Chapter 21

CONGESTION MODELING

ROBIN LINDSEY*
University of Alberta

ERIK VERHOEF*
Free University, Amsterdam

1. Introduction

Traffic congestion is one of the major liabilities of modern life. It is a price that people pay for the various benefits derived from agglomeration of population and economic activity. Because land is scarce and road capacity is expensive to construct, it would be uneconomical to invest in so much capacity that travel were congestion-free. Indeed, because demand for travel depends on the cost, improvements in travel conditions induce people to take more trips, and it would probably be impossible to eliminate congestion.

Transportation researchers have long struggled to find satisfactory ways of describing and analyzing congestion, as evident from the large number of often competing approaches and models that have been developed. Early researchers hoped to develop models based on fluid dynamics that would not only be accurate, but also universally applicable. However, unlike fluid flow, congestion is not a purely physical phenomenon, but rather is the result of peoples' trip-making decisions and minute-by-minute driving behavior. One should therefore expect the quantitative – if not also the qualitative – characteristics of congestion to vary with automobile and road design, rules of the road, pace of life, and other factors. Models calibrated in a developed country during the 1960s, for example, may not fit well a developing country at the beginning of the 21st century.

Congestion in transportation is, of course, not limited to roads: it is also a problem at airports and in the airways, at harbors, on railways, and for travelers on bus and subway networks. For modeling purposes useful parallels can often be drawn between traffic congestion and congestion at other facilities. But given

*The authors would like to thank Ken Small, Richard Arnott, and André de Palma for stimulating comments on an earlier version of this chapter. Any remaining errors, however, are the authors' responsibility alone.

Handbook of Transport Modelling, Edited by D.A. Hensher and K.J. Button

space constraints, and in the interest of maintaining focus, attention is limited in this chapter to road-traffic congestion. Broadly speaking, traffic congestion occurs when the cost of travel is increased by the presence of other vehicles, either because speeds fall or because greater attention is required to drive safely. Traffic engineering is largely concerned with traffic congestion and safety, and it should therefore be no surprise that traffic-flow theory will feature prominently in this chapter.

Traffic congestion can be studied either at a microscopic level, where the motion of individual vehicles is tracked, or at a macroscopic level, where vehicles are treated as a fluid-like continuum. Queuing theory is a form of microscopic analysis. But most of the literature on queuing is of limited relevance because it focuses on steady-state conditions (which rarely prevail in traffic) and on stochastic aspects of individual customer or traveler arrival and service times (which are arguably of secondary importance, except at junctions, for traffic flows heavy enough to cause congestion) (Hurdle, 1991). Queuing theory thus will not be treated here. Car-following theory is another form of microscopic analysis that will be mentioned. Macroscopic analysis will nevertheless occupy the bulk of attention.

This chapter is organized as follows. Section 2 concerns the modeling of homogeneous traffic flow and congestion on an isolated road under stationary conditions. It also sets up the supply–demand framework used to characterize equilibrium and optimal travel volumes. Section 3 provides an overview of macroscopic and microscopic models of non-stationary traffic flow. It then describes how trip timing can be modeled, and discusses the essence of dynamic equilibrium. Section 4 reviews the principles of static and dynamic equilibrium on a road network in a deterministic environment, and then identifies equilibrium concepts that account for stochasticity in demand and capacity. Section 5 addresses conceptual and practical issues regarding congestion pricing and investment on a network. Finally, Section 6 concludes.

2. Time-independent models

Time-independent models of traffic congestion serve as a stepping stone toward the development and understanding of more complicated and realistic time-dependent models. They may also provide a reasonable description of traffic conditions that evolve only slowly. Such traffic is sometimes called "stationary", although a precise definition of "stationary" is rather difficult (Daganzo, 1997).

Traffic streams are described by three variables: density k (vehicles per lane per kilometer), speed v (km/h), and flow q (vehicles per lane per hour). At the macroscopic level these variables are defined under stationary conditions at each point in space and time, and are related by the identity $q = kv$. Driver behavior

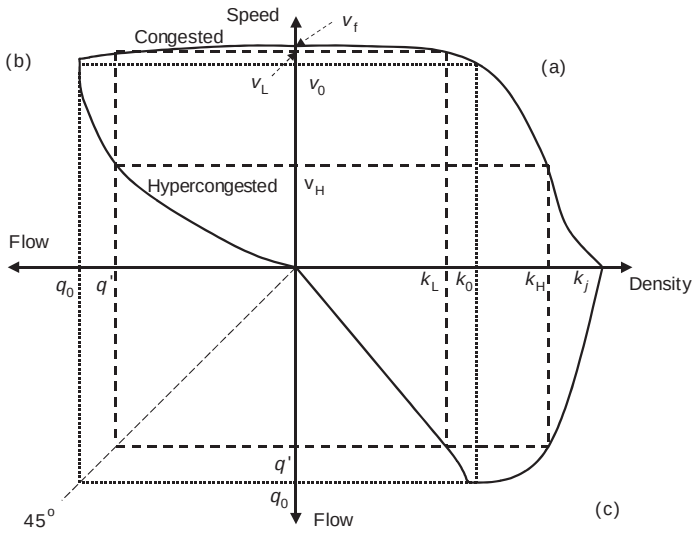

Figure 1. (a) Speed–density curve; (b) speed–flow curve; (c) flow–density curve.

creates a second functional relationship between the three variables that can be shown by plotting any one variable against another. Figure 1(a) depicts a speed–density curve, dubbed the *fundamental diagram of traffic flow* (Haight, 1963). Although studied for decades (for a literature review see May (1990)), understanding about the shape of this curve continues to evolve, as evidenced by changes to the third edition of the venerable *U.S. highway capacity manual* (Transportation Research Board, 1992). The precise shape on a given road segment depends on various factors (Roess et al., 1998, Chapters 10 and 21). These include the number and width of traffic lanes, grade, road curvature, speed limit, location vis-à-vis entrance and exit ramps, weather, mix of vehicle types, proportion of drivers who are familiar with the road, and idiosyncrasies of the local driving population.

For safety reasons speed usually declines as density increases. Nevertheless, on highways speeds tend to remain close to the free-flow speed v_f, up to flows of 1000–1300 vehicles per lane per hour. At higher densities the speed–density curve drops more rapidly, passing through the point (k_0, v_0) at which flow reaches a maximum $q_0 = k_0, v_0$, and reaching zero at the jam density k_j, where speed and flow are both zero. Speed–flow and flow–density curves corresponding to the speed–density curve in Figure 1(a) are shown in Figure 1(b) and 1(c), respectively. Note that any flow $q' < q_0$ can be realized at either a low density and high speed (k_L, v_L), or at a high density and low speed (k_H, v_H). Economists refer to the upper branch

of the speed–flow curve as *congested* and to the lower branch as *hypercongested*. In the engineering literature the upper branch is variously referred to as *uncongested*, *unrestricted*, or *free flow*, and the lower branch as *congested*, *restricted*, or *queued*. The term "queued" is apposite for the hypercongested branch in that queuing usually occurs in this state (see Section 3). But congestion also occurs on the upper branch whenever speed is below the free-flow speed. For this reason, the economics terminology will be used here.

