics become more cumbersome and typically no closed-form analytical solutions can be obtained.

**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 flexibility in schedules 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. 37

To analyze such problems, we define a network structure consisting of $M$ origin-destination pairs or “markets” (denoted $m=1,\ldots,M$), $R$ routes (denoted $r=1,\ldots,R$), and $L$ directed links (denoted $l=1,\ldots,L$). “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 $\delta_{rm}$ to denote whether route $r$ serves market $m$ (in which case $\delta_{rm}=1$), and another set $\delta_{lr}$ 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

37 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.”
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.

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 \( \delta_{rm} = 1 \):

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

where

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

are the link and market flows, respectively, and where \( \rho \) 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 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 entails integrals of average cost functions \( c_\ell(\cdot) \) 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), which occurs when adding a new link to a congested network causes equilibrium travel times to increase. This can happen because the objective function whose minimization yields the UE conditions, described above, is something 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! These paradoxes are described in an intuitive manner by Arnott and Small (1994). Both paradoxes occur because user prices are not set optimally. For example, as Chapter 4 will discuss, optimal pricing for a network as described in equation (3.40) involves link-based tolls that bridge the gap between average and marginal cost. Including these in the user equilibrium conditions (3.40) makes these correspond to the necessary first-order conditions for minimizing (minus) social surplus. The Braess paradox can then no longer occur.

An important extension of the UE concept is the *stochastic user equilibrium* (SUE), which introduces randomness into route choice (Daganzo and Sheffi, 1977). The equilibrium condition then becomes that no traveler can reduce expected travel cost by unilaterally changing route. Typically, SUE models apply logit or probit choice models in which systematic utility is determined by generalized cost and randomness can represent travel-time uncertainty or idiosyncratic preferences. Given many users, the law of large numbers causes aggregate flows to be deterministic. A similar approach can be used for multi-modal networks.

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. This 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 inflows. These interactions are asymmetric because travel time does not depend on future inflows. An example of such a model is METROPOLIS (De Palma, Marchal, and Nesterov, 1997).

A second example allows 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 spaces are 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 a common property resource, absent pricing, is overused. With either parking or congestion, some of these resources may be more valuable than others — for example the most popular routes in the case of congestion, the best-

---

38 This is accomplished by interpreting them as the solution to a mathematical optimization problem known as *variational inequality* (Dafermos, 1980).
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, the more desirable locations being overused the most.

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. Anderson and de Palma allow for this factor by postulating that vehicles parking at location $x$ slow down all vehicle wishing to park

---

39 Shoup (2005, pp. 303-304) describes quantitative estimates of this effect in the UK and Calcutta.
closer to the CBD than $x$ (since they must pass location $x$ to do so). This sets up another 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 second-best optimal provision of parking spaces, given inefficiently low parking prices, is greater than first-best optimal provision, even though these spaces remove some capacity from the streets that could otherwise be used to improve traffic flow.

Thus models of parking exhibit properties depending sensitively on fine details of the situation. From a research perspective, this is a young field open to significant new findings.

As we will see below, 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.

### 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. 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).

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. About $0.017 of this is for fuel taxes, which we subtract here but add later as a

---

40 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), and Newbery (2005).

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

42 Davis and Diegel (2004), Table 10.11.
private payment toward fixed roadway costs, which is how they are normally viewed in the US.\footnote{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: \textit{ibid.}, Table VM-1.} 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.

\textit{(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 \textit{capital recovery factor} to the price of a new car.\footnote{The capital recovery factor $\rho$ 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 $K$, $\rho K$ 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 $\frac{r}{(1-\delta)^T}$, where $\delta = 1/(1+r)$. For a clear derivation, see DeNeufville and Stafford (1971), ch. 8. If interest is compounded continuously, the formula becomes $\frac{r}{(1-e^{-rt})}$.} 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.\footnote{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: \textit{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.} 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
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. 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. 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

---

46 Mean hourly wages for metropolitan areas in 2003 were $18.29, from US BLS (2004), Table 1-1.
47 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.\footnote{We update using hourly earnings in private industries, which grew by 9.64 percent between 2000 and 2003 (US CEA, 2005, Table B-47).} 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;\footnote{Transport for London (2004), pp. 2-3.} 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.\footnote{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.} We also omit the additional costs to freight that may occur due to paid

\[ V/V_K = 0.84 \]
drivers’ higher values of time and the inventory costs of delays to expensive vehicles and their loads.  

