McGraw-Hill's HANDBOOK OF TRANSPORTATION ENGINEERING

Chapter 12. TRAFFIC CONGESTION

Authored by
Kara Kockelman, PhD CE, MCP
Professor of Transportation Engineering
The University of Texas at Austin
6.9 ECJ, Austin TX 78712
512-471-0210; FAX: 512-475-8744
kkockelm@mail.utexas.edu

The following is a pre-print and the final publication can be found as Chapter 12 in the *Handbook of Transportation Engineering*, McGraw Hill, January 2004.

TABLE OF CONTENTS

Introduction

Congestion is everywhere. It arises in human activities of all kinds, and its consequences are usually negative. Peak demands for goods and services often exceed the rate at which that demand can be met, creating delay. That delay can take the form of supermarket check-out lines, long waits for a table at a popular restaurant, and after-work crowds at the gym. Yet the context in which we most often hear of congestion posing a serious problem, to ourselves and to our economy, is the movement of people and goods.

The average American reports traveling 78 minutes a day, over 80 percent of which is by automobile.[1] (NHTS 2001) The Texas Transportation Institute (TTI) estimates that over 45 percent of peak-period travel or roughly one-third of total vehicle miles traveled occur under congested conditions in many U.S. metropolitan areas. (Shrank and Lomax 2002) These include the predictable places like Los Angeles, Washington DC, and Atlanta; they also include places like San Diego (California), Tacoma (Washington) and Charlotte (North Carolina). Though crime, education, taxes, and the economy certainly are key issues for voters and legislators, polls regularly report congestion to be the number one local issue. (See, e.g., Scheibal 2002, Knickerbocker 2000, and Fimrite 2002.)

Non-personal modes of transportation are certainly not immune to congestion, either. Intercity trucking carries almost 30 percent of freight ton-miles shipped in the U.S. every year (BTS 2002), and 72 percent of the value shipped. (CFS 1997) These trucks are subjected to the same roadway delays, resulting in higher priced goods, more idling emissions, and frayed nerves. The gates, runways, and traffic control systems of many popular airports are tested daily. And seaport berths, rail tracks, canals, and cables all have their limitations. As soon as demand exceeds supply, goods, people, and information must wait in queues that can become painfully long. Though not stuck in queues, others find themselves waiting at the destinations, for expected shipments, friends, family members, and colleagues that fail to arrive on time.

Engineers, economists, operations researchers, and others have considered the problem of congestion for many years. The confluence of growing traveler frustration, technological innovations, and inspirational traffic management policies from around the world provide added momentum for the modifications needed to moderate and, ideally, eliminate this recurring problem and loss. This chapter examines congestion's defining characteristics, its consequences, and possible solutions.

Defining Congestion

Notably, congestion is not always undesirable. Some "congested" experiences can be positive, and these tend to occur at one's destination – rather than *en route*. Myers and Dale (1992) point out that orchestrated congestion in public spaces, such as theatre entry plazas, enhances public interaction, enables better land use mixing, and calms vehicular traffic. Taylor (2002) observes

[1] NPTS-reported trip-making involves spatially very distinct locations, such as home and work, school and shopping center. In reality, we are moving much more regularly, between bedrooms, around offices, and along supermarket aisles. Such travel, while substantial, is probably much less impacted by congestion.

that congested city centers are often signs of vibrant city activity and prosperity. These forms of congestion are to some extent desirable and are not the concern of this chapter.

This chapter stresses instead the undesirable form of congestion: the kind that impedes travel between two points, effectively adding access costs to a desired destination. The travel itself is not enjoyable; it is instead a necessary expenditure.[2] This form of congestion is a slowing of service. Queues (or lines) of travelers will form if demand exceeds capacity. But these are not necessary for congestion to occur. All that is necessary is that the service speed is less than the "free-flow" or maximum speed, which exists when demand is light, relative to capacity.

All transportation systems are limited by a capacity service rate. Operators at manual toll booths and transponder readers for electronic toll collection (ETC) cannot reliably serve more than a certain number of vehicles per hour. Commuter rail lines eventually fill up, along with their train cars. Port cranes exhibit functional limits, as do canals, runways, and pipelines. No system is immune; all physical pathways are constrained in some respect.

When systems slow down, delays arise. Delay generally is defined as the difference between actual travel time and travel time under uncongested or other acceptable conditions. The *Highway Capacity Manual 2000* (TRB 2000) defines signalized intersection delay as the sum of delay under uniform arrivals (adjusted for a progression factor), incremental delays (to account for randomness in arrival patterns), and any initial queue delays (to recognize spillovers from prior cycles). While vehicle detection and intersection automation can dramatically reduce signal delays, they cannot eliminate them. Any time two or more vehicles (users) wish to use the same space at the same time, delay – else a crash – will result. One must cede that space to the other. The mechanism may be a signal, a queue, pricing, rationing or other policies.

In general, then, congestion is the presence of delays along a physical pathway due to the presence of other users. Before discussing strategies to combat such delays, this chapter examines the general costs and consequences of congestion, its causes, and its quantification.

The Consequences of Congestion

Automobile congestion has myriad impacts, from wasted fuel and added emissions to frayed nerves, more expensive goods, and elevated crash rates. Its clearest impact is delay, or lost time. Across the U.S. this may average 20 hours per year per person. In Los Angeles, it is estimated to exceed 60 hours, which translates to 10 minutes a day or one-sixth of one's average travel time. (Schrank and Lomax 2002) Presented this way, even Los Angeles' numbers may seem acceptable. Why then does this issue so consistently top opinion polls as our communities' number one policy issue? There are many reasons. One is that dense car traffic is more difficult to navigate, even if speeds stay high. Such travel is unpleasant and tiring in many ways. Another is that congestion tends to be unpredictable, even when it is recurring. As a result, peak-period travelers – including trucks and buses – regularly arrive early or late at their destinations,

[2] Mokhtarian and colleagues have been examining the extent to which travel may be desirable in and of itself, resulting in "excess travel." (Mokhtarian and Salomon 2001, Redmond and Mokhtarian 2001, Salomon and Mokhtarian 1998) Richardson (2003) found reasonably high proportions of travelers in Singapore to exhibit a zero value of travel time, causing him to conclude that, among other things, travel on the air-conditioned transit system in that humid city can be a relatively enjoyable (or at least refreshing) experience.

creating frustration. These suboptimal arrival times carry a cost: missed meetings and deliveries, loss of sleep, child care fees for late pick up, children waiting around for classes to begin, a supervisor's growing intolerance of missed work. Researchers have tried to quantify these additional costs.

Bates et al.'s (2001) review of travel time reliability research finds that every minute of *lateness* is regularly valued at two to five times a minute of travel time. And every minute *early* is valued at almost a minute of travel time (around 80 percent). In general, variation in travel time (as measured by standard deviation) is worth more – to the typical traveler – than the average travel time. So, even if one's commute trip *averages* 25 minutes, if it exhibits a 10 minute standard deviation[3], it typically is not preferred to a guaranteed travel time of 35 minutes. Using loop detector data for samples of freeway sections in cities across the United States, Lomax et al. (2003) estimated early-departure "buffer times", in order to ensure on-time (or early) arrivals in 95% of one's tripmaking. For Austin, Texas, the required buffer was estimated to be 24% of the average travel time; in Los Angeles, it was 44%. In Lomax et al's 21-city data set, congestion is highly correlated with high buffer times (and thus low reliability). Clearly, congestion is costly – in a variety of ways.

Another reason for society's impatience with congestion may relate to equity and an ability to prioritize consumption or purchases. On a "free" system of public roadways, every trip is treated the same. In other words, more valuable trips experience the same travel times as less valued trips. Persons and goods with very low values of time pay the same time-price as others, even though the monetary value of and/or willingness to pay for their trips can differ by orders of magnitude. Few options are available, and they can require significant adjustments: changes in one's home, work, school, or other locations; moved meeting times; and entirely forgone activities. In one exceptional circumstance, Silicon Valley pioneer Steve Jobs elected to purchase a helicopter to reduce his San Francisco Bay Area commute.

As mentioned earlier, some travelers find a moderate level of congestion acceptable, even desirable. Moreover, an evolution of in-vehicle amenities, from radios, air conditioning, and reclining seats, to tape and CD players, stereo-quality sound systems, cellular phones, video players, and heated seats, plays a role in reducing the perceived costs of congestion. At the same time, increasing presence of two-worker couples, rising incomes, just-in-time manufacturing and delivery processes, and complication of activity patterns for adults and children have resulted in a variety of travel needs and time constraints that can make delays more costly and stressful. The market for scarce roadspace breaks down during peak travel hours in many places. So what are the costs?

The TTI studies estimate Los Angeles' congestion costs to exceed $14 billion each year, or more than $1000 per resident. This figure is based on speed and flow estimates across the region's network of roads, where every hour of passenger-vehicle delay is valued at $12.85, every mile of

[3] The standard deviation is the square root of the expected value of squared differences between actual and mean travel times: $\sigma_t = \sqrt{E\left((t-\mu_t)^2\right)}$. For a normal or Gaussian distribution of travel times that averages 25 minutes, a standard deviation of 10 minutes implies a 16 percent probability that a trip will exceed 35 minutes.

truck travel at $2.95, and each gallon of gasoline consumed (while delayed) at $1.39[4]. (Schrank and Lomax 2002) Together with costs from 74 other major U.S. regions (using the same unit-cost assumptions), the annual total reaches $70 billion. This is about 4¢ per vehicle mile traveled, or double U.S. gas taxes (which total roughly 40¢/gallon, or 2¢/mile). Essentially, the U.S. Highway Trust Fund could be trebled through the addition of these estimated costs. Of course, these neglect congestion in other travel modes and other U.S. locations, as well as environmental impacts, schedule delay, delivery difficulties, inventory effects, frustration, crash and other costs. Carbon monoxide and hydrocarbon emissions are roughly proportional to vehicle *hours* of travel (Dahlgren 1994); oxides of nitrogen also rise with slowed traffic, further worsening air quality. (Beamon 1995) Congestion stymies supply chains and diminishes agglomeration economies. (Weisbrod et al. 2001) Taken all together, such costs may rival or even exceed the TTI 75-city cost estimates.