Following Walters (1961), the speed–flow curve can be used for economic analysis by interpreting flow as the quantity of trips supplied by the road per unit of time. A trip-cost curve can be generated of the form $C(q) = c_0 + \alpha L/v(q)$, where α is the unit cost of travel time, L is trip distance, $v(q)$ is speed expressed in terms of flow, and c_0 denotes trip costs other than in-vehicle travel time, such as monetized walk access time and fuel costs (if these costs do not depend on congestion). The trip-cost curve (Figure 2) has a positively sloped portion corresponding to the congested branch of the speed–flow curve, and a negatively sloped backward-bending portion corresponding to the hypercongested branch. A flow of q' can be realized at a cost C_L on the positively sloped portion, as well as at a higher cost C_H on the negatively sloped portion. Because the same number of trips is accomplished, the latter outcome is inefficient.

If flow is also interpreted to be the quantity of trips "demanded" per unit of time, then a demand curve $p(q)$ can be combined with $C(q)$ to obtain a supply–demand diagram. Candidate equilibria occur where $p(q)$ and $C(q)$ intersect. In Figure 2 there are three intersection points x, y and z, with flow congested at x and hypercongested at y and z. There has been a heated debate in the literature (for

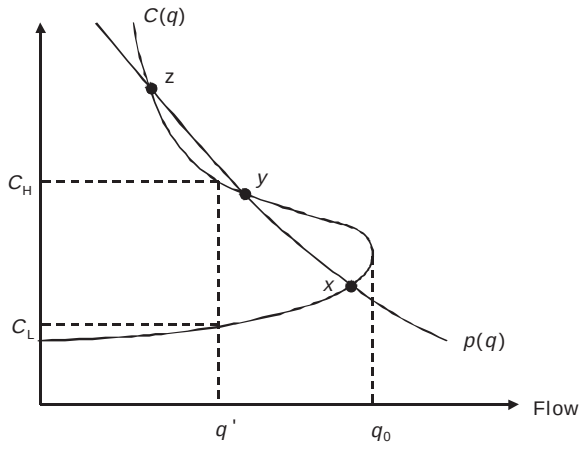

Figure 2. Backward-bending travel-cost curve $C(q)$ and travel-demand curve $p(q)$.

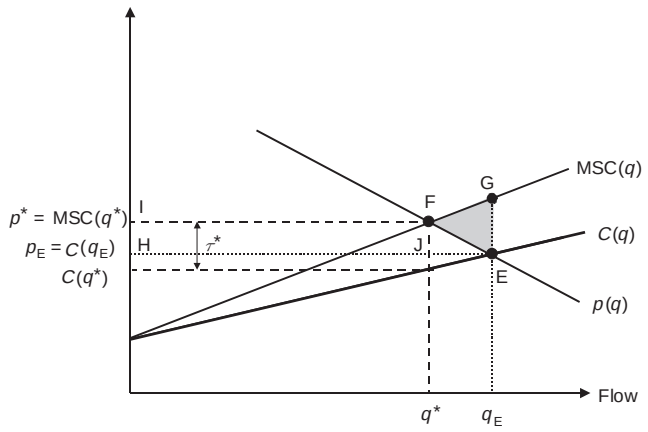

Figure 3. Equilibrium road usage q_E, optimal road usage q^*, and optimal congestion toll τ^*.

reviews and recent contributions to this debate see Chu and Small (1997), Verhoef (1999), and McDonald et al. (1999)) about whether hypercongested equilibria are stable, and also whether it makes sense to define the supply and demand for trips in terms of flow. The emerging view seems to be that hypercongestion is a transient phenomenon that can be properly studied only with dynamic models (see Section 3).

For economic analysis (e.g., Button, 1992), it is common to ignore the hypercongested branch of the speed–flow curve and to specify a functional form for $C(q)$ on the congested branch directly, rather than beginning with a speed–density function and then deriving $C(q)$. Given $C(q)$, the socially optimal usage of the road and the congestion toll that supports it can be derived, as shown in Figure 3. As in Figure 2, the unregulated equilibrium flow q_E occurs at point E, the intersection of $C(q)$ and $p(q)$. Now, since "external benefits" of road use are not likely to be significant (benefits are normally either purely internal or pecuniary in nature), $p(q)$ specifies both the private and the marginal social benefit of travel. Total social benefits can thus be measured by the area under $p(q)$. Analogously, $C(q)$ measures the cost to the traveler of taking a trip. If external travel costs other than congestion, such as air pollution or accidents, are ignored, then $C(q)$ measures the average social cost of a trip. The total social cost of q trips is then $\mathrm{TC}(q) = C(q)\,q$, and the marginal social cost of an additional trip is $\mathrm{MSC}(q) = \mathrm{f}\,\mathrm{TC}(q)/\mathrm{f}\,q = C(q) + q\,\partial C(q)/\partial q$.

The socially optimal number of trips q^* occurs in Figure 3 at point F where $\mathrm{MSC}(q)$ and $p(q)$ intersect. The optimum can be supported as an equilibrium if travelers are forced to pay a total price of $p^* = \mathrm{MSC}(q^*)$. Because the price of a trip is the sum of the individual's physical travel cost and the toll $p = C(q) + \tau$, the

requisite toll is $\tau^* = \text{MSC}(q^*) - C(q^*) = q^* \partial C(q^*)/\partial q$, where $q^* \partial C(q^*)/\partial q$ is the marginal congestion cost imposed by a traveler on others. This toll is known as a Pigouvian tax, after its spiritual father Pigou (1920).

Imposition of the toll raises social surplus by an amount equal to the shaded area FGE in Figure 3. Nevertheless, travelers end up worse off if the toll revenues are not used to benefit them. The q^* individuals who continue to drive each suffer a loss per trip of $p^* - p_E$, resulting in a collective loss equal to area HIFJ, and the $q_E - q^*$ individuals who are priced off the road, either because they switch to another mode or give up traveling, suffer a collective loss equal to area JFE. These losses are the root of the long-standing opposition to congestion tolling in road transport. Transportation analysts and planners are now trying to devise ways of spending toll revenues so as to improve the acceptability of pricing (Small, 1992b).

3. Time-dependent models

Time-dependent or dynamic traffic models allow for changes in flow over time as well as over space. The most widely used dynamic macroscopic model is the *hydrodynamic model* developed by Lighthill and Whitham (1955) and Richards (1956) (the LWR model) (for a review see Daganzo (1997)). The essential assumption of the LWR model is that the relationship in stationary traffic between speed and density, shown in Figure 1, also holds under non-stationary conditions. The model is completed by imposing the condition that vehicles are neither created nor destroyed along the road. If x denotes location and t time, and if the requisite derivatives exist, this results in a partial differential equation, $\partial q(t, x)/\partial x + \partial k(t, x)/\partial t = 0$, known as the *conservation equation*. In cases where q and k are discontinuous, and therefore not differentiable, a discrete version of the conservation equation still applies, as shown in the following example.