(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 \( \beta \) and \( \gamma \) 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. 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. 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

\[ \frac{7.0}{(22.5/3)} = 0.93. \]

---

51 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.

52 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.

53 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.\(^5\) 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

---

\(^5\) 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 Consumer 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.\footnote{Blincoe \textit{et al.} (2003, Appendix A) use estimates of willingness to pay for “quality adjusted life years” to estimate willingness to pay for both fatalities and injuries. They do not use these to recompute their estimated total accident costs.}

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 required by risk alone 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, depending on specification. 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.\footnote{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 using the average of the growth in the CPI-U component of the Consumer Price Index (17.3\% for 1996-2003, 12.9\% for 1998-2003, 6.9\% 2000-2003) and the growth in hourly earnings (27.6\% for 1996-2003, 18.1\% and 9.6\%, respectively) (US CEA, 2005), Tables B-60, B-47). This is equivalent to assuming that VSL grows at half the rate of real wages (see text below) and using the price index to restate it in 2003 dollars.} 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 (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.57

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.58 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 simple question of whether adding more vehicles to the road raises accident costs for existing

57 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.

58 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.
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.\textsuperscript{59} The resulting effect on average accident costs is ambiguous, and for this reason it is sometimes assumed to be zero.\textsuperscript{60}

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.\textsuperscript{61} 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 \textit{mc} minus private \textit{ac}) is 53 percent of private \textit{ac}—a serious impediment to efficient resource allocation. This percentage can

\textsuperscript{59} See Traynor (1994) or Fridstrøm \textit{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 \textasciitilde 0.5; zero for car crashes between intersections; and positive for multi-car crashes at intersections, with elasticity \textasciitilde 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.

\textsuperscript{60} The results of Lindberg (2001, Table 6) for urban cars, given his value of \( \theta = 0.5 \) 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 twp-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.

\textsuperscript{61} 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.\textsuperscript{62}

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.\textsuperscript{63} 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.\textsuperscript{64} 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.

If one accepts that the accident externality is large, one cannot necessarily conclude that the best way to deal with it is through a Pigouvian user charge, which will be discussed in

\textsuperscript{62} We compute this as $28/(76-\theta)$, with $\theta=0.5$ 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 $\theta r(a+b)$.

\textsuperscript{63} An even larger ratio is estimated by Miller, Spicer, and Levy (1999), based on older data.

\textsuperscript{64} 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).
Chapter 4. Reciprocal externalities are also addressed in tort law, criminal law, and insurance regulation. See, for example, Boyer and Dionne (1987) and White (2004). Whether such arrangements provide efficient incentives to avoid causing accidents, as idealized Pigouvian charges would, depends to a large extent on the degree to which these 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.

(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.65 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.

(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

65 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).
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.66

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 oldest respondents (aged 70 or above).” By contrast, Viscusi and Aldy (pp. 50-53) conclude that VSL does decline with age.67

Considerations of discounting and possible age dependence

---

66 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.

67 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
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.68

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).69 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.70 An identical number can be derived from a somewhat older study by McCubbin and Delucchi (1999).71

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 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).

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

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

70 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.

71 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).
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. 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. 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.

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 optimal policy toward environmental externalities from

---


73 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.
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—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. In the latter case, the resulting function is not smooth but rather is the lower envelope of several distinct short-run curves. Haikalis and Campbell (1962) empirically estimate such a cost function assuming that just six highway designs are possible; they find a scalloped-shape long-run average cost curve that generally rises as highway design moves through several arterial types of increasing capacity, then falls steeply after the cost-minimizing design shifts to grade separation. In other words, in broad terms they find scale diseconomies at low volumes and scale economies at high volumes. Choice among discrete highway designs can also be made using cost-benefit analysis, discussed in Chapter 5.