Congestion and Crashes

The consequences of congestion for crash frequency and severity are intriguing. The lowest crash rates (crashes per vehicle mile traveled per lane) tend to occur at intermediate levels of flow (for example, level of service C[5]). (See, e.g., Gwynn 1967 and Brodsky and Hakkert 1983.) Controlling for traffic density – rather than flow – also is key, since low flows can occur under both uncongested (high speed) and congested (low speed) conditions. Garber and Subramanyan's recent work (2002) for weekday crashes on four highways indicates a steep reduction in police-reported crash rates (i.e., crashes per vehicle mile traveled) when densities are about half of critical density (where critical density corresponds to capacity flow rates)[6]. Thus, crash rates generally appear to rise as congestion sets in. However, speeds tend to fall under such conditions, especially when demand exceeds capacity. And lower collision speeds mean less severe traffic crashes. (See, e.g., Evans 1991; Kockelman and Kweon 2002.)

Of course, crashes themselves generate congestion, by distracting other drivers and blocking lanes and shoulders. And drivers frustrated by congestion may take risks that offset some of the benefits of reduced speeds, such as tailgating, using shoulders as traffic lanes, cutting across dense on-coming traffic, and speeding up excessively when permitted. Such behaviors are typical of "road rage", and may worsen congestion – and result in crashes. Little is formally known regarding the magnitude and nature of such indirect safety effects of congestion.

Quantifying Congestion

The explicit measurement or quantification of congestion has many uses. Such measures help communities identify and anticipate traffic problems, by location, severity, and time of day. Their magnitude and ranking provides a basis for targeted investment and/or policy decisions. They also as useful inputs for air quality models, which require travel speed and distance

[4] Passenger-vehicle occupancies were assumed to be 1.25 persons/vehicle. 5 percent of congested-period VMT was assumed to be by trucks. Truck time delays were multiplied by congested speeds and $2.95 per mile traveled, in order to provide truck delay costs, which are on the order of $100 per hour.

[5] Freeway level of service C implies conditions where speeds are near free-flow speeds but maneuverability is "noticeably restricted". (*Highway Capacity Manual 2000* [TRB 2000], p. 13-10)

[6] Many slight crashes may occur under congested conditions, and go unreported to police. Thus, it is possible that total (reported and unreported) crash rates stay stable or even rise under congested conditions.

information. The following discussion provides a definition of congestion, along with methods for its estimation based on travel time formulae.

Congestion and The Highway Capacity Manual

For transmission of people, goods, or data, an important distinction exists between capacity (i.e., maximum-flow) speeds and free-flow speeds. In the case of roadways, congestion can set in and speeds can fall well before capacity is reached. Chapter 23 of the *Highway Capacity Manual 2000* (TRB 2000) provides estimates of capacity and speeds for a variety of basic freeway segments. A traffic density of 45 passenger cars per lane-mile (pc/mi/ln) corresponds to (sustainable) capacity conditions, and it distinguishes levels of service E and F. On facilities with a 75 mi/h free-flow-speed (FFS), average speeds are expected to begin falling at relatively low densities (e.g., 18 pc/mi/ln). And at capacity conditions (2400 pc/h/ln and 45 pc/mi/ln), the predicted prevailing speed is just 53.3 mi/h – well below the uncongested 75 mi/h FFS. On freeways exhibiting lower free-flow speeds, congestion is expected to set in later and speed reductions are less severe. For example, for FFS = 55 mi/h, small drops in prevailing speed arise at 30 pc/mi/ln densities, and capacity speeds are 50 mi/h (just 10 percent less than FFS). Travel on these lower speed facilities will take longer, however, even if their conditions do not qualify as "congested" – simply because their associated speeds are lower, for every level of service.

Why do speeds fall as density increases? Because the smaller the spacing between vehicles, the more likely a conflict. Safety also sets an upper limit on how fast people want to travel. Human and vehicle response times are limited: we take time to perceive threats, and our vehicles take time to slow down. Thus, drivers have maximum speed preferences, which govern when traffic is relatively light, and have minimum spacing preferences, which govern when traffic is relatively heavy[7]. Driver spacing requirements go up with speed, so there is a natural limitation on how many can traverse a given road section in any given time. The *Highway Capacity Manual* predictions for freeway lanes are just one illustration of these safety-response phenomena.

On uncontrolled (non-freeway) multilane highways, *Highway Capacity Manual* estimates of speeds and capacity flows are lower than those found on freeway lanes. (Driveways and other access points, left turns in the face of on-coming traffic, and other permitted behaviors necessitate more cautious driving.) However, the density values defining levels of service are almost unchanged, and the magnitudes of speed reduction leading to capacity conditions are minor (on the order of 1 to 3 mi/h).

On two-lane (undivided) highways, capacity flows are dramatically less (1700 pc/h/ln) and levels of service are defined by the percentage of time that vehicles follow slower vehicles (PTSF) and, in the case of Class I facilities, by average travel speed. If passing is not permitted along a section, average travel speeds are predicted to fall by as much as 4.5 mi/h.

When travel is controlled by traffic signals, signal-related delay estimates define level of service. Speed calculations are not emphasized, though simulation software such as the FHWA's

[7] Based on the traffic observations illustrated in Figures 1 and 2 of this chapter, Kockelman (2001) has estimated free-flow speed and spacing preferences for various freeway user classes.

CORSIM (1999) can generate estimates of trip start times, end times, and distances, thereby predicting operating speeds.

Significantly, *Highway Capacity Manual* methods permit no traffic speed estimation beyond level of service E, and Chapter 23's speed-flow curves and tables disappear. Beyond level of service E, traffic conditions can be characterized by speeds as high as 50 mi/h (on high-design freeways) – or by complete gridlock (0 mi/h). When demand exceeds capacity, a queue develops and straightforward speed models break down. At that point, it becomes more important to know when a traveler enters the queue than to know how many are entering it. And it is all characterized as level of service F.

Unfortunately, level of service F's oversaturated conditions are common in many regions, and on many facilities. Local bottlenecks and incidents cause demand to exceed capacity, sending congestive shockwaves back upstream. Under these conditions, upstream speeds can fall well below those prevailing under capacity conditions, even to zero. While oversaturated traffic conditions are rather unstable, and exhibit high variation, speeds can be approximated. The following two sections describe some applicable methods.

Roadway Conditions: The Case of Interstate Highway 880

The *Highway Capacity Manual* offers traffic predictions based on empirical evidence for a variety of roadway types, designs and locations. Yet conditions on specific facilities can differ rather dramatically. Loop detectors embedded in highway pavements offer continuous data-collection opportunities for key traffic variables.

Based on detector data from Interstate Highway (I.H.) 880's number-two lane, observed speeds and densities are plotted as Figure 1. Speeds (measured in mi/h) fall as density (measured in vehicles per lane mile [veh/ln/mi]) increases. Density is inversely related to vehicle spacing[8]; and, as vehicles are added to a section of roadway, density rises and spacings fall. Drivers are inhibited by reduced spacings and the growing presence of others. For reasons of safety, speed choice, and reduced maneuverability, drivers choose lower speeds.

Since speed multiplied by density is flow[9], so Figure 1's information leads directly to Figure 2's speed versus flow values. Flow rates over these detectors may reach 3,000 vehicles per hour – or 1.2 seconds per vehicle, for a brief, 30-second interval. And average speeds appear to fall slightly throughout the range of flows, from roughly 65 to 55 mi/h. Travel times are impacted as more and more vehicles enter the lane, densifying the traffic stream. At some point downstream of this detector station, demand exceeds capacity or an incident destabilizes traffic and shockwaves travel back upstream, forcing traffic into level-of-service F conditions. These oversaturated conditions correspond to speeds below 50 mi/h in this lane on this facility.

[8] Vehicles per unit distance (density) equal the inverse of distance per unit vehicle (spacing), where spacing is measured from the front of one vehicle to the front of the following vehicle (thereby including the vehicle bodies).
[9] Flow is vehicles per unit time, speed is distance per unit time, and density is vehicles per unit distance. Under stationary traffic conditions, vehicles maintain constant speeds, the mix of these vehicles (and their speeds) is unchanging, and density times space-mean speed (rather than the commonly measured time-mean speed or average of spot speeds) equals flow.

Figure 3 is a photograph of congested traffic that could have come from this same section of I.H. 880. Since passenger vehicles average 15 to 18 feet in length (Ward's 1999), the image suggests an average spacing of roughly 40 feet per vehicle (per lane). Such spacing translates to a density of 132 vehicles per lane-mile. Assuming that Figure 1's relationships are predictive of Figure 3's traffic behaviors, this density corresponds to average speeds between 5 and 15 mi/h. Since speed times density is flow, flows are likely to be around 1,300 veh/h/ln, or about half of capacity.

As illustrated by Figure 2, a flow of just 1,300 veh/h/ln could also correspond to a much higher speed (about 60 mi/h) – and a much lower density (perhaps 25 veh/ln/mi – or a 211 ft/veh spacing). This contrast of two speeds (and two densities) for the same level of output (i.e., flow) is disturbing. Densities beyond critical (capacity-level) density and speeds below critical speed identify a loss: the restricted roadspace could be more fully utilized and traveler delays could be avoided, if only demand and supply were harmonized.

This "Tragedy of the Commons" (Hardin 1968) plays itself out regularly in our networks: heavy demand, downstream bottlenecks and capacity-reducing incidents force upstream travelers into slow speeds. And facilities carry lower than capacity flows. In extreme cases, high-speed freeways as well as downtown networks become exasperating parking lots. To optimally avoid underutilization of scarce resources, one must have a strong understanding of demand versus supply. And bottlenecks are to be avoided, if the benefits can be shown to exceed the costs.