To illustrate how the LWR model behaves, suppose that traffic on a roadway is initially in a congested stationary state A with density k_A, speed v_A, and flow q_A, as shown in Figure 4(a). Inflow at the entrance then falls abruptly from q_A to q_B, moving traffic to a new state B at another point on the same flow–density curve. State B will propagate downstream as a *shock wave* with some speed w_{AB}. Vehicles upstream in state B catch up to the shock wave at a speed $v_B - w_{AB}$, and thus leave state B at a flow rate $(v_B - w_{AB})k_B$. Given conservation of vehicles, this must match the rate at which they enter state A: $(v_A - w_{AB})k_B$. Equating the two rates, and recalling that $q_i = k_i v_i$, i = A, B, one obtains $w_{AB} = (q_A - q_B)/(k_A - k_B)$. This wave speed corresponds to the slope of a line joining states A and B on the flow–density curve in Figure 4(a). The wave speed is slower than the speed of vehicles in either state, v_A and v_B.

The trajectories of representative individual vehicles in this thought experiment are shown by arrows in the *time–space diagram* (Figure 4(b)). Prior to the change

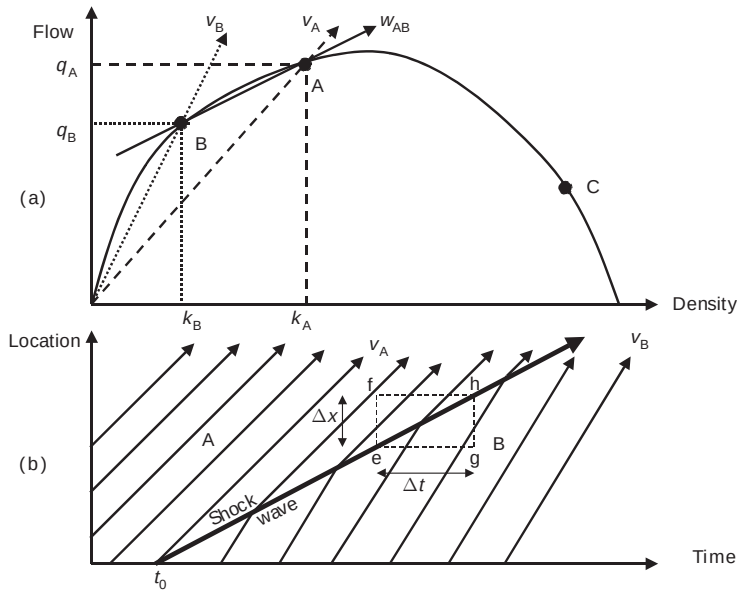

Figure 4. (a) Transition from A to B on the flow–density curve; (b) trajectories in the time–space diagram. Adapted from May (1990, Figure 11.1).

in inflow, vehicles are moving to the north-east at speed v_A. If the time and location axes are scaled appropriately, vehicle trajectories have the same slope as vehicle speeds in Figure 4(a). At time t_0 the inflow falls to q_B, and the trajectories of incoming vehicles increase in slope to v_B. As vehicles reach the shock wave, shown by the thicker line extending north-east from point $(t_0, 0)$, they slow down to v_A. Because vehicles slow down instantaneously, their trajectories are kinked where they cross the shock wave. Thus, throughout the time–space diagram vehicles are either traveling at speed v_A in traffic of density k_A, or at speed v_B in traffic of density k_B. Intermediate densities and speeds never develop in this particular thought experiment. Note, finally, that the horizontal spacing between vehicle trajectories is greater in state B than in state A because $q_B < q_A$.

A discrete version of the conservation equation can be derived by referring to the small rectangle with length Δt and height Δx, shown by dashed lines in Figure 4(b). The number of vehicles entering the rectangle from side eg($q_B\,\Delta t$) and side ef($k_A\,\Delta x$) must equal the number exiting from side fh($q_A\,\Delta t$) and side gh($k_B\,\Delta x$). This gives $(q_B - q_A)\Delta t + (k_A - k_B)\Delta x$, or $\Delta q/\Delta x + \Delta k/\Delta t = 0$, where $\Delta q \ldots q_B - q_A$ and $\Delta k \ldots k_A - k_B$ because the density changes from k_B to k_A when moving downstream across the wave boundary.

The shock wave in this example is a *forward-recovery* shock wave because it signals a reduction in density that propagates downstream. If the transition were in the opposite direction, from B to A, a *forward-forming* shock wave would result, conveying an increase in density downstream with a speed $w_{BA} = (q_B - q_A)/(k_B - k_A)$ that is the same as w_{AB}. In contrast to the situation depicted in Figure 4(b), vehicles would have the higher speed v_B to the north-west of the shock wave, and the lower speed v_A to the south-east. Vehicles would accelerate upon crossing the shock wave in response to the reduction in density from k_A to k_B.

Finally, consider a transition from state C in Figure 4(a) to state B. Because a line (not shown) joining B and C on the flow–density curve has a negative slope, a *backward-forming* shock wave would result that propagates upstream of the roadway entrance. Several other types of shock wave are also possible (see May, 1990, Section 7.4 and Chapter 11).

Shock-wave analysis is useful for studying discrete changes in traffic conditions such as temporary capacity reductions. But the accuracy of shock-wave analysis is limited by the assumption of the LWR model that a given speed–density relationship holds exactly at each point in time and space, regardless of what conditions drivers may have recently encountered, or what conditions they may anticipate by looking ahead. Moreover, the LWR model assumes that vehicles can adjust speed instantaneously; i.e., with (physically impossible) infinite acceleration or deceleration, as manifest in Figure 4(b) by the kinks in vehicle trajectories at the shock-wave boundary. The LWR model also does not account for differences between drivers in desired speed that create incentives to pass. And the model cannot explain instabilities in traffic flow such as stop-and-go conditions (Daganzo, 1997, Section 4.4.6).

A further drawback of the LWR model is that deriving a solution, either using shock-wave diagrams or analytically using the speed–density relationship and conservation equation, is tedious on inhomogeneous roadways or when inflow varies continuously over time. For the sake of tractability, various simplifications of the model have been formulated, three of which are mentioned below.

One simplification, widely used for analyzing bottlenecks and called the *bottleneck model* here, is to assume that the congested branch of the speed–flow curve at the bottleneck is horizontal up to maximum flow or capacity s. If the incoming flow exceeds s, traffic flows through the bottleneck at rate s, and the excess flow accumulates in the form of a queue propagating upstream as a backward-forming shock wave. Some recent empirical evidence (e.g., Cassidy and Bertini, 1999) indicates that discharge rates from bottlenecks fall after queue formation and then partially recover. The constant-flow assumption nevertheless appears to serve as a reasonable approximation to observed behavior.