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 from a short-run cost function and 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.\(^{74}\) Second, we assume 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, all of which may be quite important in practice (Larsen, 1993). Third, we assume that initial capital cost is linear in capacity:

\[
K(V_K) = K_0 + K_1 \cdot V_K
\]  

(3.41)

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

---

\(^{74}\) 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).
We restrict the average short-run variable cost over a given time period of duration $q$ to be a function of volume-capacity ratio $V/V_K$ during that period, as it is in every case considered in Section 3.4:

$$c(V) = c_0 + c_g \left(\frac{V}{V_K}\right)$$  \hspace{1cm} (3.42)

where $c_g(\cdot)$ describes congestion-related average user cost and $c_g(0) = 0$. Depending on the model, $V$ may represent static flow, time-averaged flow, or desired arrival rate; thus short-run total variable cost over that period is $c \cdot V \cdot q$. Several such periods 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 $c_0$.

Adding (3.41) and (3.42), short-run total cost may be written as:

$$C(V, q; V_K) = \rho K_0 + c_0 Q + C_g(V, q; V_K)$$

where $Q=V \cdot q$ is total vehicle-trips. The first two terms are unaffected by capacity, so we can and will ignore them in deriving investment criteria. We thus focus on the tradeoff between congestion and capacity in the third, congestion-related, term, which includes those parts of both short-run variable cost and capital cost that vary with capacity. 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), the resulting congestion-related part of the long-run total cost function (cost per day) is:

$$\tilde{C}_g(V, q) = \min_{V_K} \left\{ \sum_h q_h \cdot V_h \cdot c_g (V_h / V_K) + \rho K_1 V_K \right\}$$  \hspace{1cm} (3.43)

where $V$ is the vector of flows $V_h$ and where $q$ is the vector of durations $q_h$. The first-order condition for minimization in (3.43) leads to the following investment rule:

$$\rho K_1 = \sum_h q_h \cdot V_h \cdot \frac{\partial c_g (\cdot)}{\partial V_K} = \sum_h q_h \cdot \left(\frac{V_h}{V_K}\right)^2 \cdot c_g^\prime (\cdot),$$  \hspace{1cm} (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 \( V \) and substituting into the minimand in (3.43) gives the long-run cost function as well as short-run costs 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 across 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, whose assumption of constant returns makes their results independent of demand, 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.

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 \( Q = V \cdot q \); 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 \( \alpha T_f \cdot a \cdot (V/V_K)^b \) from the BPR cost function (3.23). Applying (3.44) and solving for \( V_K \), we obtain:

\(^{75}\)This applies their assumed relationship, \( 60/S = 1 + (V/V_K)^{2.4} \) (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.
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:

$$
\tilde{C}_g(V, q) = c_{hyp} \cdot V^{1/(b+1)} \text{ with: } c_{hyp} = (\rho K_1)^{b/(b+1)} \cdot (\alpha T_f a b)^{1/(b+1)} \cdot (1 + b^{-1}).
$$

We see that the congestion-related part of the long-run cost function exhibits no economies or diseconomies of scale 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 $V q^{1/2}$.

**Dynamic Congestion Model with Endogenous Scheduling**

Let us now turn to dynamic bottleneck congestion. Because capacity expansion will not affect the nature of the individuals' scheduling decisions, $\tilde{c}_g$ as given by (3.38) remains the relevant average congestion cost over all users. The daily congestion-related total cost is therefore again in the form (3.43) with just one time period, but now with $V_d$ replacing $V_h$, with $q$ replacing $q_h$ (and with equilibrium peak duration $t_q$–$t_q'$ of course exceeding $q$ if queuing occurs), and with $c_g$ replaced by $\tilde{c}_g$ from (3.38):

$$
\tilde{c}_g(V_d / V_K) = \begin{cases} 
0 & \text{if } V_d \leq V_K \\
\delta \cdot q \cdot \left( \frac{V_d}{V_K} - \frac{1}{2} \right) & \text{otherwise.}
\end{cases}
$$

There are two possible solution regimes. If $\rho K_1$ 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 $\rho K_1$ 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:
\[ V_K^* = \left( \frac{\delta}{\rho K_1} \right)^{1/2} \cdot V_d \cdot q. \]

In this regime, optimal capacity is proportional to \( Q = V_d \cdot q \); the proportionality constant is greater if the composite scheduling-cost parameter \( \delta \) is large or if the capacity-expansion cost \( \rho K_1 \) 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:

\[ \tilde{C}_g(V_d \cdot q) = V_d \cdot q \cdot \left( c_{bot} - \frac{\delta}{2} \cdot q \right) \quad \text{with:} \quad c_{bot} = 2\left(\rho K_1 \delta \right)^{1/2}. \]

