Arthur De Vany
Simon Fraser University

Thomas R. Saving
Texas A&M University

Competition and Highway Pricing for Stochastic Traffic*

This paper models the competitive supply and pricing of highways with random traffic. It adds the missing theory of the firm to Knight's highway problem and shows that competition delivers efficient prices. The classic efficient-pricing results of Knight and Walters are shown to be valid in a world with uncertain traffic flows. Competitive prices exceed marginal cost by an amount equal to expected marginal congestion cost. The mean throughput of each highway firm is less than maximum throughput. Even though the expectation of excess capacity is positive, random traffic jams will occur on an efficiently priced highway. Pricing of mixed traffic is also analyzed.

The received theory of highway pricing, as so well developed by Walters, is a certainty, or certainty equivalent, theory which does not address the issues raised when there is randomness in traffic flows.[1] When traffic is random, one wonders if an equilibrium flow exists, and if it does what form of congestion toll is appropriate. Is a uniform toll which does not vary with the number of vehicles on the highway efficient? How does uncertainty affect the level of capacity which is to be provided and its optimal rate of use? These questions are examined here within a model of a competitive market for highways functioning under uncertainty.

To construct a competitive model of the highway market might seem a useless exercise. Are not highways publicly owned and managed because private provision of highways would be inefficient? Our answer is that public ownership has been a convenient way to solve the free-rider problem (excluding nonpayers and monitoring

*The authors wish to acknowledge the thoughtful comments of the referee. This research was supported by the National Science Foundation through grant SOC 7917884 to the Texas A&M Research Foundation. While on leave from Texas A&M University, De Vany's work was supported by the Center for the Study of the Economy and the State, and the Law and Economics Program, both of the University of Chicago.

1. Much of this received theory has been developed as an extension of Knight's (1924) discussion of Pigou (1920). See, for example, Walters (1961), Mohring (1970), Gould (1972), and Keeler and Small (1976).

(*Journal of Business*, 1980, vol. 53, no. 1)

0021-9398/80/5301-0005$01.37

use), but, as the cost of exclusion and monitoring falls, private provision of highways becomes feasible. It is, therefore, interesting to study competition as an alternate to public provision of highways. Of more immediate relevance is the fact that the pricing that emerges from competition is under certain conditions efficient and, therefore, is of normative significance for management of the public roadways. Moreover, to the extent that highway authorities move in the direction of direct pricing, they will face the problems of traffic diversion, mixed traffic, and random traffic jams, all of which are studied here in a competitive setting.

In many respects our excursion into uncertainty brings us back to the principles set forth in Alan Walters's (1961) classic highway paper. Our work corroborates his results while generalizing them and also extends the analysis to the choice of highway capacity. We retain a feature of his model which has been dropped in much of the subsequent highway pricing literature by allowing transit time (in our case, expected transit time) to influence the demand for use of the highway. This feedback from demand through congestion and back to demand turns out to be a very convenient way to handle the quality-of-service problem and results in a model which determines prices and highway capacity simultaneously.

The Walters paper drops one of the central features in Knight's original formulation, which is the competitive allocation of vehicles over alternate roadways, in favor of using a welfare function to derive efficient prices. We retain Knight's competitive process by characterizing a competitive allocation of vehicles to roadways under uncertainty. We go a step beyond Knight, following a vein recently explored by Gould, to supply the missing theory of the firm to Knight's model. In this manner, Knight's model is closed and we obtain a complete model of the competitive equilibrium in the market for highway services.

There are two basic contributions of the paper. First, it derives classic results on efficient highway pricing in a setting that is more realistic than other studies and also derives efficient capacity levels. Second, it offers some new insights into the operation of competition in the area of uncertainty and congestion. We find that a fixed toll in the face of random arrivals is expected cost minimizing and that under competition, the price equals marginal cost plus a congestion toll. The congestion toll reflects the cost of the incremental capacity required to maintain expected transit time for the highway's users when the expected output rate is increased by one unit. The sum of these congestion tolls collected from all users just suffices to cover the cost of highway capacity. Contrary to the certainty theory, efficiently priced highways do experience traffic jams; they are random, however, and the expectation of output is less than the maximum throughput rate of the highway. Each competitive firm operates with excess capacity on average but experiences overloads which occur on a random basis.