An example of queue evolution in the bottleneck model is shown in Figure 5. Curve $D(t)$ denotes the cumulative number of vehicles that have passed or departed the bottleneck since some initial reference time. Curve $A(t)$ denotes the

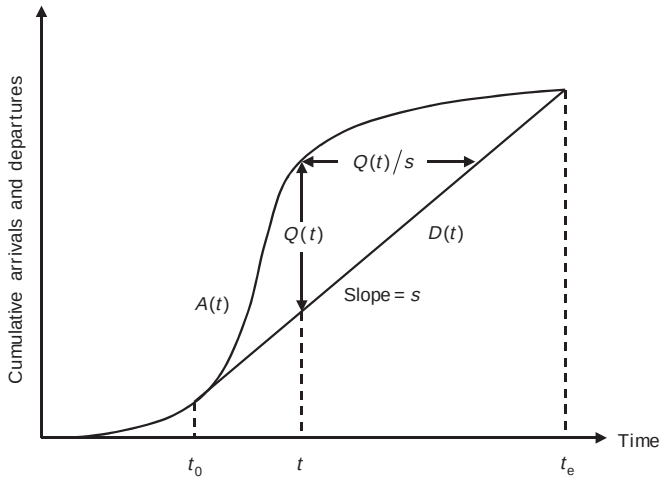

Figure 5. Cumulative arrivals and departures curves and queue evolution.

number of vehicles that have arrived at the tail of the queue upstream. (This terminology is not universal; some of the literature refers to the curve here labelled $A(t)$ as cumulative departures from the origin, and to the curve here labelled $D(t)$ as arrivals at the destination.) Prior to time t_0 the arrival flow is less than s, so that no queue forms and $D(t)$ and $A(t)$ coincide. Between t_0 and t_e a queue exists. The vertical distance $Q(t)$ between $A(t)$ and $D(t)$ measures the number of vehicles in the queue at time t. The horizontal distance $Q(t)/s$ measures time spent queuing by a vehicle that arrives at time t. Total queuing time for all vehicles is simply the area between $A(t)$ and $D(t)$.

Cumulative-count diagrams such as Figure 5 are commonly used to predict queues caused by scheduled maintenance or accidents. They can trace the growth and decay of several queues in sequence, and can deal with situations in which the capacity of the bottleneck changes, or depends on the length of the queue, so that, unlike in Figure 5, $D(t)$ is non-linear. Such diagrams can describe the impact of a "moving bottleneck" such as a slowly moving truck (Gazis and Herman, 1992).

One frequently overlooked fact is that queues are not dimensionless points, but rather occupy (sometimes kilometers of) road space, and relatedly that vehicles in the queue are not stationary but moving slowly forward. Vehicles arriving at the location of the tail of a queue would take time to reach the bottleneck, even with no queue present. Individual-vehicle delay is thus less than $Q(t)/s$, and total delay is less than the area between $A(t)$ and $D(t)$. Because travel costs are generally assumed to depend on delay, rather than queuing time *per se*, failure to account for the physical length of queues can lead to an overestimate of travel-time losses.

Accounting for the length of queues is also important if queues can spill back and block upstream junctions, or entry and exit ramps. Still, for some purposes it is unnecessary to keep track of the physical length of queues, and models of networks (see Section 4) sometimes work with so-called *vertical queues*.

A second variant of the LWR model, hinted at by Walters (1961, p. 679) and adopted by Henderson (1977), embodies the assumption that on uniform roadways a vehicle travels at a constant speed determined by the speed–density curve and the density prevailing when the vehicle enters. This means that shock waves travel at the same speed as vehicles and therefore never influence other vehicles. We call this here the *no-propagation model*. One problem with this model is that a vehicle departing under low-density conditions may catch up with and overtake a vehicle that departed earlier when the density was higher. Yet overtaking is not allowed in the original LWR model, has no behavioral foundation if drivers and vehicles are identical, and is likely to be impossible anyway when congestion is heavy.

A third variant of the LWR model, here called the *instantaneous propagation model*, was adopted by Agnew (1977) and Mahmassani and Herman (1984). It entails the assumption that density (and hence speed) remains uniform along the roadway. An increase in input flow, for example, is immediately absorbed by an equal increase in density everywhere along the road. This implies that shock waves propagate with infinite speed, and relatedly that vehicles can be affected by traffic behind them, contrary to what is assumed in other traffic-flow models. While the instantaneous propagation model is unrealistic, it has been adopted for individual links or road segments in some network models (see Section 4). Instantaneous propagation may be descriptive of congestible facilities such as computers, where there is no spatial analog of distance and where speed or time of service does not depend on order of entry into the system.

We now turn our attention briefly to microscopic models that treat vehicles as discrete entities, rather than elements of a continuum. Microscopic models are used to describe traffic behavior on lightly traveled roads where passing and lane changing are possible. Such models predict that, consistent with what is observed, the congested branch of the speed–flow curve is horizontal at zero flow. This is because in very light traffic a vehicle can almost always pass another vehicle without delay, while if it is delayed it is usually due to a conflict with just one other vehicle (see Daganzo, 1997, Section 4.2.3).

Microscopic models are also useful for tracking the progress of vehicles along heavily congested roads, through signalized or non-signalized intersections, and on networks. For example, May et al. (1999) use a microsimulation computer model to generate aggregate speed–flow relationships for an area. Such relationships can be combined with demand curves to predict traffic volumes, either as stationary equilibria or on a temporally disaggregated basis, as in Chu and Small (1997).

A widely used class of microscopic models are car-following models, which were developed in the 1950s and 1960s. Such models usually describe the motion of vehicle $n + 1$ (the "follower") in a traffic stream as a function of the motion of vehicle n (the "leader") immediately ahead. A relatively general formulation (see May, 1990, Section 6.2) is given by the differential equation:

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

where x denotes location, one dot a first derivative, two dots a second derivative, Δt a reaction-time lag, and c, l and m are non-negative parameters. The left-hand side of eq. (1) is the response of the follower in terms of lagged acceleration. The right-hand side is the stimulus, which is an increasing function of the follower's speed, a decreasing function of the distance to the leader, and proportional to the difference in the two vehicles' speeds. Eq. (1) describes stable behavior if a small perturbation in the speed of one vehicle in the stream is attenuated as it propagates along the chain of vehicles that follow, so that safe headways between vehicles are maintained. Stability turns out to prevail if the product $c\ \Delta t$ is not too large; i.e., if responses are rapid (small Δt) but gentle (small c).

Under stationary traffic conditions, car-following models imply a relationship between density (the inverse of vehicle spacing) and speed that can be described by the LWR or other macroscopic model. But car-following models are more realistic in recognizing that vehicles accelerate or decelerate at finite rates, and drivers react with time lags. Such models can also specify the response of a vehicle to the motion of vehicles two or more positions ahead in the traffic stream, in recognition of the fact that drivers may look at traffic conditions well downstream to give themselves more time to react. Under rapidly changing traffic conditions, where the LWR model may fail to perform adequately, a car-following model with the same stationary behavior can be used instead.

Both macroscopic and microscopic models are being used to address various traffic-flow and congestion phenomena that await definitive treatments. Phase transitions are one alleged phenomenon whereby free-flowing traffic can spontaneously break down for no obvious reasons and persist in a self-maintained congested state for long periods (Kerner and Rehborn, 1997). Such behavior – which is disputed (Daganzo et al., 1999) – poses a challenge to traffic managers seeking to maintain smoothly flowing traffic.