The condition for transition between the two regimes is unenlightening. In both regimes, total congestion-related costs show scale economies with respect to duration \( q \): average long-run congestion-related cost \( \tilde{C}_g/(V_d \cdot q) \) 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

---

\(^77\) In the limiting special case where \( q \to 0 \) while keeping \( V_d q = Q \) finite, only the second regime applies, and its cost becomes \( C_g(Q) = c_{bot} Q \), showing no scale economies or diseconomies in \( Q \).
other ways besides physical road expansion to increase road capacity. Two main types of information technology are discussed here: automated highway systems and advanced travelers 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 26 to 85 percent on urban highways according to TRB (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 manual traffic (TRB, 1998). 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, there is a risk of merely shifting congestion from highways to city streets, leaving costs unchanged. 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, so may not improve overall network performance much. Fourth, legal considerations and driver resistance to relinquishing control may be barriers to implementation. Other considerations include costs, 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. A major purpose is to reduce the effects of unexpected
incidents on travel times, thereby improving both expected travel time and reliability. Emmerink and Nijkamp (1999) provide an overview.

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), for example, show that information provision may be welfare-reducing in certain networks.

Figure 3.12, based on Verhoef et al. (1996), shows why. 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 \( mcs \) 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 \( c_s \).

Let us define expected social surplus as expected total benefits minus expected total cost. Its change due to perfect information is measured by first determining in both states the change in area under the demand function, then subtracting the change in total cost (which is the area under the marginal cost function). Averaging these 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).

\(^{78}\) A sufficient condition is that the intercept of the average cost function be no lower in state 1 than in state 0.
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 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 and degree of urbanization, which affect such factors as the number and types of structures (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. It is common to classify terrain as flat, rolling, or mountainous. Urbanization may be accounted for by classifying locations into categories such as suburban, central city, and central business district; or by including explicit parameters such as urban density in the cost function.
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; i.e., as the inverse of the elasticity of capital cost with respect to capacity (or with respect to width as a proxy for 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 intersections, especially when they require complex signals or overpasses (Kraus, 1981), or from a rising 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 of 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 $a$. They estimate scale economies of $1/a=1.03$ (standard error 0.39), which may be taken as weak evidence for no scale economies or diseconomies. Land costs are estimated as fractions of construction costs, the fractions ranging from 26.7 to

---

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.

80 Keeler, and Small, 1977, Table 1, using their loglinear specification. As their discussion on p. 7 makes clear, the estimates of $a_0$ reported in their paper should have minus signs.
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 of 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. Using British data on costs and design standards (U.K. Department of the Environment, 1973), he estimates overall scale economies at 1.19 for a circular urban area of 10-mile radius containing a specified highway network.82 Kraus’s analysis indicates that intersections are an important component of urban highway cost, and the impact of one highway’s width on the cost of other highways crossing is a significant source of scale diseconomies for the network, 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. If we follow the cost allocation to vehicle classes in US FHWA (1997, Table V-21), 51 percent of these costs are attributable to automobiles, which comes to $0.051 per automobile vehicle-mile for the entire US.83 Given the evidence that scale economies, if any, are small, this

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.

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 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
average value for physical capital is a reasonable estimate 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 $0.067/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.

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 at 2003 prices. 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

---

(2004, Table 921). 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).

84 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).
optimal structure height depends on land costs. 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, these figures imply an annual fixed cost per parking space at 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, 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. 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.89

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.

85 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.

86 Shoup assumes a 40-year life, and US OMB (1992) recommends 7% real interest for project evaluation; the corresponding capital recovery factor is 0.075. For annual operating costs per stall, we use the midpoint of the ranges of annual costs implied by Cambridge Systematics et 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.

87 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).

88 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.

89 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.
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.

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. More broadly, it could be justified if other industries that compete for investment have price markups above social marginal cost similar in magnitude to their gaps between average and 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 infiltrated, 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, 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, for example by converting parking lots to other uses or by arranging for existing parking structures to be shared with nearby new users.

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 relative advantages and disadvantages of various 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 et al. (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.
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. The suspected relationship between transit-subsidy programs and service expansion beyond these markets has been confirmed statistically for the U.S. (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 we suggested in Section 3.2 may be an atypically high-cost system; 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\(^9\) in contrast to the results of Wunsch, described earlier, for a broad cross-section of European cities.