A chain is only as strong as its weakest link, and downstream restrictions can have dramatic impacts on upstream roadway utilization. Bridges are common bottlenecks; they are relatively costly to build[10] and expand and thus often are constructed with fewer, narrower lanes and limited shoulders. Given a tradeoff in construction costs, travelers' willingness to pay, and traveler delays, theoretically, there is an optimal number of lanes to build in any section of roadway. Reliable information on demand (or willingness to pay) and associated link travel times (for estimates of delay) is necessary. Real-time roadway pricing, robust models of travel demand, and instrumentation of highways (for traffic detection) are key tools for communities aiming to make optimal investment – and pricing – decisions. This next section describes methods for estimating delays.

Link Performance Functions and Delay Estimation

Delay and the social value of this delay (such as the willingness of travelers to pay to avoid such delay) depend on the interplay of demand, supply, and willingness to pay. One must quantify congestion through link and network performance functions, and transform the resulting delays into dollars.

Travel time (per unit distance) is the inverse of speed. Thus, delays rise as speeds fall. And, as described earlier, speeds fall as density rises. Density rises as more and more users compete for limited roadspace, entering the facility and reducing inter-vehicle spacings, causing speeds to fall. Fortunately, densities can rise fast enough to offset reduced speeds, so that their product (i.e., flow) rises. But flow can increase only so far: in general, it cannot exceed capacity. As

[10] Based on freeway construction-cost data from the Texas Department of Transportation, Kockelman et al. (2001) estimated bridge lanes to cost five to ten times as much as regular lanes.

soon as demand for travel across a section of roadway exceeds capacity, flow at the section "exit" will equal capacity, and a queue will form upstream, causing average travel times across the congested section to rise. Unfortunately, these travel times tend to rise exponentially, as a function of demand for the scarce roadspace. Thus, only moderate additions to demand can dramatically impact travel times. And if the resulting queuing impedes other system links, the delay impacts can be even more severe.

Referred to as the "BPR (Bureau of Public Roads) Formula" (FHWA 1979), the following is a common travel-time assumption:

$$t(V) = t_f\left(1 + .15\left(\frac{V}{C}\right)^4\right) \quad (1)$$

where $t(V)$ is actual travel time, as a function of demand volume V, t_f is free-flow travel time, and C is "practical capacity", corresponding to approximately 80 percent of true capacity.[11] Figure 4 illustrates the BPR relationship for a particular example where true capacity is 10,000 veh/h, C is 8,000 veh/h, and t_f is 1 minute (for example, the time to traverse a one-mile section at a free-flow speed 60 mi/h). If demand exceeds capacity by 30 percent ($V = 1.3C$), travel times will be twice as high as those under free-flow/uncongested conditions. The resulting 2 min/mi pace implies speeds of just 30 mi/h. It also implies one minute of delay for every mile of travel, with delay naturally defined as follows:

$$delay(V) = t_f - t(V) = .15t_f\left(\frac{V}{C}\right)^4 \quad (2)$$

Is a one-to-one correspondence of delay to free-flow travel time common? Shrank and Lomax (2002) estimate 54 sec of delay for every minute of peak-hour travel in the Los Angeles region. Kockelman and Kalmanje's (2003) survey results indicate that Austin, Texas commuters perceive almost 60 sec of delay for every 60 sec of their commute travel. Thus, it may be in some regions that peak-period delays regularly exceed capacity by 30 percent or more.

Of course, this 30-percent figure for a doubling of travel times is based on the BPR formula. Researchers have proposed modifications to this formula. Horowitz (1991) suggested replacing the two constants (0.15 and 4) with 0.88 and 9.8 (for use with 70 mi/h-design speed freeways) and 0.56 and 3.6 (for 50 mi/h freeways). Dowling et al. (1997) recommended 0.05 and 10 (for freeways) and 0.20 and 10 (for arterials). For comparison purposes, Dowling et al.'s freeway formula is included in Figure 4, as the "Modified BPR" curve. The two differ dramatically when demand exceeds capacity by more than 50 percent.

Beyond basic modifications to the BPR formula, there are other options. Akçelik's (1991) formula is wholly distinct, and recognizes demand duration. The duration of queuing has important consequences for total queue lengths and thus overall delays – which depend on when one can expect to enter the queue.

[11] True capacity is understood to be the maximum service flow (MSF) under the *Highway Capacity Manual*'s (TRB 2000) level of service E. This practical capacity variable is a source of regular error and confusion in applications. Many (e.g., Garber and Hoel 2001) substitute a roadway's true capacity flow rate for C, resulting in an underprediction of travel times.

Modifications in BPR factors and the underlying formulae can have dramatic impacts on travel time estimates, travel demand predictions and policy implications.[12] Yet actual delay relationships remain poorly understood. Vehicles occupy space and roadway sections back up, spilling over onto other links in the network. The complexity of networks makes it difficult to measure travel times[13] – and even more difficult to ascertain "demand."

Delay Example: A Temporary Lane Loss

Relying again on Figure 4 and Equation 1's BPR formula, consider the impact of a loss of one lane. If capacity of 10,000 veh/h corresponds to a four-lane high-design freeway, then the loss of one of these four lanes (by a crash or creation of a construction work zone, for example) results in an effective capacity of 7,500 veh/h. At a demand of 13,000 veh/h, travel times will jump by 115 percent, from 2.0 min/mi to 4.3 min/mi. This is now 330 percent longer than travel time under free-flow conditions. Under this dramatic situation, speeds would be just 14 mi/h – far less than the free-flow speed of 60 mi/h and well below the four-lane speed of 30 mi/h.

The reason that travel times rise so dramatically once demand exceeds capacity is that a roadway (like an airport or any other constrained facility) can accommodate no more flow. It behaves much like a funnel or pipe that can release only so much fluid per unit of time. Any additional users will be forced to form a slow-moving queue, backing up and impacting the rest of the system (by blocking off-ramps and on-ramps, or driveways and cross-roads, upstream of the limiting section). This is a classic bottleneck situation, where demand exceeds supply. Capacity-reducing incidents can affect supply instantly, leading to essentially the same low-speed, high-delay conditions for which recurring bottlenecks are responsible. Unfortunately, there is no guarantee that congestion can be altogether avoided; supply disruptions, through incidents and the like, can occur at most any time.

Recurring and Non-recurring Congestion

The above example of a temporary lane loss may be recurring or non-recurring, predictable or unpredictable. Recurring congestion arises regularly, at approximately the same time of day and in the same location. It results from demand exceeding supply at a system bottleneck, such as a bridge, tunnel, construction site, or traffic signal. Non-recurring congestion results from unexpected, unpredictable incidents. These may be crashes, jack-knifed trucks, packs of slow drivers, or foggy conditions.

Schrank and Lomax's (2002) extensive studies of regional data sets on travel, capacity, and speeds suggest that non-recurring incidents account for roughly half of total delay across major U.S. regions. However, these percentages do vary. They depend on the levels of demand and supply, crash frequency, and incident response. For example, in the regularly congested San

[12] Nakamura and Kockelman's (2002) welfare estimates for selective pricing on the San Francisco Bay Bridge were very dependent on the Bridge's travel time function. Outputs of Krishnamurthy and Kockelman's (2003) integrated land use-transportation models of Austin, Texas were most affected by the exponential term in the BPR formula.
[13] Lomax et al. (1997) recommend the use of probe vehicles to ascertain operating speeds. Loop detectors are presently only popular on freeway lanes and can assess only local conditions; frequent placement of loops is necessary to appreciate the extent of upstream queuing. Video cameras and sophisticated image processing techniques offer hope for future traffic data collection and robust travel-time estimation.

Francisco Bay Area, with its roving Freeway Service Patrols, incidents account for 48 percent of total delay. In the New York-Eastern New Jersey region, this estimate rises to 66 percent.

Evaluating the Marginal Costs of Travel

Whether a traveler opts to enter a facility that is congested for recurring or non-recurring reasons, he or she pays a price (in travel time, schedule delay, and other costs) to use that facility. Because travel times rise with demand, his/her entry onto the facility (or at the back of the queue) also marginally increases the travel cost for others entering at the same time or just behind. This imposition of a cost, to be borne by others, is called a *negative externality*. Essentially, use of a congestible facility reduces the quality of service for others. This reduction in service quality is an external cost in the form of travel time penalties that others bear. Of course, all users bear it equally. So is it a problem?

Any time users "pay" a cost lower than society bears to permit the added consumption (in this case, use of a space-restricted facility), the good is over-consumed and society bears more cost than it should. Economists have rigorously shown that in most any market, goods should not be allocated beyond the point where marginal gain (or value to society) equals marginal cost to furnish the good. Marginal gains for most goods are well-specified by consumers' willingness to pay. And marginal costs are typically absorbed by suppliers of those goods. In the case of road use, costs arise in many forms – and they are absorbed by many parties: infrastructure provision and maintenance costs are absorbed by federal, state, and local agencies (and passed on through fuel and property taxes); travel time costs are absorbed by travelers; environmental damages are absorbed by humans and animals (on, off, and far from the facilities themselves); and crash losses are felt by a variety of individuals (through pain, suffering, delay, and EMS-related taxes).

The focus of this chapter is congestion, and therefore travel times in excess of free-flow travel times. A road's available space is fixed, and under congested conditions fellow travelers absorb the costs of delays arising from additional users. What are these marginal costs? Every link-performance function, $t(V)$, implies these.

At a particular level of demand *V*, the marginal cost of an additional user, $MC(V)$, is the change in total travel costs due to that added user. Total costs $TC(V)$ are average cost per user, $AC(V)$, multiplied by the number of users, *V*. And travel time (per user), $t(V)$, is the average cost. Using this logic, the standard BPR travel time function, and a little calculus (for continuous differentiation of the total-time formula), one has the following results:

$$AC(V) = t(V) = t_f\left(1 + .15\left(\frac{V}{C}\right)^4\right)$$

$$TC(V) = V \cdot t(V) \qquad (3)$$

$$MC(V) = TC(V+1) - TC(V) \approx \frac{\partial TC}{\partial V} = AC(V) + 0.6t_f\left(\frac{V}{C}\right)^4$$

This last of three equations, the marginal cost of additional users, clearly includes the cost that the additional user experiences directly, $AC(V)$. But it also includes a second term: the unpaid

cost or negative externality that others endure in the form of higher travel times. In this BPR-based example, the externality also depends on the fourth power of the demand-to-practical capacity ratio; but it is amplified by a factor of 0.6, rather than 0.15. Thus, at certain levels of demand *V*, this second external cost will dominate the first average-cost term. It is this unpaid cost that is responsible for over-consumption of roadspace. Without assignment of ownership of the roads or some other method to ensure optimal use, excessive congestion results. If the roads are in heavy demand, the cost is more severe and the loss to society particularly striking.