Hypercongestion is another phenomenon that has attracted attention since the work by Walters (1961). Hypercongestion routinely occurs on non-uniform roadways. As described above, it occurs in queues upstream of a saturated bottleneck. One question under debate is whether hypercongestion is possible on a uniform roadway that has no intermediate entrances and is not initially hypercongested. As Newell (1988) observes, it cannot happen in the LWR model.

To see this heuristically, note first that inflow at the entrance cannot exceed the maximum flow q_0 (see Figure 1) so that hypercongestion cannot develop from upstream. Now consider a platoon of vehicles traveling at flow q_0 and corresponding density k_0. In order for these vehicles to experience hypercongestion they must encounter a shock wave from higher density traffic downstream. But this cannot happen because there is no hypercongestion on the road initially.

Another question is how to model hypercongestion on a realistic city network. Chu and Small (1997) address this using two versions of a model of a spatially homogeneous urban commuting corridor. The first version adopts the instantaneous propagation model and assumes a constant and exogenous inflow of vehicles. Hypercongestion develops when the inflow exceeds the capacity of the corridor for long enough. While an inflow exceeding capacity is impossible on an isolated road according to the LWR model, it is possible if, as Chu and Small assume, there are intermediate entrances along the route. The second version of their model features endogenous inflow (discussed below). Again, hypercongestion can occur.

Low speeds and flows, characteristic of hypercongestion, are indeed common in urban areas. This is attributable in part to conflicting traffic flows at intersections, and in part to the fact that road network capacity is limited near city centres. Hypercongestion during the morning rush hour may also be aggravated by reductions in road capacity as on-street parking spots become occupied, or as queues develop of vehicles waiting to enter off-street parking lots.

The discussion of time-dependent models thus far has focused on the behavior of vehicles once in a traffic stream, while neglecting the determinants of inflow (i.e., travel demand). In Section 2 demand was described by a demand curve that accounts for the dependence of demand on the cost of travel, but not on when travel takes place. Yet it is evident from the diurnal, weekly, and seasonal fluctuations in traffic volumes that people do care about when they travel. Indeed, if traffic were spread uniformly over time, congestion would not be a serious problem.

Since time of travel does matter, it is necessary to model how easily trips can be substituted forward or backward in time. One extreme, but common, assumption is that trips are not intertemporally substitutable, so that the demand for trips at a given time depends only on the cost of making a trip at that instant. A more general approach, pioneered by Vickrey (1969), is to assume that each individual has a preferred time t^* to complete a trip, and incurs a *schedule-delay cost* for arriving either earlier or later. It is often assumed that this cost is linear, increasing by some amount β for each additional minute early (before t^*), and by some amount γ for each extra minute late. Small's (1982) empirical estimates for morning commuting trips satisfy $\beta < \alpha < \gamma$, where (as in Section 2) α is the unit cost of travel time.

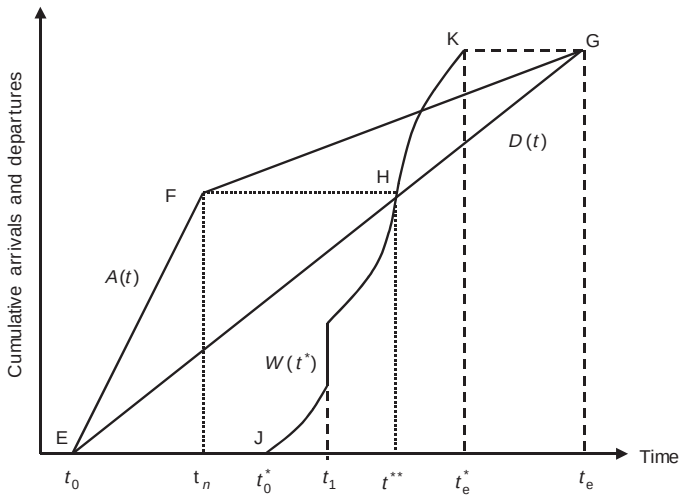

Figure 6. Equilibrium trip timing and queue evolution in the bottleneck model.

Given a schedule-delay-cost function, it is straightforward to solve for the equilibrium timing of trips along a single roadway connecting a single origin and destination. The LWR model, as well as the bottleneck, no-propagation and instantaneous-propagation versions of it, have all been used in the literature to describe traffic flow on the roadway. Equilibrium in the bottleneck model version is readily depicted using an augmented version of Figure 5, shown in Figure 6. The new element is the curve $W(t^*)$, which specifies the cumulative distribution of t^* in the population. To fix ideas, consider morning commute trips, so that t^* is desired arrival time at work. As drawn, the distribution of t^* extends from t_0^* to t_e^*, and has a "mass point" at t_1 (perhaps because a company has a large shift of workers that starts at this time). To simplify, it is assumed that commuters have the same values of α, β and γ, and demand is price inelastic: N individuals commute, one per vehicle, regardless of the trip cost. Free-flow travel times before and after passing the bottleneck are set to zero, and the queue is assumed to be vertical, as defined above (i.e., zero length).

In this setting, commuters have only one decision to make: at what t to join the queue behind the bottleneck. A Nash equilibrium is defined by the condition that no individual can reduce their trip cost by changing thier t, taking as given the travel-time choices of everyone else. An algebraic derivation of the equilibrium is given in Arnott et al. (1998); only a heuristic explanation will be given here. Trip cost is composed of schedule-delay cost, queuing-time cost, and any fixed costs independent of t. The queue upstream of the bottleneck must therefore evolve at such a rate that the sum of queuing-time cost and schedule-delay cost is

independent of t. This results in a piecewise queuing pattern, as shown in Figure 6, because schedule-delay costs are assumed linear. The four unknown times $\{t_0, t_n, t^{**}, t_e\}$ are determined by four equations. One equation simply states that the rush hour is long enough for everyone to get to work: $s(t_e - t_0) = N$. A second equation defines t^{**} as that time at which the number of individuals who want to have arrived at work equals the number who have actually done so: $W(t^{**}) = s(t^{**} - t_0)$. The third and fourth equations obtain from the condition that the individual who departs at time t_n and arrives at work on time at t^{**} incurs the same trip cost departing at t_n as they would if they departed early at t_0, or late at t_e: $\beta(t^{**} - t_0) = \alpha(t^{**} - t_n) = \gamma(t_e - t^{**})$.

As in Figure 5, total queuing time in this equilibrium is given by the area EFG between $A(t)$ and $D(t)$. Total time early is area EHJ, and total time late is area GHK. Because aggregate schedule-delay costs are the same order of magnitude as total queuing-time costs (if everyone has the same t^* they turn out to be equal) it is important to account for schedule delay in determining total travel costs. It is straightforward to compute an equilibrium travel-cost function $C(N)$ conditional on N. If demand is price elastic, N can then be solved as in Section 2, with the condition $p(N) = C(N)$, where $p(N)$ is the inverse demand curve.