I. Highway Congestion

Consider a highway between points A and B. For simplicity assume that the only points of access and egress for the highway are A and B. We can visualize the arrival at, flow through, and exit from the highway as a stochastic process. Given the availability of alternative routes and other conditions, the expected number of vehicles entering the roadway at points A and B per unit of time will depend on the toll, if any, and the time required to traverse the roadway. We can write the cost, or full price, of using the highway as

$$P = p + \eta T, \tag{1.1}$$

where P is the full price of trip AB, p is the toll, η is the value of time for highway users, and T is expected trip time. Since the analysis is independent of direction traveled we can, without loss of generality, ignore the trip BA.[2]

Assume that the time between arrivals to the highway is random and independently distributed with expected interarrival time of $(1/\alpha)$. Under these conditions, the expected number of arrivals in any interval of time t is $[t/(1/\alpha)] = t\alpha$, so that expected arrivals per unit of time is α. Given that highway users have a common value of time (or alternatively, that within classes of users the marginal value of time is equal) the arrival rate, that is, the expectation of the number of arrivals per unit of time, will depend on the full price, so that

$$\alpha = \alpha(P), \tag{1.2}$$

where $\alpha' < 0$. Essentially, α is the stochastic equivalent to the quantity demanded in a certainty setting.

The element in the free-access resource problem that differentiates it from the more usual problem considered in economics is the effect of crowding on the ability of any given resource to service new and existing users. In the case of highways, the ability of the system to service users can be summarized by transit time. All the evidence, and it is ample, indicates that the transit time remains constant for low levels of highway usage and then rises. Ultimately, when the system becomes overloaded, transit time approaches infinity. Thus, the total throughput on the highway first rises with increased traffic, then peaks out and ultimately declines.[3]

2. Note that we have abstracted away from the joint effect of return trips. That is, whether an individual takes a trip or not may be a function of transit time in both directions. Here we consider these decisions to be of a sequential nature for simplicity.

3. See Walters (1961) for an excellent discussion of the effect of congestion on throughput and transit time in a certainty setting.

To capture this congestion effect on the ability of the highway to handle traffic, we assume that the required time for vehicles to complete the trip AB is independently distributed with a common mean $(1/\mu)$. In this case, μ can be interpreted as the expected number of one-way vehicle trips per unit of time. Or, if we treat the highway as having unit length, μ is expected vehicle speed. Let μ be a function of highway size, k, and the number of vehicles using the roadway at any point in time, N, so that

$$\mu = \mu(\overset{-}{N}, \overset{+}{k}). \tag{1.3}$$

The assumed signs of the derivatives of (1.3) are consistent with the empirical evidence for highways. The negative sign for the derivative with respect to number of vehicles implies that, for fixed size, mean trip time $(1/\mu)$ is an increasing function of the number of vehicles using the roadway. Similarly, the positive sign for the highway-size derivative implies that increases in the capital stock reduce mean trip time.

It will prove convenient if we transform our discussion from mean trip time to number of vehicles completing trips, that is, throughout. From the above discussion it follows that the conditional expected throughput n, given that there are N vehicles in the system, is

$$n = N\mu(N, k). \tag{1.4}$$

The unconditional expectation of the number of vehicles completing trips will depend on the probability distribution for the number of vehicles using the roadway (the system size), N.[4]

For fixed roadway size, or capital stock, k, the expected rate at which vehicles could exit the highway is conditional on the number on the road, and is first an increasing and then a decreasing function of the number of vehicles using the roadway. Essentially, what happens is that an increase in the number of vehicles on the highway increases the number completing the trip until congestion becomes great enough. Once this critical congestion level is reached, further increases in the number of vehicles on the highway reduces the flow of vehicles completing trips. The effect can be seen by differentiating the conditional expected values of completed trips per unit of time with respect to N,

$$\frac{\partial n}{\partial N} = \mu + N\mu_N. \tag{1.5}$$

4. It is important to distinguish between the number of vehicles that could *potentially* complete trips over the highway and the number which *do* complete trips. The former, which is called the service mechanism in the queuing literature, is obviously conditional on there being an adequate number of vehicles available so that the service mechanism may function at its potential rate. The latter, which is the number which do complete trips, depends not only on the potential of the highway to handle trips but also on the number of vehicles occupying it. Below we derive the steady-state throughput of the system, which depends on arrivals as well as capacity.