Figure 5 plots the marginal cost curve above the standard BPR average cost curve. The difference between these two is that which goes unpaid by the additional users. At capacity levels of demand, this *difference* exceeds average cost by 7 percent. At demand of 13,000 veh/h, it constitutes more than double the average cost. Essentially then, for optimal operations, perceived travel times or costs *should be* equivalent to 6.23 min/mi. Yet they are only 2.04 min/mi; the added travelers are enjoying an implicit subsidy of 4.18 min/mi at the expense of other travelers.

This situation may lead one to question: Who is the last, "marginal" driver and why (and how) should only he or she be penalized, when everyone should enjoy equal rights of access to the public right of way? The answer is that *all* drivers should weigh the true marginal cost of their trip before embarking – and then pay this cost (in the form of a toll) if their marginal gain from making the trip exceeds the total cost of time and toll. This requirement is placed on buyers of any good in regular markets, even for items as basic as clothing, shelter, food, and health care[14]. Private providers are not asked to provide goods below their marginal cost. Such excessive production is unwise. In considering whether and how to combat congestion, we should ask whether society should provide space on roadways at prices below marginal cost. This question brings us to this chapter's final section.

Solutions to Congestion

Congestion may result from inadequate supply, imperfect information, flawed policies or any combination of these three. To address these possibilities, a variety of remedies have been proposed. Solutions may be supply- or demand-sided, long- or short-term, best suited for recurring or non-recurring congestion. They may be demographically, temporally and spatially extensive – or limited; they may be costly or inexpensive, mode-specific or multimodal. They may involve mode subsidies, special lanes and/or pricing policies. Virtually all produce winners and losers.

In discussing various strategies for combating congestion, this section first details supply-side remedies, which generally aim to increase capacity, by adding facilities or enhancing operations of existing facilities. On the demand side, the emphasis is a moderation or modification of prices and preferences.

[14] Education remains largely a public good. Regardless, subsidies targeting specific goods and consumer groups always can be provided (e.g., food stamps for low-income families).

Supply-side Solutions

Traditionally, the solutions for congestion have been supply sided. Engineers expected that roadway expansions and upgrades would relieve congestion. And, as long as the latent demand for the facilities did not overtake the expansions (through, for example, Downs' [1962] "triple-convergence principle" of route, time of day, and mode choice adjustments), peak-hour speeds rose and travel times fell.

The same phenomenon holds true at airports, where gates and runways may be added and aircraft may be made larger (to add seating). It holds true on railways, where engines and track may be added and headways reduced (through added trains and track as well as better coordination of train schedules); and at shipyards, where container cranes may be added or upgraded and berths may be extended. Traffic actuation and synchronization of signals, rationalization of freight networks, targeted enhancements of bottlenecks, headway-reducing vehicle-guidance technologies, and other remedies also are very helpful, for specific applications.

Points of recurring congestion generally are often well-suited for supply-side solutions. Returns on such investment will be more certain, as long as similarly sized, nearby bottlenecks do not negate the local expansion. (For example, a previously untested and thus undetected chokepoint may lie just downstream and lead to similar backups.) In cases of non-recurring congestion, solutions providing rapid detection (e.g., paired loop detectors, video processing, or transponder tag re-identification), rapid response (e.g., roaming freeway service patrols), and real-time information provision (e.g., variable message signs) will have the greatest impact.

Capacity Expansion

There are a great many ways to expand the capacity of congestible systems. As long as increases in demand – through natural growth, changing preferences, and substitution (of origins, destinations, routes, modes, departure times, and other choices) – do not overtake these expansions, service times will fall. Yet, in many markets, capacity is so constrained and pent-up demand so significant, that travel times and speeds on expanded sections of the network may not fall in any perceptible way. Investigations by Hansen and Huang (1997), Noland and Cowart (1999), Fulton et al. (2000), and Rodier et al. (2000) have resulted in high long-run elasticity predictions of demand (vehicle miles traveled) for roadspace (generally after controlling for population growth and income). Elasticity estimates of almost 1.0 (suggesting that new roadspace is almost precisely filled by new miles traveled) are not uncommon. However, several of these studies draw largely on California data, where congestion is relatively severe. In less congested regions, elasticities are expected to be lower. Even so, at roughly $1 million per added lane mile for freeway construction costs alone[15] (Kockelman et al. 2001, Klein 2001), funding constraints regularly preclude major supply-side solutions. And in regions not in attainment with air quality standards and/or wishing to limit sprawl and other features of long-distance driving, building one's way out of congestion may not be a viable option.[16]

[15] Right-of-way acquisition, traffic diversion delays during construction, and other features of major highway projects in congested areas will add further to the overall expense of such projects.

[16] Even though VMT may rise in proportion to expanded capacity, with congestion remaining high, there may be sufficient benefits accruing to warrant such expansions. For example, if households can afford better homes and enjoy more choice in stores, schools, jobs and other activities, thanks to expanded travel options, those benefits should be recognized.

Alternative Modes and Land Use

To reduce roadway congestion, there are several supply-side enhancements of *alternative* modes that can cost less than new roadways while reducing driving and emissions. Improvements and expansions of bus, rail, ferry and other services may qualify. However, transit already is heavily subsidized, per trip, in many countries. And even in downtown locations where its provision is reasonably extensive, transit ridership rates remain low in the United States. Thus, it is unlikely to attract many travelers, particularly for long trips in a U.S. context.[17]

Land use solutions also have been proposed, as a way to increase the use of alternative modes and diminish congestion. Transit use and walking are highest in high-density, mixed-used areas. (See, e.g., Pushkarev and Zupan 1977; Kockelman 1997; Cervero and Kockelman 1997.) Transit- and walking-oriented New Urbanist designs strive to motivate mode shifts, and reduce automobile reliance. But the resulting mode shifts are relatively weak, and neighborhood design – particularly in the form of higher development densities – is a poor instrument for combating roadway congestion. (See, e.g., Boarnet and Crane 2001; Taylor 2002.)

Managed Lanes

Addition of managed lanes, both high-occupancy vehicle (HOV) and high-occupancy toll (HOT) lanes are an intermediate option. HOT lanes help insure against congestion for those whose trips are highly valued while facilitating full utilization of these special lanes. (Peirce 2003) Fees can rise and fall (for example, up to 40 cts/mile) to keep the HOT lanes flowing smoothly, while carpoolers (HOV users) and transit buses ride free – and fast. Thanks to revenues generated, agencies can float bonds to help cover some of the construction and other costs, or spend the money on other services (such as increased transit service, roving freeway service patrols, and variable message signs with information on traffic conditions). (Dahlgren 2002) However, without pricing of substitute routes and services, it is difficult if not impossible to raise sufficient revenues from the private sector. Public financing is still needed. And HOV/HOT-lane construction costs generally exceed those of standard freeway lanes, due to distinguishing features (such as longitudinal barriers and special access points).

Ramp Metering

Another supply-side strategy is ramp metering (May 1964, Newman et al. 1969). In contrast to expansion of existing systems and services, the objective is *reduction* of ramp flows, to keep main freeway lanes moving safely and swiftly. This form of supply restriction can reduce travel times and improve safety (Chen et al. 2001, Klein 2001) but also penalize near-destination dwellers in favor of long-distance drivers, resulting in certain inequities. (Levinson et al. 2002) Ramp metering aims to moderate the use of key links in a system, thus impacting route choice and demand, the subject of the next section.

Demand-side Solutions

Demand-side strategies seek to impact demand directly, through policies and prices. Rather than expanding (or shrinking) existing services and facilities, one targets the relative prices of and/or access to these.

[17] There are a variety of reasons for this. Land use patterns (including dispersed, low-density origins and destinations), parking provision, and relatively low gas prices are just a few.

Parking policies offer valuable examples of demand-side strategies. Most parking is provided "free", at offices, shopping centers, schools, and elsewhere. When space is plentiful, attendants are not needed, operations costs are zero and maintenance costs may be minimal. Largely for the sake of cost-collection efficiency, parking costs are borne indirectly by users, through, for example, reduced salaries (to employees) and higher goods prices (for shoppers). Everyone bears these indirect prices, however, so there is no price-based incentive for not parking. Preferential parking and other perks for carpoolers and others who reduce total driving and parking demands provide a way to impact driving demand; however, there is a cost to these policies. Shoup's (1992, 1994) cash-out policy, now in place in California[18], requires that the cash-equivalent of parking expenses be given to those employees who do not use the parking. This form of clear remuneration makes good sense to those who agree that markets naturally clear at optimal levels when pricing and other signals are unambiguous and consistent. This argument raises the case for congestion pricing.

Congestion Pricing

The objective of congestion pricing is efficient travel choices. It is a market-based policy where selfish pursuit of individual objectives results in maximization of net social benefits. Such laissez-faire capitalism is the guiding light behind Adam Smith's (1776) Invisible Hand. When market imperfections are removed (through pricing of negative externalities [e.g., noise, emissions, and congestion], subsidy of positive externalities [e.g., public schools for educating all community members], and provision of adequate information), private pursuit of goods is optimal[19].

In 1952 Nobel Laureate William Vickrey proposed congestion pricing for the New York City Subway, through the imposition of higher fares during congested times of day. This proposal was followed by theoretical support by Walters (1954, 1961) and Beckmann, McGuire, and Winsten (1956). If peak-period pricing recognizes the negative externality (i.e., time penalty) equivalent of each new, marginal user of a facility, optimal consumption and use decisions can result. In the case of Figure 5's standard-BPR curve for travel time, the value of the negative externality, or the optimal toll, is the difference between the average and marginal cost curves:

$$OptimalToll = 0.6t_f \left(\frac{V}{C}\right)^4 VOTT_{avg} \qquad (4)$$

where $VOTT_{avg}$ is the monetary value of travel time of the average marginal traveler (roughly \$5 to \$15 per hour, depending on the traveler and trip purpose)[20]. Other variables are defined as for Eq. 1.