If individuals differ not only with respect to t^*, but also α, β and γ, the geometry of equilibrium becomes more complicated but an analytical solution is still possible. Suppose there are G groups of homogeneous individuals. Given the number of individuals N_g in group g ($g = 1, \ldots, G$), one can derive parametric equilibrium travel-cost functions of the form $C_g(N_1, N_2, \ldots, N_G)$. Then, given the demand curves $p_g(N_g)$, the equilibrium values of N_g can be solved using the G equations $p_g(N_g) = C_g(N_1, N_2, \ldots, N_G), g = 1, \ldots, G$.

4. Modeling congestion on a network

Although attention has been limited so far to isolated road segments, most trips occur on a road network. A network can be represented as a set of origin–destination (OD) pairs, a set of routes connecting each OD pair, and a set of directed links for each route, where links may be shared by more than one route.

A conceptual framework for solving stationary equilibrium traffic flows on a network was developed by Wardrop (1952). *Wardrop's first principle* states that in equilibrium, the costs of trips for a given OD pair must be equal on all used routes (i.e., routes that receive positive flow), and no lower on unused routes. If demand is price elastic, then in addition the marginal benefit of a trip for an OD pair must equal the trip cost if any trips are made, and be no bigger than the cost if no trips are made. In a Wardrop equilibrium no individual has incentive to change either his route or his decision whether to travel. A Wardrop equilibrium is, therefore, a Nash equilibrium. If no tolls are levied on the network, then the equilibrium is said

to be a *user equilibrium* or *user optimum*. Beckmann et al. (1956) showed that a user equilibrium can be formulated and solved as an equivalent optimization problem.

Because of unpriced congestion externalities, user equilibrium is generally inefficient, both in terms of the number of trips taken between each OD pair, and the allocation of demand over routes. Efficient usage occurs at a *system optimum*, which is defined by the same conditions as a user equilibrium, but with the marginal social cost of using each route in place of the average (user) cost, where the marginal social cost of a route is the sum of the marginal costs on each link comprising the route. Thus, marginal costs must be equal on all used routes between a given OD pair, and no lower on unused routes (*Wardrop's second principle*). This assures that total travel costs are minimized for the trips taken on the network. If demand is price elastic, then in addition the marginal benefit of a trip for an OD pair must equal the marginal social cost if any trips are made, and be no bigger if no trips are made.

The system optimum can be decentralized as a user equilibrium by imposing tolls on each link. If the travel cost on each link depends on flow on the link, but not on flows on other links (this rules out interactions at junctions, or between opposing traffic flows on undivided highways), then the tolls take the same form as in the one-link setting described in Section 2 (Dafermos and Sparrow, 1971). Thus, if C_l is the link cost function on link l and q_l^* is the optimal flow, then the optimal toll on link l is $q_l^* \, \partial C_l(q_l^*)/\partial q_l$.

Characterizing network equilibrium with non-stationary traffic flows, and then solving for the equilibrium, is more difficult both conceptually and computationally than with stationary flows. Questions arise about how to model congestion on individual links, and how to maintain first-in, first-out discipline if passing is not permitted. *Dynamic traffic assignment* is concerned with solving for user equilibrium routing while treating departure times as given. Finding a *dynamic network user equilibrium* requires also solving for departure times. Both problems have their system-optimal counterparts. Various dynamic generalizations of Wardrop's first and second principles have been proposed. For simple networks where routes share no links, an equilibrium can be found by first solving for the travel cost functions as described at the end of Section 3, and then applying Wardrop's principles for stationary traffic. In more complex cases, sophisticated programming methods are required. Progress has recently been made through the use of variational inequalities (Ran and Boyce, 1996; Nagurney, 1999).

Wardrop's equilibrium principles are based on the implicit assumption that drivers know the travel costs on each route and at each time exactly. To allow for less than perfect information, Daganzo and Sheffi (1977) introduced an equilibrium concept for stationary traffic called *stochastic user equilibrium* (SUE). In this framework, travelers have idiosyncratic perceptions of travel times on each

route, and seek to minimize their expected or perceived travel costs. SUE has been extended to dynamic networks by adding idiosyncratic perceptions of travel costs as a function of departure time.

SUE incorporates random behavior at the individual level, but embodies an implicit law of large numbers assumption because aggregate flows on each route and at any time are deterministic. Hazelton (1998) has introduced randomness in aggregate flows by treating vehicles as discrete and finite in number. This allows for day-to-day variations in flow, and may be useful for modeling driver learning.

In SUE, randomness originates from driver perception errors, rather than from aggregate demand or from the network itself. In practice, demand can fluctuate unpredictably because of special events, and capacity can be affected by weather, road work, and accidents. One way to allow for this type of randomness is to suppose that drivers choose routes, possibly with guidance from a *motorist information system* (see Section 6), on the basis of current travel times without attempting to predict how these times will evolve during the rest of their trips (Wie and Tobin, 1998). Another approach, termed *stochastic network stochastic user equilibrium* by Emmerink and colleagues (see Emmerink, 1998, Chapter 4), assumes that drivers minimize expected trip costs while conditioning their expectations on all information available to them.

5. Road pricing and investment

The principles of congestion pricing for stationary traffic and identical vehicles were introduced in Sections 2 and 4. These principles were extended by Dafermos (1973) to treat heterogeneous vehicles that differ in size, operating characteristics, or other aspects of behavior. Traffic engineers adjust for the greater impact of heavy vehicles on traffic by computing passenger-car equivalents, and tolls could be based on these. Alternatively, heavy vehicles can be modeled as causing reductions in road capacity. Charging on the basis of speed, with higher tolls for slower vehicles, has been studied by Verhoef et al. (1998). Surcharges might also be imposed on poor or careless drivers who tend to create greater congestion and are more prone to accidents. But unless charges can be levied non-anonymously, perhaps via *automatic vehicle identification systems*, tolling on the basis of driving behavior is impractical because it is too costly to observe.

Varying tolls over time has become practical through advances in *electronic toll collection* technology. Time variation can range from peak/off-peak tolls with a single step to continuous time variation. The optimal continuously time-varying toll in the bottleneck model can be readily deduced by inspection of Figure 6. Because the capacity of the bottleneck is independent of queue length, queuing is pure dead-weight loss. Queuing can be eliminated by imposing a toll at each instant equal to the cost of queuing time that would have obtained in the no-toll

user equilibrium. The toll is zero at the beginning of the travel period t_0 rises linearly to a maximum at t^{**}, and decreases linearly to zero again at t_e. Because the toll exactly offsets queuing-time cost, private costs of drivers are unchanged. Aggregate schedule-delay costs are also unchanged because, with a fixed bottleneck capacity, both the timing and the duration of the travel period are the same.