Now if $\mu_{NN} < 0$ and $\mu_N \simeq 0$, for small N, both of which are indicated by the evidence, then n will first rise, reach a peak, and then fall as N rises.

Once an individual enters the highway system he is essentially serving himself. It seems natural then to view the highway system as a self-service queuing model with the service rate per individual equal to $\mu = \mu(N, k)$. In this model it is assumed that the highway has the capacity to hold any number of vehicles. Thus, we ignore the queues that develop at the highway entrance. The transit time is dependent on the state of the system and, for any given state of the system N, the expected transit time is

$$E(T \mid N) = \frac{1}{\mu(N, k)}. \tag{1.6}$$

Assume that potential users of the roadway do not know the system state, but only know the probability distribution over the feasible states of the system. Given the above discussion, this probability distribution can be derived as the solution to an arbitrary Markovian birth-death process.[5] The steady-state probability distribution for system size is

$$\begin{aligned} p_N &= \left[\frac{\alpha^N}{\prod_{i=1}^{N} i\mu(i, k)}\right] p_0;\ N \geqslant 1, \\ p_0 &= \left[1 - \sum_{N=1}^{\infty} \frac{\alpha^N}{\prod_{i=1}^{N} i\mu(i, k)}\right]^{-1}, \end{aligned} \tag{1.7}$$

where p_N is the probability that N vehicles are on the roadway at any time.

From (1.7) it is clear that the existence of the steady-state probabilities for system size requires that p_0, the probability that the highway is empty, be positive. The necessary and sufficient condition for $p_0 > 0$ is that the infinite sum

$$S = \sum_{N=1}^{\infty} \prod_{i}^{N} \frac{\alpha_i}{i\mu(i, k)} \tag{1.8}$$

be positive and finite. The sum is positive since all its terms are nonnegative. A necessary but not sufficient condition for S to be finite is that each term in S be finite. The meaning of this is that the highway cannot become so congested that movement of vehicles ceases and

5. In a Markovian birth-death process, it is assumed that instantaneous changes in the system state are restricted to an increase (birth) or a decrease (death) of one. Gross and Harris (1974, pp. 92–95) give the details of the derivation.

throughput becomes zero: that is, there cannot exist an N such that $N\mu(N, k) = 0$.

The condition $N\mu > 0$ implies that the throughput of the roadway never reaches zero; that is, the freeway never literally becomes a parking lot and vehicles always exit the road, even if only at a trickle. Indeed, a positive value for the probability of the system being empty requires that, on the average, the utilization rate must be less than unity; that is, excess capacity must exist. We could have made the existence and stability conditions somewhat less restrictive by accounting for the fact that the roadway has a maximum system size and that waiting at on-ramps occurs. In this case we would have greatly increased the complexity of the problem with little or no gain in the results. Accordingly, we shall assume that the above infinite sum is finite, that is, that on the average the system has excess capacity.

Using (1.6) and (1.7) we can derive the unconditional expected transit time and number of vehicles in the system. The expected transit time is

$$\bar{T} = \sum_{N=1}^{\infty} \frac{1}{\mu(N, k)} p_{N-1}, \tag{1.9}$$

where each element in the sum is the product of the conditional transit time, given that the system is of size N and the probability that the system is of size $N - 1$ when a user arrives—implying that the system will be of size N when the potential user enters it. From (1.7) and (1.9) expected transit time is

$$T = \left[\frac{1}{\mu(1, k)} + \sum_{N=1}^{\infty} \frac{\alpha^N}{N!\mu(N + 1, k)} \prod_{i=1}^{N} \frac{1}{\mu(i, k)}\right]$$

$$\times \left[1 + \sum_{N=1}^{\infty} \frac{\alpha^N}{N!} \prod_{i=1}^{N} \frac{1}{\mu(1, k)}\right]^{-1}. \tag{1.10}$$

Thus, we can write expected transit time as a function of the arrival rate and the capital stock or roadway size, that is,

$$\bar{T} = T(\overset{+}{\alpha}, \overset{-}{k}). \tag{1.11}$$

The signs of the derivatives of the function T in (1.11) are directly obtainable from (1.10) and the assumption that $\mu_k > 0$.