---

[18] Under California Health and Safety Code Section 43845, this policy applies to businesses with 50 or more employees, in air quality non-attainment areas.

[19] Efficient travel choices will have effects on several related markets. For example, land use choices and wage decisions will become more efficient, by recognizing the true costs of accessing goods, services and jobs.

[20] The toll should equal the value of the additional travel time the last vehicle adds to the facility. This last vehicle adds $\frac{dt(V)}{dV}V$, where $V = \sum_i V_i(t, Toll)$ and *i* indexes the various classes of users demanding use of the road under that toll and travel time. The appropriate moneterarized value of travel time to interact with this added time is

Figure 6 adds a downward-sloping demand or "willingness to pay" (WTP) curve to Figure 5's travel time "supply" functions. The untolled supply curve represents the approach currently defining roadway provision, while the higher, tolled curve incorporates all marginal costs. The intersection of the demand curve with these two defines the untolled (excessively congested) and tolled (less-congested) levels of use on this roadway. The first results in 13,000 veh/h, travel times of 2.0 min/mi, and delays of 1.0 min/mi. The latter results in 10,450 veh/h, travel times of 1.44 min/mi, delays of 0.44 min/mi, and (the equivalent of) a 1.75 min/mi toll. In this example, neither situation is uncongested. Demand exceeds capacity in each instance, so queues will build for as long as demand and supply curves remain at these levels.

Imposition of the net-benefit-maximizing "optimal toll" equates benefits (measured as willingness to pay, here in the form of time expenditures) and marginal costs. Under this 1.75 min/mi toll, realized demand is 10,450 veh/h and toll revenues reach almost 18,300 min (1.75×10,450). Moreover, traveler benefits – measured as WTP in excess of travel time plus toll – exceed 25,100 min[21]. If time is converted to money at a rate of $12 per vehicle-hour (for the marginal traveler), the toll is worth 35¢/mile, revenues are $3600 per hour per mile and the traveler benefits are worth $5000/h/mi.

Without the toll, traveler benefits exceed 39,000 minutes ([8-2] ×13000/2), or $7800/h/mi. But there are no toll revenues. Thus, the $8600 of value derived under the optimally tolled situation exceeds the laissez faire approach by over ten percent.

For a ten-percent addition in value during peak times of day on key roadway sections, should communities ask their public agencies to step in and start charging on the order of 35¢ per congested mile of roadway? There are many issues for consideration. First, the revenues do not go to those who bear the cost of congestion (i.e., the delayed travelers). If these revenues are not well spent, society loses. Second, if demand is highly inelastic (i.e., steeply sloped in Figure 6), realized demand will not change much under pricing; revenues simply will be transferred from traveler benefits to the collecting agent. Third, if the cost of implementing and enforcing congestion pricing is high, any improvements in added value of the policy may be overcome.

However, the benefits of pricing on highly congested links may well exceed 10 percent. While *average* travel times may rise by a factor of two during peak periods in congested areas like Los Angeles, certain sections of roadway (such as key bridges) may experience much more severe delays. Imagine the same example in Figure 6 with a peak-period demand line that still begins at 8 min/mi but lies flatter, and crosses the $AC(V)$ curve at 17,500 veh/h, or 75 percent beyond capacity. In this case, untolled travel times are about 4.5 min/mi and marginal costs lie 13.7 min/mi above that level (a significant externality, and implicit subsidy). The optimal toll would

---

the demand-weighted average of $VOTT_i$'s: $VOTT_{avg} = \sum_i VOTT_i \cdot V_i(t, Toll) \Big/ \sum_i V_i(t, Toll)$. Given demand functions sensitive to time and toll and a link's performance function, it is not difficult to solve for the optimal toll. However, these inputs are tricky to estimate; they will be based on sample data and may involve significant error.

[21] Thanks to an assumed-linear demand function, this computation is relatively simple. It involves the triangular area under the demand curve, and above the 3.19-min generalized cost: (8-3.19) ×10450/2 = 25,132 min (per hour of flow, on this mile-long section of roadway).

be worth 3.6 min/mi (or 72¢/mile), bringing demand down to 12,500 veh/h. And tolled traveler benefits plus revenues would exceed untolled traveler benefits by a whopping 94 percent. Such a situation is probably a strong candidate for pricing.

Beyond increases in community benefits through removal of delay-related externalities, there are other advantages of congestion pricing. These include reduced emissions and gasoline consumption, as idling is reduced, closer destinations are chosen, and cleaner modes of transportation are selected. They result in healthier species, less crop and property damage, and diminished threat of global warming. Such benefits are difficult to value, but Small and Kazimi's work (1995) puts the human health costs of emissions close to 5¢/mi in air basins like that of Los Angeles.

Another key benefit is the allocation of roadspace to the "highest and best users[22]." Many users with high values of travel time (including, for example, truck drivers delivering goods for just-in-time manufacturing processes) presently are doing whatever they can to avoid congested roadways. Those with little or no value of travel time (e.g., high school students traveling to a shopping mall) are taking up that scarce space. By introducing monetary prices, the types of travelers on our roadways would shift towards those with more money and less time.

Of course, high values of time often correspond to higher wages and wealth, so communities and policymakers are understandably worried about the regressive impacts of congestion pricing – relative to the status quo. This is particularly true in the short term, when home, work, school and other location choices are relatively fixed. Pricing will have land use effects that are difficult to predict; short-term activity location and timing inflexibilities suggest that restrained introduction of pricing will be necessary, and special cases of heavily impacted low-income travelers (for example, single parents with fixed work and child-care times that coincide with rush hour) may require credits and/or rebates. Toll revenues can be spent on alternative modes and other programs to benefit those most negatively impacted by such policies. For example, revenues from London's downtown cordon toll of 5 pounds (roughly $8) go largely toward transit provision. This experiment seems to be succeeding: speeds have doubled (from 9.5 to 20 mi/h), bus ridership is up 14 percent, and only 5 percent of central London businesses claim to have experienced a negative impact. (*Economist* 2003)

Equity Considerations: Minimum-Revenue Pricing, Rationing and Reservations

While various forms of congestion pricing already exist, in Southern California, Florida, Singapore, Milan, Rome, Trondheim and now London, equity and revenue-distribution implications rouse public concern (Button and Verhoef, 1998). These considerations also have spawned a number of creative policy proposals. They include Dial's (1999) "minimal revenue pricing" (where one route is kept essentially free, between every origin-destination pair), DeCorla-Souza's (1995) "Fast and Intertwined Regular (FAIR) Lanes"[23], and Daganzo's (1995) alternating tolled days across users.

[22] "Highest and best use" is a normative economic term traditionally used to describe a market-derived (top-bid) use. A social or community-derived highest and best use may differ.

[23] FAIR lanes involve demarcating congested freeway lanes into tolled Fast lanes and untolled Regular lanes. Regular lane drivers using electronic toll tags would receive credit for use of the facility; accumulated credits can be redeemed for use of the Fast lanes.

Penchina (Forthcoming) has demonstrated that demand must be relatively inelastic for Dial's (1999) proposal to have important advantages over marginal cost tolling (such as lower tolls, more stable tolls, and fewer tolled links). Given the variety of travel-choice substitutes people face for many activities (e.g., time of day, destination, and mode), relatively inelastic cases may be unusual.

Nakamura and Kockelman (2002) examined Daganzo's idea in the context of the San Francisco-Oakland Bay Bridge to assess winners and losers under a variety of alternate pricing and link-performance scenarios. Net benefits were significant, but, as with almost any policy, they identified some losers – unless revenues are specially targeted and link-performance functions are favorable.

Another policy one might consider is reservation of scarce road capacity, to allocate crossing of key network links (such as bridges) by time of day. This policy could involve a cap on free reservations and tolls for additional use. Airports grant (or sell) gate rights to airlines for exclusive use. And airlines sell specific seat assignments to passengers, effectively guaranteeing passage. Singapore rations vehicle ownership, through a Vehicle Licensing System wherein only a certain number of 5-year licenses are auctioned off each year (on line, through sealed bids). Coupled with the most extensive road-pricing system in the world, this city-state is a clear leader in congestion-fighting policies.

Owing largely to budget constraints, U.S. transportation agencies are looking more and more at toll roads as a congestion-fighting and roadway-provision option. These may be variably priced throughout the day, to recognize the premium service they provide when demand is high (yet tolled travel times remain reasonable). In addition, smooth-flowing high-occupancy toll (HOT) lanes encourage peak-period travelers to shift to buses and carpooling while allowing better use of such lanes through tolling of single-occupancy vehicles (SOVs).[24] Toll roads and HOT lanes may be the intermediate policies that spark widespread congestion pricing in the U.S. Yet the issue of equity and redistributive impacts remains.

Ideally, any selected solution will be cost-effective and relatively equitable. There is one possible solution that has the potential for near-optimal returns while ensuring substantial equity and efficiency. It is being called "credit-based congestion pricing". (Kockelman and Kalmanje 2003)

Credit-Based Congestion Pricing

Under a credit-based congestion-pricing plan, the tolling authority collects no net revenues (except, perhaps, to cover administrative costs), and travelers willing to reschedule their trips, share rides, or switch modes actually can receive money by not exhausting allocated cash credits. All collected tolls are returned in the form of per-driver rebates to licensed drivers in a region or regular users of a corridor.