---

\(^9\) 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).
The biggest cost factor accounting for these differences is the capital cost of the 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 his 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 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 very 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. There is also strong evidence, as noted earlier, that many rail systems in the US have been approved based upon misleading projections of their costs and ability to attract riders.
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, Peromago Press plc.

(b) Santa Monica Freeway (Los Angeles). Adapted 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 $V \leq V_B$ and $S = 110.8 - \frac{33}{1000} \cdot (V - V_B)$ for $V_B < V \leq V_k$. The two HCM segments are given by $S = 110$ for $V \leq V_B$ and $S = 110 - \left[\frac{730}{28} - \frac{(V-1450)^2}{900}\right]^{0.6}$ for $V_B < V \leq V_k$. 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

- Cumulative queue entries: slope $V_a(t)$
- Cumulative queue exits: slope $V_b(t)$
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 queueing 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

<table>
<thead>
<tr>
<th></th>
<th>Rapid rail</th>
<th>Light rail</th>
<th>Bus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Boyd(^a)</td>
<td>Allport(^b)</td>
<td>Allport</td>
</tr>
<tr>
<td>Capital Cost:(^c)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per Route-Mile ($M/yr)</td>
<td>5.95</td>
<td>3.48</td>
<td>0.70</td>
</tr>
<tr>
<td>Per Peak Vehicle ($K/yr)</td>
<td>94.3</td>
<td>57.3</td>
<td>76.0</td>
</tr>
<tr>
<td>Operating Cost:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Per Route-Mile ($M/year)</td>
<td>0</td>
<td>0.834</td>
<td>0.209</td>
</tr>
<tr>
<td>Per Peak Vehicle ($K/yr)</td>
<td>0</td>
<td>57.9</td>
<td>41.0</td>
</tr>
<tr>
<td>Per Convoy-Hour(^e) ($)</td>
<td>7.18</td>
<td>37.69</td>
<td>49.44</td>
</tr>
<tr>
<td>Per Vehicle-Mile(^e) ($)</td>
<td>7.39</td>
<td>2.67</td>
<td>2.83</td>
</tr>
</tbody>
</table>

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 0.065); 12-year lifetime for bus (capital recovery factor 0.113). 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

<table>
<thead>
<tr>
<th></th>
<th>Minibus</th>
<th>Regular Bus</th>
<th>Tram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit Costs: (^a)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>per peak convoy(^b) per year ((PV))</td>
<td>196</td>
<td>557</td>
<td>1782</td>
</tr>
<tr>
<td>per convoy-hour ((VH))</td>
<td>3.4</td>
<td>13.8</td>
<td>44.6</td>
</tr>
<tr>
<td>per convoy-mi ((VM))</td>
<td>0.35</td>
<td>0.48</td>
<td>3.57</td>
</tr>
<tr>
<td>Unit costs divided by capacity (n) (for seated plus standing people):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>per person per year ((PV \cdot n))</td>
<td>7.84</td>
<td>16.88</td>
<td>23.45</td>
</tr>
<tr>
<td>per person-hour ((VH \cdot n))</td>
<td>0.14</td>
<td>0.42</td>
<td>0.59</td>
</tr>
<tr>
<td>per person-mi ((VM \cdot n))</td>
<td>0.0142</td>
<td>0.0146</td>
<td>0.0470</td>
</tr>
<tr>
<td>Percentage of category in total cost: (c_2 \cdot PV)</td>
<td>66.0</td>
<td>65.2</td>
<td>61.2</td>
</tr>
<tr>
<td>(c_3 \cdot VH)</td>
<td>16.5</td>
<td>24.8</td>
<td>23.4</td>
</tr>
<tr>
<td>(c_4 \cdot VM)</td>
<td>17.5</td>
<td>10.0</td>
<td>15.4</td>
</tr>
<tr>
<td>Employees per convoy</td>
<td>7.5</td>
<td>16.7</td>
<td>51.9</td>
</tr>
</tbody>
</table>

Source: Abbas and Abd-Allah (1999), Tables 1, 4

Notes:
\(^a\) Monetary units are Egyptian pounds (EGP) for fiscal year 1996-97. 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