The expected system size is simply

$$\bar{N} = \sum_{N=0}^{\infty} N p_N, \tag{1.12}$$

which from (1.7) and (1.9) is

$$\bar{N} = \alpha\bar{T}. \tag{1.13}$$

The relation (1.13) states simply that, in the steady state, the expected transit time must equal the time required for the expected number in the system to arrive. If an individual enters the roadway and it contains the steady-state expected number of vehicles, he also expects it to contain the same number as he exits. But, if the arrival rate per unit of time is α, then it must take $(\bar{N}/\alpha)$ units of time for $\bar{N}$ individuals to arrive and enter the roadway.[6]

From (1.11) and (1.13) we can write the steady-state expected transit time and system size as functions of the toll and roadway capacity,

$$\bar{T} = f(\overset{-}{p}, \overset{-}{k}) \tag{1.14}$$

and

$$\bar{N} = g(\overset{-}{p}, \overset{+}{k}). \tag{1.15}$$

The derivatives of (1.14) and (1.15) are derived by treating (1.11) and (1.13) as implicit functions and imposing the condition that the expected transit time used by demanders be the actual expected transit time.[7] The signs of the derivatives of (1.14) and (1.15) are intuitive in that increases in the toll given roadway capacity decrease both the expected transit time and the expected number in the system. An increase in roadway size increases the expected number in the system while reducing expected transit time.[8]

We are now in a position to derive the steady-state demand condition facing a roadway firm operating between points A and B. From the original specification of demand (1.2) and the solution for the steady-state expected transit time (1.14), we can write the steady-state demand function

$$\alpha^* = \alpha^*(\overset{-}{p}, \overset{+}{k}). \tag{1.16}$$

Treating demand (1.2) as an implicit function, the derivatives of (1.16) are

$$\frac{\partial \alpha^*}{\partial p} = \frac{\alpha'}{1 - \alpha'\eta T_\alpha} < 0,$$
$$\frac{\partial \alpha^*}{\partial k} = \frac{\alpha'\eta T_k}{1 - \alpha'\eta T_\alpha} > 0,$$

6. In the queuing literature, (1.13) is referred to as Little's law, and it is quite generally valid. See Gross and Harris (1974, pp. 59–64) for a discussion of Little's law and of its general applicability.

7. With any form of rational expectations the differences between the expected transit time on the left-hand side of (1.11) and the demanders' expected transit time will go to zero.

8. If there is an infinite-capacity alternate route, then the demand elasticity with respect to full price would be infinity. Under these conditions, arrivals at the roadway would increase until expected transit time returned to its previous level. In this case, changes in roadway capacity will not improve traffic conditions on the roadway, and $\partial T/\partial k$ would equal zero.

where T_α and T_k are the partial derivatives of (1.11) with respect to α and k.

II. Competition

To construct a competitive market for highway services, consider the following setting. Let each firm own an identical roadway with transit time equal to $1/\mu(N, k)$ and let each firm charge a fixed toll which does not vary with the number of vehicles on the road. Assume that once a vehicle enters a particular firm's road it is free to choose any lane; however, vehicles may not leave the firm's roadway until reaching the exit. Let there be many alternative roads, running between A and B, each supplied by a separate firm.

We assume that drivers do not have information on the numbers of vehicles using each road at any point in time and thus enter each roadway ignorant of the number of cars already on it. Since each roadway will hold an infinite number of vehicles, there is always room for additional vehicles. Each road is labeled, and through repetitive use of the roadways, drivers learn the tolls and expected transit times of each. Drivers are indifferent among roads and seek to minimize total transit cost, which we treat as being the full price. The allocation of drivers over the roads will equalize expected full prices over roads in that if any road has consistently shorter transit time for the same toll as other roads, then drivers will expand use of the faster roadway until its transit times are consistent with other roadways.

In this way each firm becomes a full-price taker; whatever price it charges (within the feasible range), its resultant flow of vehicles will be such that its expected full price is equal to the market-determined full price. Full-price taking behavior implies that, in equilibrium, the steady-state expected full price must be identical across roadways. The equality of full price across roadways will be imposed on the firms by the distribution of arrivals to firms. If a roadway has a lower full price than the market-clearing full price, then full-price minimizing behavior on the part of roadway users will result in increased arrivals until the expected full prices across roadways are equal. Thus, high-priced roads must have traffic flows small enough so that transit time is sufficiently low to compensate for the high price.[9]

9. Since the steady-state full prices are identical across roadways, then consumers are indifferent as to which roadway they travel. But equality of full prices also implies that arrivals at roadways with n lanes must exceed arrivals at roadways with m lanes whenever $n > m$. One mechanism which yields this distribution of customers is to treat customers as randomly choosing lanes rather than roadways. This problem is exactly the same as the fast-food problem with differing numbers of servers. More customers show up at the outlet with the most servers, and yet customers can be said to be indifferent between outlets. Actually, customers are indifferent between servers.