A credit system for roadway use that provides for trading (and user rebates) addresses public concerns to a much greater extent than pricing alone – and thus could generate considerably more support. Credit banking and trading are becoming widespread in other domains. In 1999, the OECD found nine programs involving tradable permit schemes in air pollution control, five

[24] Toll collection can become an issue, however, since HOVs must be distinguished so that they are not charged.

in land use control, five in water pollution, 75 in fisheries management, and three in water resources. (Tietenberg 2002) At present, many more such applications exist. And, as long as the cost of trading credits is not high, *any* distribution strategy will result in an efficient use of the constrained resource. (Tietenberg 2002)

A National Academy of Science committee recently proposed a tradable credits systems as an alternative to the current Corporate Average Fuel Economy standards (TRB 2002). For each gallon of fuel expected to be consumed by a new vehicle over its lifetime, the manufacturer would need a credit. The annual allocation of credits to a manufacturer would be based on the company's production level and the government's target for fuel consumption per vehicle.

Clearly, there is a distinction between industrial applications involving the trading of emissions, fuel economy credits, and the consumer-oriented credit-based congestion-pricing application considered here. In the case of roadway networks, where the temporal and spatial attributes of the congested resource are critical, appropriate link pricing must also exist. Roadway pricing, therefore, is somewhat more complex than emissions or fisheries regulation; one cannot simply cap total vehicle miles traveled (VMT) and allow VMT credit use at any time of day or at any point in the network. Continual monitoring of traffic conditions and recognition of key links and times of day are needed. However, just as companies achieved the imposed reductions in SO_2 emissions more easily than had been foreseen initially, travelers also may have greater flexibility in their demand for peak-period travel than some anticipate. The potential for and presence of such flexibility is key.

Congestion Pricing Policy Caveats

While credit-based and other forms of pricing have the ability to reduce congestion to (and beyond) "optimal" levels, there are implementation issues that require careful consideration. First, the costs of the technology can pose serious hurdles. Many commercial vehicles carry GPS systems, but these presently cost hundreds of dollars. If the price of congestion is less than $100 per vehicle in a region per year (as it is in most regions), GPS systems may not make economic sense. If much cheaper local radio-frequency systems are used, their roadside readers (presently costing tens of thousands of dollars) are likely to be selectively distributed, leaving much of the network unpriced and some locations still congested. In either case, private third-party distribution of identification codes and formal legal protections are probably needed to ensure privacy expectations are met. Second, policy administration is rife with special needs. If only a few regions per state merit pricing, visitors may need to buy or rent identification systems for travel through and around the priced systems.[25] For enforcement purposes, video-image processing of rear license plates is probably necessary at key network points in order to identify vehicles without transponders or system violators. Lists of license plate holders would then be necessary, and regions and states would have to share information (as they presently do, for serious offenses).

Finally, much travel time is access time, at trip ends. The drive from one's home to the first arterial may have little to no travel time improvement available; similarly, the walk and elevator ride from one's parking space to one's office may already be minimized. These access times are

[25] Short-term visitors may be granted access without penalty, depending on time of stay and use of network.

not insignificant. Thus, even if congestion on all arterial roads is removed, total peak-period travel times may fall only 25% or less, for most travelers, even in highly congested regions.[26]

Conclusions

Congestion results from high demand for constrained systems and tends to manifest itself in the form of delay. A relative scarcity of roadspace and other forms of transportation capacity has lead to substantial congestion losses in many travel corridors, across many regions, at many times of day. The delays may be severe or moderate; the congestion may be recurring and anticipated or non-recurring and unpredictable.

Society pays for congestion not just through higher travel times and crash rates, uncertain and missed schedules, additional emissions, and personal frustration, but also higher costs for goods and services. After all, commercial delivery services must confront the same traffic delays that personal vehicle occupants face. These delays translate to lowered productivity and more expensive deliveries and commutes, resulting in higher prices for everyone.

The relationship between demand and delay is not clear-cut. As demand approaches capacity, travel times tend to rise. When demand exceeds capacity, queues form and travel time impacts are pronounced. Small additions to demand can generate significant delays – and minor reductions can result in significant time savings. Removal of bottlenecks, toll road provision, and subsidy of alternative modes are proving popular mechanisms to combat congestion. But more effective solutions are needed, in most cases. Individual travel choices remain inefficient until travelers recognize and respond to the true costs of using constrained corridors and systems.

The good news is that new technologies are available to address the congestion issue. But robust estimates of individual demand functions and the marginal costs of additional users are needed, to take congestion policies to the next level. Congestion pricing promises many benefits. And *credit-based* congestion pricing, as well as other strategies, may substantially offset the burdens on automobile-dependent low-income populations, particularly in the short term, when location and other requirements are relatively fixed for certain activity types.

Congestion is not just a roadway phenomenon. It affects nearly all pathways, including air, rail, water, and data ports. Fortunately, many of the strategies and technologies for resolving roadway congestion also apply to these other domains. Greater recognition of the negative externalities involved in oversaturation of our transportation systems will guide us to the most effective strategies for coping with congestion, with the promise of more efficient travel for everyone.

Acknowledgements

The author is grateful for the excellent suggestions of Tim Lomax, Steve Rosen, Sukumar Kalmanje, and Annette Perrone.

[26] Taylor (2002) makes this point in his *Access* article "Rethinking Traffic Congestion."

References

Akçelik, R. 1991. Travel time functions for transport planning purposes: Davidson's function, its time-dependent form and an alternative travel time functio*n. Australian Road Research 21* (3), 44-59.

Bates, John, John Polak, Peter Jones, and Andrew Cook. 2001. The Valuation of Reliability for Personal Travel. *Transportation Research* 37E (2): 191-229.

Beamon, Benita. 1995. Quantifying the effects of road pricing on roadway congestion and automobile emissions. PhD Dissertation. Department of Civil and Environmental Engineering. Georgia Institute of Technology. Atlanta, Georgia.

Beckmann, M., McGuire, C. B., and Winsten, C.B. 1956. *Studies in the Economics of Transportation*, Chapter 4, Yale University Press.

Boarnet, Marlon G., and Randall Crane. 2001. *Travel by Design: The Influence of Urban Form on Travel*. Oxford, Oxford University Press.

Brodsky, H., and Hakkert, A.S. 1983. Highway Accident Rates and Rural Travel Densities. *Accident Analysis and Prevention* 15 (1): 73-84.

BTS. 2002. *National Transportation Statistics 2002*. Bureau of Transportation Statistics, U.S. Department of Transportation. Washington, D.C.

Button Kenneth J., and Erik T. Verhoef (Eds.) 1998. *Road Pricing, Traffic Congestion and the Environment.* Edward Elgar Publishing: Cheltenham, UK.

Cervero, Robert, and Kara Kockelman. 1997. Travel Demand and the Three Ds: Density, Diversity, and Design. *Transportation Research* 2D (3): 199-219.

Chen, Chao, Zhanfeng Jia, and Pravin Varaiya. 2001. "Causes and Curves of Highway Congestion." *IEEE Control Systems Magazine* 21 (4): 26-33.

Commodity Flow Survey (CFS). 1997. U.S. Department of Transportation, Bureau of Transportation Statistics, and U.S. Department of Commerce, Economics and Statistics Administration, U.S. Census Bureau. Washington, D.C.

Daganzo, C. 1995. A Pareto optimum congestion reduction scheme. *Transportation Research B 29*, 139-154.

Dahlgren, Joy. 1994. "An Analysis of the Effectiveness of High Occupancy Vehicle Lanes." Dissertation, Department of Civil Engineering, University of California at Berkeley.

Dahlgren, Joy. 2002. "High-Occupancy/Toll Lanes: Where Should They be Implemented." *Transportation Research A*, Vol. 36, pps 239-255.

DeCorla-Souza, Patrick. 1995 "Applying the Cashing Out Approach to Congestion Pricing." *Transportation Research Record.* No. 1450: 34-37.

Dial, Robert B. 1999. "Minimal Revenue Congestion Pricing Part I: A Fast Algorithm for the Single-Origin Case." *Transportation Research* 33B: 189-202.

Dowling, R., W. Kittelson, J. Zegeer, and A. Skabardonis. 1997. *NCHRP Report 387,* "Planning Techniques to Estimate Speeds and Service Volumes for Planning Applications." TRB, National Research Council, Wahsington, D.C.

Downs, A. 1962. The Law of Peak-Hour Expressway Congestion. *Traffic Quarterly* 16: 393-409.

Economist. 2003. Congestion Charge: Ken's Coup (March 22nd): 51.

Evans, L. 1991. *Traffic Safety and the Driver*. Van Nostrand and Reinhold, New York.

Federal Highway Administration (FHWA). 1979. *Urban Transportation Planning System* (UTPS). Washington: U.S. Department of Transportation.

Federal Highway Administration (FHWA). 1999. CORSIM, Version 4.2. Federal Highway Administration, U. S. Department of Transportation (Distributed through McTrans, University of Florida).

Fimrite, P. 2002. "Traffic Tops List of Bay Area Banes, Weak Economy is Number 2 Bane, Survey Shows." San Francisco Chronicle. At http://sfgate.com/cgi-bin/article.cgi?file=/c/a/2002/12/05/MN51835.DTL#sections. (Accessed on 5 December 2002).

Fulton, L., Meszler, D.; Noland, R.; Meszler, D.J., and Thomas, J. 2000. "A Statistical Analysis of Induced Travel Effects in the U.S. Mid-Atlantic Region." *Journal of Transportation and Statistics* (3) 1, 1-14.

Garber, Nicholas, and Sankar Subramanyan. 2002. "Feasibility of Incorporating Crash Risk in Developing Congestion Mitigation Measures for Interstate Highways: A Case Study of the Hampton Roads Area." Virginia Transportation Research Council, Final Report VTRC 02-R17. Charlottesville, Virginia.

Garber, Nicholas J., and Lester A. Hoel. 2002. *Traffic and Highway Engineering*, Third Edition. Pacific Grove, California; Brooks-Cole.

Gwynn, D.W. 1967. Relationship of Accident Rates and Accident Involvements with Hourly Volumes. *Traffic Quarterly* 21 (3): 407-418.

Hansen, M. and Huang, Y. 1997. "Road Supply and Traffic in California Urban Areas." *Transportation Research-A*, 31 (3), 205-218.

Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162 (1968): 1243–1248.

Horowitz, Alan J. 1991. Delay Volume Relations for Travel Forecasting based on the 1985 Highway Capacity Manual. Federal Highway Administration. Report FHWA-PD-92-015.