<table>
<thead>
<tr>
<th>Type of cost</th>
<th>Private (Average)</th>
<th>Social Average</th>
<th>Social Marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable costs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Costs borne mainly by highway users in aggregate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) Operating &amp; maintenance</td>
<td>0.114</td>
<td>0.114</td>
<td>0.114</td>
</tr>
<tr>
<td>(2) Vehicle capital</td>
<td>0.162</td>
<td>0.162</td>
<td>0.162</td>
</tr>
<tr>
<td>(3) Travel time</td>
<td>0.286</td>
<td>0.286</td>
<td>0.367</td>
</tr>
<tr>
<td>(4) Schedule delay &amp; unreliability</td>
<td>0.088</td>
<td>0.088</td>
<td>0.163</td>
</tr>
<tr>
<td>Costs borne substantially by non-users</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5) Accidents</td>
<td>0.108</td>
<td>0.130</td>
<td>0.165</td>
</tr>
<tr>
<td>(6) Government services</td>
<td>0.005</td>
<td>0.018</td>
<td>0.018</td>
</tr>
<tr>
<td>(7) Environmental externalities</td>
<td>0</td>
<td>0.017</td>
<td>0.017</td>
</tr>
<tr>
<td><strong>Short-run variable costs</strong></td>
<td>0.763</td>
<td>0.815</td>
<td>1.006</td>
</tr>
<tr>
<td><strong>Fixed costs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8) Roadway</td>
<td>0.017</td>
<td>0.077</td>
<td></td>
</tr>
<tr>
<td>(9) Parking</td>
<td>0.006</td>
<td>0.252</td>
<td></td>
</tr>
<tr>
<td><strong>Short-run fixed costs</strong></td>
<td>0.023</td>
<td>0.329</td>
<td></td>
</tr>
<tr>
<td><strong>Total costs</strong></td>
<td>0.786</td>
<td>1.144</td>
<td>1.006</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>By Type of Cost</th>
<th>By Type of Accident</th>
<th>Cost ($/veh-mi)</th>
<th>Cost ($/veh-mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTP of death, injury</td>
<td>Fatality</td>
<td>0.095</td>
<td>0.070</td>
</tr>
<tr>
<td>Productivity</td>
<td>Disabling injury</td>
<td>0.013</td>
<td>0.023</td>
</tr>
<tr>
<td>Medical expenses</td>
<td>Other injury</td>
<td>0.008</td>
<td>0.032</td>
</tr>
<tr>
<td>Property damage</td>
<td>Prop. Damage only</td>
<td>0.007</td>
<td>0.004</td>
</tr>
<tr>
<td>Legal, police, fire</td>
<td>Unknown</td>
<td>0.003</td>
<td>0.002</td>
</tr>
<tr>
<td>Insurance admin</td>
<td></td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>Traffic delay</td>
<td></td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>TOTAL</strong></td>
<td><strong>0.130</strong></td>
<td><strong>0.130</strong></td>
</tr>
</tbody>
</table>

*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 Transit Systems

<table>
<thead>
<tr>
<th>Construction Cost (2003 US dollars)$^a$</th>
<th>per route mile ($ millions)</th>
<th>per daily trip$^b$ ($ thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heavy rail (average of 4 systems)</td>
<td>177.4</td>
<td>30.51</td>
</tr>
<tr>
<td>Light rail (average of 4 systems)</td>
<td>55.1</td>
<td>29.57</td>
</tr>
<tr>
<td>Exclusive busway:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ottawa</td>
<td>46.8</td>
<td>3.00</td>
</tr>
<tr>
<td>Pittsburgh</td>
<td>24.7</td>
<td>5.66</td>
</tr>
<tr>
<td>Shared carpool lane:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>El Monte Busway, Los Angeles</td>
<td>14.9</td>
<td>3.79</td>
</tr>
<tr>
<td>Shirley Hwy, northern Virginia</td>
<td>15.8</td>
<td>2.98</td>
</tr>
<tr>
<td>I-66, northern Virginia</td>
<td>27.8</td>
<td>8.55</td>
</tr>
<tr>
<td>Houston transitways (average)$^c$</td>
<td>6.6</td>
<td>5.71</td>
</tr>
</tbody>
</table>

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.