Given that full price is market determined, then for each firm it follows that

$$P = p + \eta T(a, k), \tag{2.1}$$

where $T(a, k)$ is the individual firm equivalent of (1.11). The steady-state arrival rate can be solved for from (2.1) by assuming that the waiting time expected by demanders is the actual expected waiting time. Thus, the steady-state arrival rate can be written as

$$a = a(\overset{-}{p}, \overset{+}{k}), \tag{2.2}$$

where p, k, and a refer to firm rather than industry levels of prices, capacities, and arrivals.

The signs of the derivatives of (2.2) can be derived by totally differentiating the full price, holding it constant at its market equilibrium level. These derivatives are[10]

$$\begin{aligned} \frac{\partial a}{\partial p} &= -\frac{1}{\eta T_a} < 0, \\ \frac{\partial a}{\partial k} &= -\frac{T_k}{T_a} > 0. \end{aligned} \tag{2.3}$$

The price derivative reflects the fact that an increase in the toll must decrease arrivals by an amount sufficient for the reduced transit-time cost to restore the firm's full price to the market level. Similarly, an increase in the capital stock must increase the arrival rate by an amount sufficient to restore the transit-time cost to its initial level.

Having characterized the demand process at the firm, we now proceed to examine the firm's profit-maximizing choices of price and capacity. For this to be a reasonable problem to address, we require some limitation in firm size. Such a limitation does not come from the demand process since the full-price taking firm can increase output without limit by adding more lanes. Moreover, there are some economics of scale to be gained by having all lanes owned by a single firm, thereby affording drivers the opportunity to change lanes and more effectively distribute traffic to the least congested lanes. We assume that the economies of pooling traffic on a multilane highway dissipate beyond a certain size because lane jockeying adversely affects traffic flow when a large number of lanes are involved. If, in addition, we assume that the costs of monitoring highway use—collecting tolls and excluding free riders—and maintaining the roads rise more than proportionately with size, then we will have a limit on firm size.

10. This approach requires that the marginal and average value of transit time be equal and constant. If they are not, then the analysis of Spence on the relation between the marginal valuation schedule for product quality can be applied.

We can capture these ideas with a simple cost structure. Let total expected cost be

$$C = C(a, k), \tag{2.4}$$

where a is expected output and k is capital (lanes) as before. Suppose monitoring and maintenance costs depend on mean output, on the grounds that the firm does not vary monitoring resources with fluctuations in traffic, fixing them instead on the basis of expected traffic, and that maintenance cost depends only on the number of vehicle trips and not on how they are distributed. Then it is reasonable to assume that the marginal cost of expected traffic is positive and rising. Suppose further that capital (lanes) may be added at constant or rising cost, but that the costs of managing and monitoring larger highways are rising so that the marginal cost of additional capacity is also positive and rising. Then $C(\cdot)$ is convex and provides a limit on firm size.

If firms maintain a fixed toll and allow transit time to fluctuate randomly, expected profits are

$$\pi = pa(p, k) - C(a, k). \tag{2.5}$$

Since full price is market determined, firms can choose any two of the three variables that determine full price: toll, capacity, and arrivals. The maximization of expected profits with respect to the toll and roadway capacity requires that

$$p = C_a + a(-a_p)^{-1}, \tag{2.6}$$

$$(p - C_a) = C_k(a_k)^{-1}. \tag{2.7}$$

Condition (2.6) suggests that, in equilibrium, the competitive price is the sum of direct marginal cost plus a congestion charge.

The congestion charge is equal to the value of the additional time imposed on existing users by the addition of one more steady-state user. Alternatively, the charge measures the gain to remaining roadway users should there be a reduction of one user in the steady state. This interpretation becomes clearer if we use (2.3) to rewrite (2.6) as

$$p = C_a + a\eta T_a. \tag{2.8}$$

In this case, ηT_a is the value of the marginal expected transit time per user. The quantity $a\eta T_a$ is just equal to what existing users would pay a prospective user not to use the facility and is the competitive congestion toll.