Klein, Lawrence. 2001. *Sensor Technologies and Data Requirements for ITS*. Boston, Artech House.

Knickerbocker, Brad. 2000. "Forget Crime - But Please Fix the Traffic." *Christian Science Monitor*. (February 16)

Kockelman, Kara. 1997. "Travel Behavior as a Function of Accessibility, Land Use Mixing, and Land Use Balance: Evidence from the San Francisco Bay Area." *Transportation Research Record No. 1607*: 116-125.

Kockelman, Kara. 2001. Modeling Traffic's Flow-Density Relation: Accommodation of Multiple Flow Regimes and Traveler Types. *Transportation* 28 (4): 363-374.

Kockelman, Kara, and Sukumar Kalmanje. 2003. “Credit-Based Congestion Pricing: A Policy Proposal and the Public’s Response.” Paper presented at the 10th International Conference on Travel Behaviour Research; Lucerne, Switzerland. August.

Kockelman, Kara, and Young-Jun Kweon. 2002. “Driver Injury Severity and Vehicle Type: An Application of Ordered Probit Models.” *Accident Analysis and Prevention* 34 (3): 313-321.

Kockelman, Kara K., Randy Machemehl, Aaron Overman, Marwan Madi, Jacob Sesker, Jean (Jenny) Peterman, Susan Handy. 2001. “Frontage Roads in Texas: A Comprehensive Assessment.” University of Texas at Austin, Center for Transportation Research Report FHWA/TX-0-1873-2.

Krishnamurthy, Sriram, and Kara Kockelman. 2003. “Propagation of Uncertainty in Transportation-Land Use Models: An Investigation of DRAM-EMPAL and UTPP Predictions in Austin, Texas.” *Transportation Research Record*.

Levinson, David, Lei Zhang, Shantanu Das, and Atif Sheikh. 2002. Ramp Meters on Trial: evidence from the Twin Cities Ramp Meters Shut-Off. Paper presented at the 81st Meeting of the Transportation Research Board. Washington, D.C.

Lomax, Tim, Shawn Turner, and Richard Margiotta. 2003. “Monitoring Urban Roadways in 2001: Examining Reliability and Mobility with Archived Data.” Federal Highway Administration Report FHWA-OP-02-029. Washington, D.C.

Lomax, Tim, Shawn Turner, Gordon Shunk, Herbert S. Levinson, Richard H. Pratt, Paul N. Bya, and G. Bruce Douglas. 1997. *NCHRP Report 398*, “Quantifying Congestion: Volume 1, Final Report.” TRB, National Research Council, Washington, D.C.

May, Adolf D. 1964. “Experimentation with Manual and Automatic Ramp Control.” *Highway Research Record* 59: 9-38.

Mokhtarian, Patricia, and Ilan Salomon. 2001. “How Derived is the Demand for Travel?” *Transportation Research* 35A: 695-719.

Myers, Barton, and John Dale. 1992. Designing in Car-Oriented Cities: An Argument for Episodic Urban Congestion. In *The Car and the City: The Automobile, the Built Environment, and Daily Urban Life*. Martin Wachs and Margaret Crawford (eds.). Ann Arbor: University of Michigan Press.

Nakamura, Katsuhiko, and Kara Kockelman. 2002. “Congestion Pricing & Roadspace Rationing: An Application to the San Francisco Bay Bridge Corridor.” *Transportation Research* 36A (5): 403-417 (2002).

National Household Travel Survey (NHTS). 2001. United States Department of Transportation, Federal Highway Administration. Washington D.C.

Newman, L., A. Dunnet and J. Meirs. 1969. “Freeway Ramp Control: What It Can and Cannot Do.” *Traffic Engineering* (June): 14-25.

Noland, R.B., and W.A. Cowart. 2000. “Analysis of Metropolitan Highway Capacity and the Growth in Vehicle-Miles of Travel. Paper presented at the 79th Annual Meeting of the Transportation Research Board, Washington D.C.

Oak Ridge National Laboratory (ORNL). 2001. 1995 NPTS Databook. Prepared for the U.S. DOT, Federal Highway Administration. ORNL/TM-2001/248.

Peirce, Neal. 2003. “Congestion Insurance: “HOT” Lanes’ Amazing Promise.” Washington Post. March 5.

Penchina, C. M. Forthcoming. “Minimal-Revenue Congestion Pricing: Some More Good-News and Bad-News.” Accepted for publication in *Transportation Research B.*

Pushkarev, Boris S., and Jeffrey M. Zupan. 1977. *Public Transportation and Land Use Policy*. Bloomington, Indiana: Indiana University Press.

Richardson, Tony. 2003. “Some Evidence of Travelers with Zero Value of Time.” Paper presented at the 82nd Annual Meeting of the TRB. Washington, D.C. (January)

Redmond, Lothlorien S., and Patricia L. Mokhtarian. 2001. The positive utility of the commute: modeling ideal commute time and relative desired commute amount. *Transportation* 28: 179-205.

Rodier, C.J., J.E. Abraham, R.A. Johnston, and J.D. Hunt. 2000. “Anatomy of Induced Travel Using an Integrated Land Use and Transportation Model in the Sacramento Region.” Paper presented at the 79th Annual Meeting of the Transportation Research Board, Washington D.C.

Salomon, Ilan, and Patricia L. Mokhtarian. 1998. “What Happens when Mobility-Inclined Market Segments Face Accessibility-Enhancing Policies?” *Transportation Research* 3D (3): 129-140.

Schrank, David, and Timothy Lomax. 2002. *The 2002 Urban Mobility Report*. Texas A&M University, Texas Transportation Institute.

Scheibal, S. 2002. “New Planning Group Kicks-off Effort with Survey on What Area Residents Want.” *Austin American Statesman.* Austin, Texas. (August 26).

Shoup, D. 1992. *Cashing Out Employer-Paid Parking,* Report No. FTA-CA-11-0035-92-1. U.S. Department of Transportation. Washington, D.C.

Shoup, Donald. 1994. Cashing out employer-paid parking: A precedent for congestion pricing? In *Curbing Gridlock, Peak-Period Fees to Relieve Traffic Congestion*. National Academy Press, Volume 2, 152-200. Washington, D.C.

Small, Kenneth A., and Camilla Kazimi. 1995. On the Costs of Air Pollution from Motor Vehicles. *Journal of Transport Economics and Policy* 29 (1): 17-32.

Smith, Adam. 1776 (Original text). *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: Dent & Sons (1904 pub date).

Taylor, Brian D. 2002. “Rethinking Traffic Congestion.” *Access* 21 (Fall): 8-16.

Tietenberg, Tom. 2002. "The Tradable Permits Approach to Protecting the Commons: What Have We Learned?" Chapter 6 in The Drama of the Commons. Washington, D.C.: National Research Council, National Academy Press.

TRB. 2000. *Highway Capacity Manual 2000*. Transportation Research Board, National Research Council, Washington D.C.

TRB. 2002, *Effectiveness and Impact of Corporate Average Economy (CAFÉ) Fuel Standards*, National Academy Press, Washington.

Walters, A. A. 1961. "The Theory and Measurement of Private and Social Costs of Highway Congestion," Econometrica, Volume 19, No. 4.

Walters, A. A. 1954. "Track Costs and Motor Taxation," *Journal of Industrial Economics*.

Ward's Automotive Yearbook 1998. 1999. Ward's Communication, Intertec Publishing Corporation.

Weisbrod, Glen, Donald Vary, and George Treyz. 2001. *Economic Implications of Congestion, NCHRP Report 463*. National Cooperative Highway Research Program, Transportation Research Board. Washington, D.C.

LIST OF FIGURES:

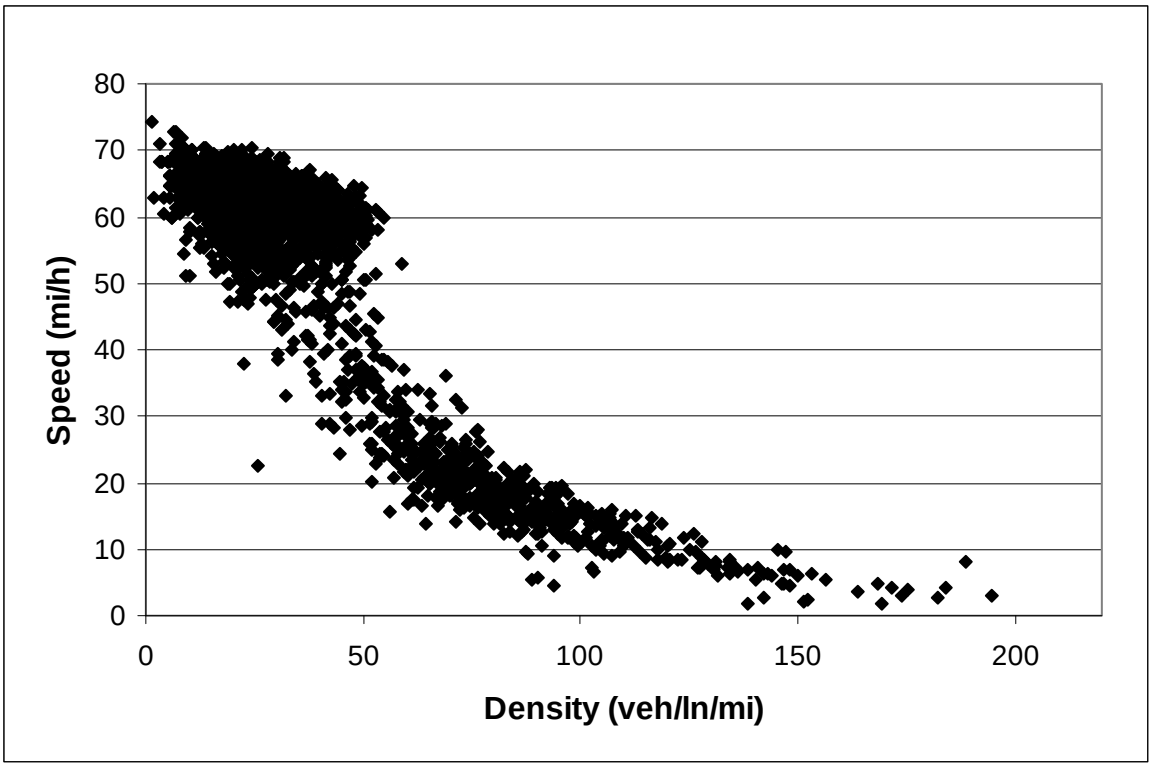

Figure 1. Speed versus Density
(Lane 2 Observations from Northbound I.H. 880; Hayward, California)

Note: This northbound freeway section consists of five lanes near Hayward, California, and the number one (or left-most) lane is an HOV lane. The plotted data come from a rainy day and a dry day in early 1993. The single-station dual inductive-loop detectors' 30-second data have been multiplied by 120 to represent equivalent hourly values. The data are indicative of this segment's general operations; however, these two days' data exhibit more congested points than typically observed during that period. (Should this be a footnote in the text/discussion, or an endnote for this Figure?)