This invariance of private-travel costs and schedule-delay costs to the tolling regime is specific to the bottleneck model. In the LWR model and its no-propagation and instantaneous-propagation variants, flow varies with speed. The optimal time-varying toll causes departures to spread out, which reduces travel-time costs by raising travel speeds (though not to free-flow levels), but increases total schedule-delay costs by a partially offsetting amount. Chu (1992) demonstrated these results using a modified version of the no-propagation model in which the speed of a vehicle is determined by the density prevailing when it exits the road. He also shows that, unlike with the no-propagation model, overtaking does not occur in either the no-toll or optimally tolled equilibria of the reformulated model.

Research is underway on how to derive and implement system-optimal time-varying tolls on a network. Among the challenges that have to be addressed are how to calculate the marginal social cost of a trip, how to make the driver pay this cost using link-based tolls or path-based tolls, and how to apprise individuals about tolls sufficiently far in advance to influence their travel decisions. Another complication is that Pigouvian tolls are efficient only in a first-best world with efficient pricing throughout the economy. This requires not only that congestion pricing be applied network-wide by time of day, type of vehicle, etc., but also that environmental and other externalities be internalized, that other modes of travel be efficiently priced, and so on. In practice, first-best conditions are not satisfied even approximately. For one thing, infrastructure costs and political constraints are likely to rule out tolling except on major roads.

Transportation economists have devoted considerable attention over the years to second-best pricing of transit in the face of unpriced automobile congestion. More recently, there has been some work on optimal second-best pricing on simple traffic networks with stationary traffic flows. Verhoef et al. (1996) considered a single OD pair connected by two congestible routes, one of which was untolled. They showed how the second-best toll on the tolled route differed from the first-best toll, and showed that it may be negative in order to discourage usage of the untolled route. Glazer and Niskanen (1992) consider the relationship between the price of parking and traffic congestion, and discuss how first-best parking fees should be modified when traffic congestion is not priced. Much remains to be understood about second-best tolling on large-scale networks with non-stationary traffic.

Although economically appealing, road pricing remains politically controversial, and awaits widespread implementation. Building roads has been the traditional response to growing congestion. But construction of new roads is increasingly constrained by shortages of public funds and land space, and by environmental concerns. It is apparent that a combination of demand restraint, improvements in existing roads, and selective construction of new ones, will be required in the future. In deciding how much to invest in roads, it is important to recognize that optimal capacity depends generally on how demand is regulated, and specifically on the tolling regime (Small, 1992a, Sections 4.1 and 4.4). To see this, consider a stationary traffic setting and suppose that capacity is increased, which shifts the travel-cost curve ($C(q)$ in Figure 3) to the right. Absent tolling, equilibrium will be established at the new intersection of $C(q)$ with the demand curve. If demand is highly elastic, travel volume will increase until the cost of a trip is only slightly below its previous level, and the investment will yield little benefit. The increase in volume comes from so-called *latent demand*: trips attracted from other routes or modes, or new trips that were deterred by congestion. With tolling, however, the increase in volume is restrained, and the investment may yield an appreciable welfare gain. By contrast, if demand is relatively (but not completely) inelastic then latent demand is less of a force. Because more trips are taken without tolling, the investment is likely to yield a greater benefit without than with tolling.

In the case of non-stationary traffic, the analysis is complicated by the fact that imposition of a time-varying toll reduces travel costs for any given capacity. But the effect of demand elasticity on the relative returns from investment with and without tolling remains qualitatively the same.

Given the increasing popularity of the user pay principle, it is natural to ask: to what extent do optimal congestion tolls pay for optimal capacity? Mohring and Harwitz (1962) showed using a static model that congestible facilities (of which roads are one instance) are exactly self-financing if three conditions hold:

(1) capacity is adjustable in continuous increments,
(2) capacity can be expanded at constant marginal cost, and
(3) trip costs are homogeneous of degree zero in usage and capacity (i.e., doubling N and s leaves costs unchanged).

Although the empirical evidence on (2) and (3) is equivocal, it appears that these conditions hold at least approximately in a range of circumstances (Small, 1992a, Sections 3.4 and 3.5; Hau, 1998). Condition (1) does not hold on a single road because the number of lanes is discrete and lanes must be large enough to accommodate vehicles. But capacity can still be varied by widening lanes, by improving vertical and horizontal alignments, and by resurfacing. And at the scale of a road network, capacity may be almost perfectly divisible. Furthermore, the self-financing theorem extends to dynamic models (Arnott et al., 1993), and in

present-value terms when adjustment costs and depreciation are allowed (Arnott and Kraus, 1995).

The self-financing theorem concerns optimal highway investment in a first-best world. Just as care must be taken in setting tolls when travel is not optimally priced on all of the road network, so must investment decisions be made with caution. This is illustrated dramatically by the famous "Braess paradox" (Braess, 1968), whereby adding a link to an untolled network can actually increase total travel costs. Various other paradoxes can also arise with unpriced (or underpriced) congestion (see Arnott and Small, 1994).

6. Conclusions

As this review should make clear, there is no single best way to model traffic flow and congestion. The level of detail at which driver behavior should be modeled depends on the objectives of the analysis. For the purpose of studying land use, for example, a model of stationary traffic flow may be adequate, and this requires only a relationship between speed and density. Non-stationary traffic phenomena, such as the rush hour, hypercongestion and passing, are more complex and may call for a microscopic rather than macroscopic approach. As is true of most scientific endeavors, there is a trade-off in modeling between realism and tractability. With today's computers it is possible to simulate the minute-by-minute progress of many thousands of vehicles on a large-scale network. Still, the complexities of simulation models and the sheer volume of output they can generate may obscure basic insight. A role thus remains for simple models that are amenable to analytical and/or graphical solution.

Many policies have been adopted to combat congestion, both on the supply side (e.g., building new roads, restriping lanes) and in managing demand (e.g. priority lanes, metering highway entrance ramps, parking restrictions and license plate rules). Attention has been limited in this review to congestion pricing, in part because of its close links with the fundamental diagram of traffic flow and with network equilibrium conditions. In his discussion of congestion, Walters (1987) came out strongly in favor of congestion pricing, but was pessimistic about its prospects for implementation. Thanks to continuing technological advances and shifts in political attitudes, the perspective at the end of the 21st century seems rather more sanguine, as evidenced by the assessments of various authors (see, e.g., Button and Verhoef, 1998).

Intelligent transportation systems (ITS) are another technology that holds promise for alleviating congestion. ITS include: advanced traffic management systems, which optimize traffic signals and freeway ramp controls; advanced vehicle control systems, which allow closely spaced platoons of vehicles to operate at high speeds; and motorist information systems, which provide real-time

information and advice to individuals about travel conditions. ITS can help people to avoid heavily congested routes, to reschedule trips, and to choose between travel modes. But to the extent that ITS do succeed in improving travel conditions, they are likely to stimulate more travel because of latent demand. Congestion pricing may therefore be a complement to, rather than a substitute for, information technology. In any case, congestion and efforts to model and control it will endure for the foreseeable future.