From (2.7), it follows that capital is chosen so that the transit-time constant marginal revenue equals the incremental capital cost associated with a unit increase in steady-state output, holding transit time constant. Moreover, using (2.7) and (2.8), we have

$$C_k(a_k)^{-1} = a\eta T_a. \tag{2.9}$$

Using (2.3), we can write (2.9) as

$$C_k = -a\eta T_k, \tag{2.10}$$

so that in equilibrium, the marginal cost of capital equals the value of its marginal product.

We have then three views of, or rationales for, the congestion toll that would be levied by the competitive firm. The toll is a levy exacted by the firm on additional users to compensate existing users for their increased transit time, it is a bribe which existing users would pay to the firm not to allow additional users, or it is a capital charge made by the firm for the additional capacity it must provide in order that the new users do not increase the transit time of existing users. Under each of these views, the firm sees additional users as imposing an opportunity cost or a direct resource cost on the firm and internalizes this crowding externality as a cost which, in turn, is levied on the highway users through the congestion toll.

The average cost of providing roadway for expected throughput a is

$$AC = \frac{C(a, k)}{a}. \tag{2.11}$$

For constant transit time (quality), average expected costs will be a minimum when

$$C_a + C_k(a_k)^{-1} = AC. \tag{2.12}$$

But from (2.8) and (2.9), we have

$$p = C_a + C_k(a_k)^{-1}, \tag{2.13}$$

so that in competitive industry equilibrium, constant-quality average cost is a minimum and the tolls cover the cost of capital.[11]

In the steady state, the rate at which vehicles arrive must equal the rate at which they exit the highway, that is,

$$a^* = N^*\mu(N^*, k), \tag{2.14}$$

where * denotes steady state. Thus, in the steady state, we have

$$N^* = \frac{a^*}{\mu(N^*, k)}, \tag{2.15}$$

and

$$T^* = \frac{1}{\mu(N^*, k)}. \tag{2.16}$$

11. At the industry level, the zero profit condition requires that new firms can enter at the same costs of existing firms. If this is not the case, then firms who previously entered will earn rents on their roadways, but tolls will still cover costs of the marginal firm.

Changes in roadway capacity must then increase arrivals until system size rises enough to offset the effect of capital on vehicle speed. The empirical implication, for which we all have frustrating evidence, is that in a system with many alternate routes, that is, a competitive system, the opening of new lanes increases the number of users with little or no effect on vehicle speed.

A striking feature of the competitive outcome is that each firm achieves an output flow which is less than the maximum achievable traffic flow. Walters (1961) first showed that "desirable traffic flow is always less than capacity flow," and it is interesting to observe that competition achieves this result even under uncertainty. We can prove this by first noting that throughput n is

$$n = N\mu(N, k). \tag{2.17}$$

But in the steady state, $n = a^*$. The maximizing throughput price is p such that $(\partial a^*/\partial p) = 0$, which is inconsistent with competitive equilibrium. Competition requires that for each firm $(\partial a^*/\partial p) < 0$, so that the expected level of excess capacity will be positive.[12] In other words, mean highway throughput is less than maximum throughput, or engineering-rated capacity. The fact that the expectation of excess capacity is positive does not imply that the maximum throughput is never reached, but only that this cannot be the expected system state.[13] Thus, traffic jams do occur, in contrast to the certainty theory, but they are random and infrequent enough so that mean output is less than capacity.

III. Pricing and Priorities for Mixed Traffic

Up to this point the analysis has been carried out assuming that traffic is homogenous, that is, that all vehicles have the same expected transit time. It is clear, however, to even the casual user of the highway system that trucks and autos have different expected transit times. Moreover, the impact of additional truck traffic on auto transit time differs from the impact of additional auto traffic on truck transit time. We can capture the essential differences in these various types of traffic by focusing on three characteristics—time value of load, mean service time (the reciprocal of speed), and the effect that each form of traffic has on itself and other traffic forms at various levels of highway saturation.

12. The competitive industry produces the maximum output for any given level of quality. Thus, there is no excess capacity at constant quality.

13. Recall that it is necessary and sufficient for the existence of a steady-state solution that the system have a nonzero probability of being empty. For the state dependent service case, the service rate falls below the arrival rate for large N so that $p_0 = 0$ and the system explodes.