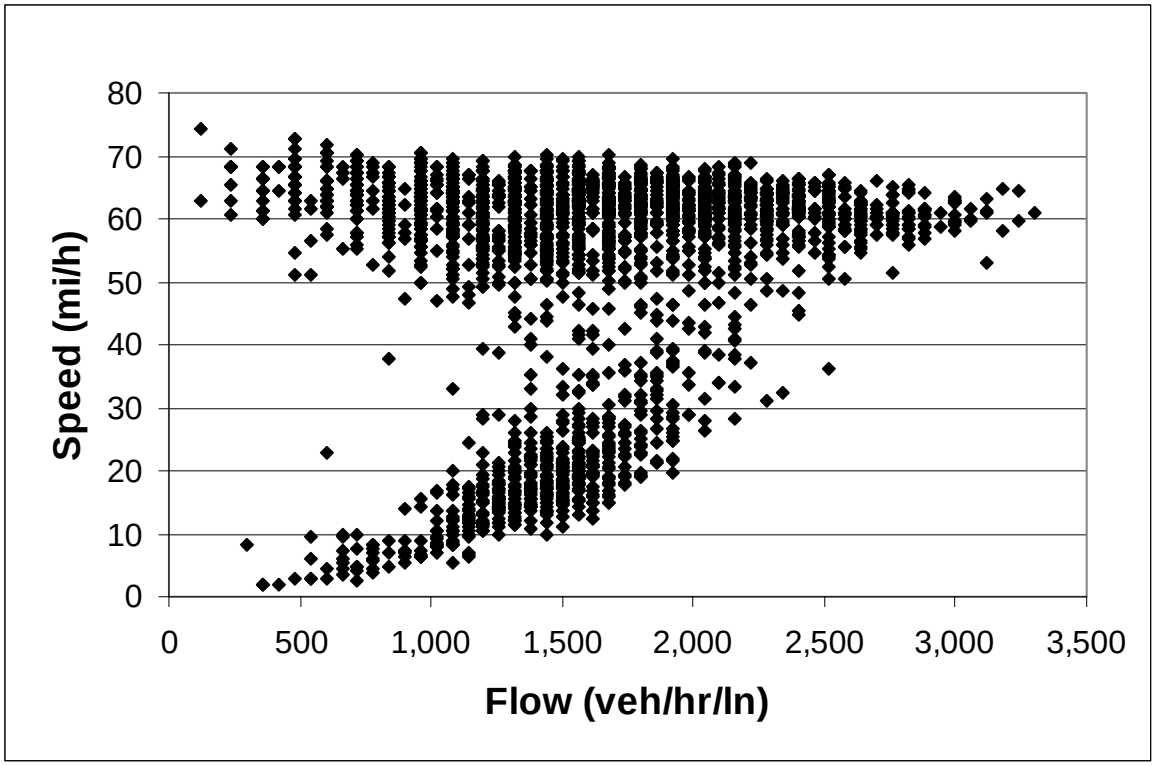

Figure 2. Speed versus Flow
(Lane 2 Observations from Northbound I.H. 880; Hayward, California)

Figure 3. Image of Congestion

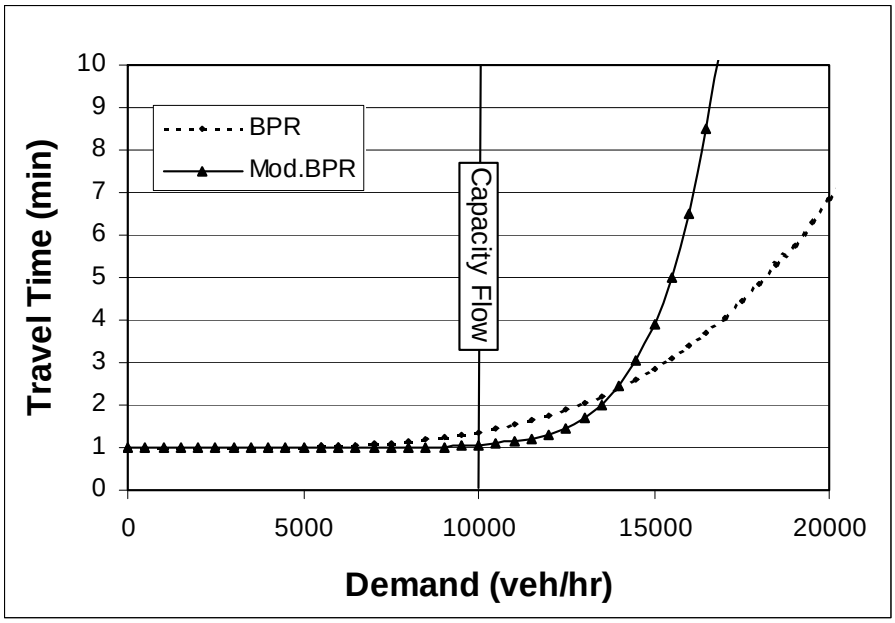

Figure 4. Travel Time versus Demand: BPR & Modified-BPR Formulae
(Capacity = 10,000, C = 8,000 veh/h, t_f = 1 min)

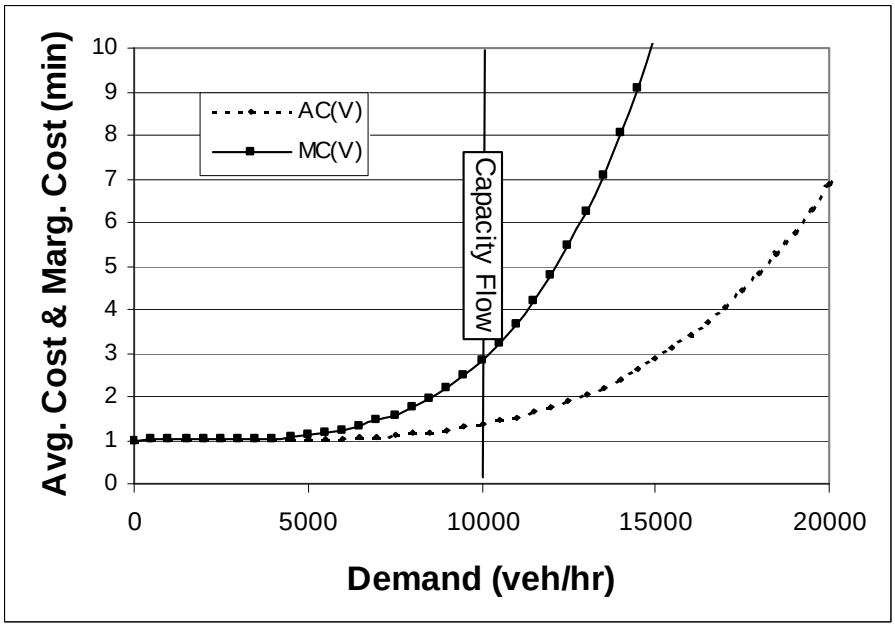

Figure 5. Average and Marginal Cost of Demand
(BPR Formula, Capacity = 10,000, C = 8,000 vph, t_f = 1 min)

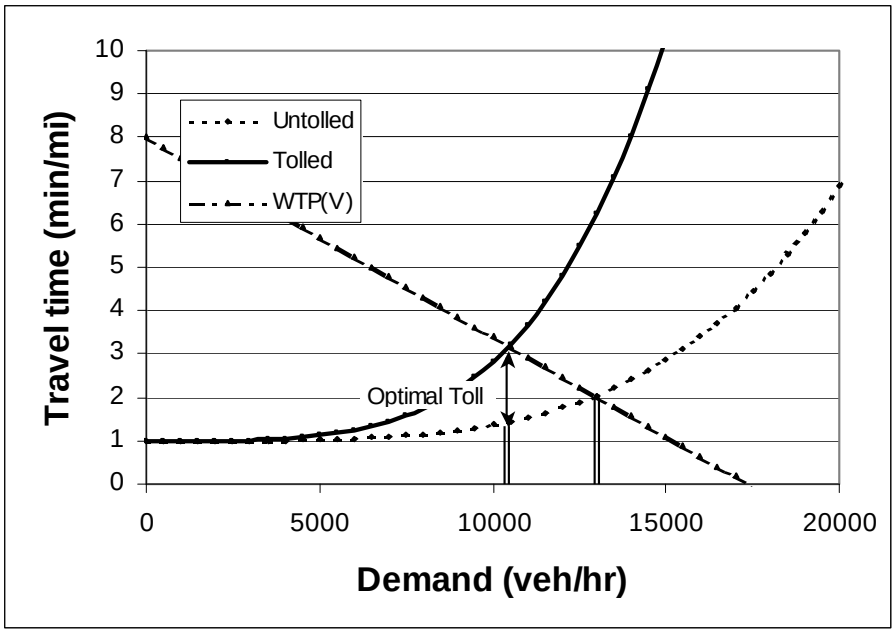

Figure 6. Demand versus Supply: Tolled and Untolled Cases
(Tolled & Untolled Equilibrium Demand levels are 10,450 and 13,000 veh/h; Optimal Toll is equivalent to 1.75 min/mi)