References

Agnew, C.E. (1977) "The theory of congestion tolls", *Journal of Regional Science*, 17(3):381–393.
Arnott, R. and M. Kraus (1995) "Self-financing of congestible facilities in a growing economy", Department of Economics, Boston College.
Arnott, R. and K.A. Small (1994) "The economics of traffic congestion", *American Scientist*, 82:446–455.
Arnott, R., A. de Palma and R. Lindsey (1993) "A structural model of peak- period congestion: A traffic bottleneck with elastic demand", *American Economic Review*, 83(1):161–179.
Arnott, R., A. de Palma and R. Lindsey (1998) "Recent developments in the bottleneck model", in: K.J. Button and E.T. Verhoef, eds., *Road pricing, traffic congestion and the environment: Issues of efficiency and social feasibility*. Cheltenham: Edward Elgar.
Beckmann, M., C.B. McGuire and C.B. Winsten (1956) *Studies in the economics of transportation.* New Haven, CT: Yale University Press.
Braess, D. (1968) "Über ein Paradoxen des Verkehrsplanung", *Unternehmenforschung*, 12:258–268.
Button, K.J. (1992) *Transport economics*. Aldershot: Edward Elgar.
Button, K.J. and E.T. Verhoef, eds. (1998) *Road pricing, traffic congestion and the environment: Issues of efficiency and social feasibility*. Cheltenham: Edward Elgar.
Cassidy, M.J. and R.L. Bertini (1999) "Some traffic features at freeway bottlenecks", *Transportation Research B*, 33:25–42.
Chu, X. (1992) "Endogenous trip scheduling: A comparison of the Vickrey approach and the Henderson approach", *Journal of Urban Economics*, 37:324–343.
Chu, X. and K.A. Small (1997) "Hypercongestion", Department of Economics, University of California at Irvine, Irvine Economics Paper 97–98–12.
Dafermos, S.C. (1973) "Toll patterns for multiclass-user transportation networks", *Transportation Science*, 7:211–223.
Dafermos, S.C. and F.T. Sparrow (1971) "Optimal resource allocation and toll patterns in user-optimised transport networks", *Journal of Transport Economics and Policy*, 5(2):184–200.
Daganzo, C.F. (1997) *Fundamentals of transportation and traffic operations*. New York: Elsevier Science.
Daganzo, C.F. and Y. Sheffi (1977) "On stochastic models of traffic assignment", *Transportation Science*, 11:253–274.
Daganzo, C.F., M.J. Cassidy and R.L. Bertini (1999) "Possible explanations of phase transitions in highway traffic", *Transportation Research A*, 33:365–379.
Emmerink, R.H.M. (1998) *Information and pricing in road transportation*. Berlin: Springer-Verlag.
Gazis, D.C. and R. Herman (1992) "The moving and 'phantom' bottleneck", *Transportation Science*, 26:223–229.
Glazer, A. and E. Niskanen (1992) "Parking fees and congestion", *Regional Science and Urban Economics*, 22:123–132.
Haight, F.A. (1963) *Mathematical theories of traffic flow*. New York: Academic Press.
Hau, T.D. (1998) "Congestion pricing and road investment", in: K.J. Button and E.T. Verhoef, eds., *Road pricing, traffic congestion and the environment: Issues of efficiency and social feasibility*. Cheltenham: Edward Elgar.

Hazelton, M.L. (1998) "Some remarks on stochastic user equilibrium", *Transportation Research B*, 32:101–108.
Henderson, J.V. (1977) *Economic theory and the cities*. New York: Academic Press.
Hurdle, V. (1991) "Queuing theory applications", in: *Concise encyclopedia of traffic and transportation systems*. Oxford: Pergamon Press.
Kerner, B.S. and H. Rehborn (1997) "Experimental properties of phase transitions in traffic flow", *Physical Review Letters*, 79:4030–4033.
Lighthill, M.J. and G.B. Whitham (1955) "On kinematic waves. II: A theory of traffic flow on long crowded roads", *Proceedings of the Royal Society, London, Series A*, 229:317–345.
Mahmassani, H.S. and R. Herman (1984) "Dynamic user equilibrium departure time and route choice on idealised traffic arterials", *Transportation Science*, 18(4):362–384.
May, A.D. (1990) *Traffic flow fundamentals*. Englewood Cliffs, NJ: Prentice Hall.
May, A.D., S.P. Shepherd and J.J. Bates (1999) *Supply curves for urban road networks*. Leeds/Oxford: Institute for Transport Studies, University of Leeds/John Bates Services.
McDonald, J.F., E.L. d'Ouville and L.N. Liu (1999) *Economics of urban highway congestion and pricing. Ttransportation research, economics and policy*. Dordrecht: Kluwer.
Mohring, H. and M. Harwitz (1962) *Highway benefits*. Evanston, IL: Northwestern University Press.
Nagurney, A. (1999) *Network economics: A variational inequality approach*, revised 2nd edition. Dordrecht: Kluwer.
Newell, G.F. (1988) "Traffic flow for the morning commute", *Transportation Science*, 22(1):47–58.
Pigou, A.C. (1920) *Wealth and welfare*. London: Macmillan.
Ran, B. and D. Boyce (1996) *Modeling dynamic transportation networks: An intelligent transportation system oriented approach*, 2nd edition. Berlin: Springer-Verlag.
Richards, P.I. (1956) "Shock waves on the highway", *Operations Research*, 4:42–51.
Roess, R.P., W.R. McShane and E.S. Prassas (1998) *Traffic engineering*, 2nd edition. Upper Saddle River, NJ: Prentice Hall.
Small, K.A. (1982) "The scheduling of consumer activities: Work trips", *American Economic Review*, 72:467–479.
Small, K.A. (1992a) *Urban transportation economics. Fundamentals of pure and applied economics*. Chur: Harwood.
Small, K.A. (1992b) "Using the revenues from congestion pricing", *Transportation*, 19(4):359–381.
Transportation Research Board (1992) *Highway capacity manual*, 3rd edition. Washington, DC: National Academy Press, TRB Special Report 209.
Verhoef, E.T. (1999) "Time, speeds flows and densities in static models of road traffic congestion and congestion pricing", *Regional Science and Urban Economics*, 29:341–369.
Verhoef, E.T., P. Nijkamp and P. Rietveld (1996) "Second-best congestion pricing: The case of an untolled alternative", *Journal of Urban Economics*, 40(3):279–302.
Verhoef, E.T., J. Rouwendal and P. Rietveld (1999) "Congestion caused by speed differences", *Journal of Urban Economics*, 45:533–551.
Vickrey, W.S. (1969) "Congestion theory and transport investment", *American Economic Review (Papers and Proceedings)*, 59:251–260.
Walters, A.A. (1961) "The theory and measurement of private and social cost of highway congestion", *Econometrica*, 29(4):676–697.
Walters, A.A. (1987) "Congestion", in: *The new Palgrave: A dictionary of economics*, Vol. 1. New York: Macmillan.
Wardrop, J. (1952) "Some theoretical aspects of road traffic research", *Proceedings of the Institute of Civil Engineers*, 1(II):325–378.
Wie, B.-W. and R.L. Tobin (1998) "Dynamic congestion pricing models for general traffic networks", *Transportation Research B*, 32:313–327.