Let the service rate γ for automobiles over a roadway AB be related to N_M, the number of autos on the highway; N_T, the number of trucks; and k, the capacity of the highway, as

$$\gamma = \gamma(N_M, N_T, k). \tag{3.1}$$

Likewise, the mean service rate τ for trucks is given by

$$\tau = \tau(N_M, N_T, k). \tag{3.2}$$

Both γ and τ indicate how many trips either type of vehicle could be expected to complete per unit of time were they to be continuously in operation. Mean trip times are $1/\gamma$ and $1/\tau$, so that whenever the service rate falls, trip time rises.

Notice that the service rates depend upon the numbers of both types of vehicles present in the system. The effects of traffic on service rates may be summarized as follows:

$$\gamma_{Ni}, \tau_{Ni} < 0, i = M, T. \tag{3.3}$$

The greater the number of vehicles of either type on the highway, the lower is the service rate for each. It is generally agreed that service rates do not differ greatly for low congestion levels; that is, $\gamma \simeq \tau$ at low ratios of vehicles relative to capacity. As the volume of vehicles grows relative to capacity, the service rates for both autos and trucks fall, but the transit time of trucks relative to autos grows. This means that the difference $\gamma - \tau$ is positive and larger the greater is the total number of vehicles in the system, although the effect varies with composition of vehicles. In general then, $\gamma \geqslant \tau$.

By assuming that demand for roadway use for each vehicle type depends only on that vehicle type's full price, we can write the steady-state arrival rates of autos and trucks as functions of prices and capacity

$$\begin{aligned} c &= c(\overset{-}{p_M}, \overset{+}{p_T}, \overset{+}{k}), \\ t &= t(\overset{+}{p_M}, \overset{-}{p_T}, \overset{+}{k}). \end{aligned} \tag{3.4}$$

Expected profits are then

$$\pi = p_M c(p_M, p_T, k) + p_T t(p_M, p_T, k) - \phi(c, t, k), \tag{3.5}$$

where ϕ is the expected cost function. As above, the firms can control capacity and arrivals or capacity and prices but not all three since the full prices are market determined. The maximization of profits requires that prices be set so that

$$p_M = \phi_c + c\eta_M(\partial T_M/\partial c) + t\eta_T(\partial T_T/\partial c), \tag{3.6}$$

$$p_T = \phi_t + c\eta_N(\partial T_M/\partial t) + t\eta_T(\partial T_T/\partial t). \tag{3.7}$$

Thus, the equilibrium prices must equal the direct marginal cost plus congestion charges equal to the sum of the value of the time lost by both autos and trucks because of an additional steady-state auto or truck arrival.

It is clear that the user charges must take the composition of vehicles in the system into account as well as the total number of vehicles. The reason for this lies in the magnitude of the cross-congestion effect. If, for example, the effect of an additional auto on truck transit time is greater than the effect of an additional truck on truck transit time when the proportion of trucks exceeds one-half, then efficiency would require a higher congestion toll for autos than trucks during those periods when trucks comprise the bulk of the traffic, such as nighttime on cross-country highways. During rush hour on urban expressways, trucks would pay higher tolls than autos.

In queuing systems where the service rates differ between classes of users, a priority system which assigns highest priority to the user with the highest service rate can reduce the total waiting time. However, if the users with low priority have the higher value of time, then it is not necessarily desirable to give priority to those users with the lowest service time. In hilly terrain, where the service rate for trucks is much lower than for cars, efficiency may require that trucks be restricted to the right lane when climbing. Downhill, the service rates do not differ much, and this restriction might not be useful except to reduce lane switching, which can reduce the throughput of the highway.

IV. "Price Taking" as a Norm for the Highway Authority

We hypothesized competition in the operation of highways to derive efficient road-user charges. It is immaterial if competition is feasible in the provision of highways, since the efficient prices we derived should be charged by a highway authority who seeks to maximize the public good, here interpreted to be an efficient allocation of resources. The theory of competitive pricing has often been pointed to as providing a normative model of behavior for the manager of a publicly owned enterprise. Since highway authorities are typically regional monopolists, it is clear that one must consider the implications of this monopoly power when highway pricing is advocated.

In this context, the preceding analysis provides normative rules for pricing and sizing the highway system, and the allocative results of those rules are efficient. For example, suppose the highway authority is considering adding a road segment to its system and that this road will divert users from other roads. Should the benefits of the added segment be counted as the user charges received for its use less the revenue lost on other roads? Competitive theory tells us that diversions from other firms are ignored when a firm makes a decision, and this rule is a guide to the right decision in the present case.

If the authority calculates the marginal revenue of a new highway, it will underestimate benefits because the marginal benefits revealed to it are those revealed to a monopolist who internalizes the decreasing returns to highway capacity. The benefits of the new roadway capacity should be calculated as its average revenue. If the addition is sizable, then the addition will reduce the average valuation of capacity and should be valued at this lower level. In other words, road capacity should be expanded until the average revenue of capacity equals the cost per mile.

The price-taking behavior described above leads to optimal tolls and capacity if all roads are priced. If, as is the more usual case, some if not all roads are priced at zero, then rules derivable from "second-best" considerations become relevant. In these cases, traffic diverted from other roadways does matter. In particular, Mohring (1970) has shown that if all roadway tolls are required to be equal, then the price-taking behavior will, in general, not be optimal. Thus, a price-taking decision rule will only apply if the road in question is to be priced, and then only if the only constraint on other roadways is that their tolls are restricted to zero.[14]

V. Summary and Conclusions

The above results that highway tolls should exceed the direct marginal cost by a congestion charge are similar to results obtained in previous work using a certainty model. There are, however, some significant differences between the certainty models and our work that affect conclusions concerning the operation of the highway system under optimal pricing. The certainty theory assumes the existence of a steady state, whereas, for our stochastic model, the steady state exists only if there is "excess" capacity in the highway system. In the certainty theory, prices which reflect marginal congestion cost result in traffic flows well below the saturation level of the highway. Under uncertainty this nice result cannot be achieved since the efficient prices result only in an expected traffic flow which is less than the saturation level. Accordingly, a positive probability of saturation exists, and therefore the highways will on occasion become "parking lots."

Because of the positive probability of a traffic jam, even with an optimal pricing structure, information systems become valuable. The value of such systems lies in their impact on the arrival rate. Once the

14. The problem here is under what conditions should the highway authority desire to meet the first-best optimizing conditions even though it cannot meet them for all roadways. This is the general problem of the second best and is handled exceptionally well by Davis and Whinston (1965). In that work, they prove that if the additional constraints (e.g., tolls restricted to zero) contain only variables for the deviant, then, except for the deviants, all other first-best conditions should be met. Thus, even if some roadways have tolls restricted to zero, it does not follow that competitive pricing is not optimal on those roadways with tolls.

system reaches a saturated state, the only way it can return to the range of stability is for the arrival rate to fall. Information on the state of the system can result in just such a decrease in the arrival rate.

A number of freeway systems have various rules regarding the allowed behavior of certain classes of vehicles, principally trucks and cars. The reason for distinguishing these two vehicle types is that they have differential impact on the service rate of the system and one another. Moreover, this differential impact varies by terrain and roadway. Our results indicate that an efficient pricing structure of automobiles versus trucks depends on the relative values of time and service rates. The auto toll is greater the greater is the marginal impact of automobile traffic on the truck service rate and the greater is the value of time for trucks. Similarly, the truck toll is greater the greater is the marginal impact of trucks on the automobile service rate and the greater is the value of time of automobiles.

References

Davis, O. A., and Whinston, A. B. 1965. "Welfare economics and the theory of second best." *Review of Economic Studies* 32 (January): 1–14.

Gould, J. R. 1972. "Externalities, factor proportions and the level of exploitation of free access resources." *Economica* 39 (November): 383–402.

Gross, Donald, and Harris, Carl M. 1974. *Fundamentals of Queueing Theory.* New York: Wiley.

Keeler, Theodore E., and Small, Kenneth A. 1977. "Optimal peak-load pricing, investment, and service levels on urban expressways." *Journal of Political Economy* 85 (February): 1–25.

Knight, F. H. 1924. "Some fallacies in the interpretation of social cost." *Quarterly Journal of Economics* 38 (November): 582–606.

Mohring, H. 1970. "The peak load problem with increasing returns and pricing constraints." *American Economic Review* 60 (September): 693–705.

Pigou, A. C. 1920. *The Economics of Welfare.* London: Macmillan.

Walters, A. A. 1961. "The theory and measurement of private and social cost of highway congestion." *Econometrica* 29 (October): 676–99.