PRICE REGULATION AND OPTIMAL SERVICE STANDARDS

The Taxicab Industry

By George W. Douglas[1]

Economic regulation of transport involves a problem common to many regulated industries, the specification of an appropriate quality level or service standard. In many industries the problem is ignored, or treated only implicitly or heuristically. In regulation of transport, however, it can play a crucial role in the determination of the price level and in the resulting equilibrium of the industry. Typically, even if the regulatory authorities do not explicitly deal with service quality, one can perceive within the context of the competitive structure of the transport industry an endogenous relationship between price level and quality of service. Usually this derives from the capacity of the system, its headways and the stochastic parameters of the demand it serves. To explore this problem, we will consider a hypothetical market for taxicabs.

The taxicab industry, in its usual urban setting (where a taxi is engaged by hailing one that is vacant), provides a service in which an important element of service quality, waiting time to engage a taxi, is not amenable to differentiation. That is, an impatient customer cannot effectively transmit to the supplier his willingness to pay for a reduced wait, but must accept the stochastic wait given by the equilibrium of the market. The crux of the regulatory problem is that, a priori, there is no "normal" utilisation rate for taxis, and no concomitant "normal" level of service quality, as measured by the expected waiting time.[2] Rather, one can show that the market can reach an equilibrium in response to any price within a wide range of prices, each equilibrium being characterised by a different level of common service quality. The regulators' dilemma, then, is the selection of a price, and implicitly a single service standard, for a population of customers with diverse preferences for service quality.

The model described below is not meant to portray accurately any specific taxicab market, but rather to serve as a basis for discussion of the central problem of this paper, the selection of an optimal level of price and service quality.[3] It is clear that other transport industries have a similar problem; the capacities and headways of a

[2]A previous analysis of a taxicab market considered the utilisation rate to be given exogenously. See Daniel Orr, "The 'Taxicab Problem'; A proposed solution," in *Journal of Political Economy*, Volume 77, No. 1, January 1969, pp. 141–147.

[3]The model does roughly approximate the market in Washington, D.C., where fares are regulated yet entry is free. Other markets quite often have constrained entry, resulting in significant capitalisation of the rents of the existing franchises. This constraint is clearly suboptimal.

116

system affect its costs, and hence its fares and its total use.[4] While there exist other dimensions of service quality, and, within limits, conceivable institutional arrangements which could allow some quality differentiation, the following analysis is purposely drawn with strong constraints, to focus on the problem described above.

A CRUISING TAXICAB MARKET

Consider a taxicab market with the following characteristics: A taxi is engaged by visual contact between the customer and driver, anywhere along the city streets. The fare schedule is set by regulatory authority, but entry is free.

Demand

The demand for taxi trips is assumed to be related to the trip price, and to a proxy for service quality,

$$Q = f(P, T)$$

The aspect of service quality of particular interest here is the expected delay time incurred in engaging a taxi. For convenience of analysis other variables affecting trip demand, such as prices on alternative modes, income, and other quality dimensions, will be held constant. The trip length and the distribution of speeds *en route* are assumed to remain constant over the various equilibrium states investigated; hence we may use as a proxy for the number of trips Q, defined as the total time each taxicab in service is occupied during a reference period ("occupied taxi-hours").

P is a proxy for the fare schedule, here assumed to be a simple linear relationship with distance or time of trip.

$$\text{Trip price} = P \times (\text{time of trip})$$

As pointed out above, to secure a taxi in this market a customer must accept a schedule delay, or wait, which is determined by the equilibrium of the market as a stochastic distribution, with mean T. (In the further interests of simplicity, we assume initially a homogeneity of space and density, so that we need describe only one delay distribution.) We assume that the partial derivative of trips demanded with respect to expected wait, T, is negative.

We assume also that the nature of this market effectively precludes quality differentiation in this dimension. Regardless of a customer's preferences and willingness to pay to avoid schedule delay, he must accept the stochastic delay distribution generated by the market equilibrium.[5]

---

[4]The issue is of great significance, for example, in the air transport industry. The United States' Civil Aeronautics Board is currently reviewing the level and structure of air fares and their relation to load factors.

[5]Various ways could be devised to relax this constraint. The simplest, perhaps, would be the identification of certain taxi stands, or locations, at which a premium would be charged. Customers willing to pay this premium would be likely to receive better service as taxis are attracted to these locations. This operates informally in existing markets as taxi drivers congregate at those hotels where the customers are likely to provide large tips.

117

Production Costs

We assume that the supply of taxicabs and drivers in the industry is perfectly elastic, and can be given as a cost per hour of service time, c.[6] We assume that the cost of operation of cruising taxicabs is independent of the division of service time between "occupied" and "vacant".

Total cost, then, is given by

$$TC = c(Q + V),$$

where V is the number of "vacant taxi-hours"; $(Q + V)$ is the total number of taxi-hours in service.

The delay distribution is a function of the density of *vacant* taxicabs in the area. We assume that the taxicabs move randomly through the streets when vacant. The rate at which they pass any point i is positively related to the density and speed of the vacant taxis. The time between contacts (the delay) is the inverse of the rate of contact; hence we would expect the mean of the delay distribution, T, to be given by (letting V be a proxy for density of vacant taxis and S being their average speed):

$$T = g(V, S)[7]$$

---

[6]This simplying assumption is made for convenience only. Alternatively, following Orr, one could read $c = C'(q)$ where $q = Q + V$. The analysis and conclusions that follow remain the same.

[7]The approximate distribution of delay times can be derived: Let $\bar{A}$ be defined as the number of street miles in the reference area. Let $A(t)$ represent the area searched since some initial time $t = t_0$. The "gross search rate" is equal to VS, where V is the number of vacant taxis and S is their speed. The "net search rate", adjusting for redundant coverage, then is

$$\frac{dA(t)}{dt} = \left[\frac{\bar{A} - A(t)}{\bar{A}} \right] VS$$

Let the area *not* searched at time t be $B(t) = \bar{A} - A(t)$. The rate of change of $B(t)$ is equal to the (negative) net search rate,

$$\frac{dB(t)}{dt} = - \left[\frac{\bar{A} - A(t)}{\bar{A}} \right] VS$$

$$\frac{dB(t)}{dt} = - B(t) \frac{VS}{\bar{A}}$$

Therefore, $B(t)$ is of the form

$$B(t) = \bar{A}e^{-\rho t},$$

$$\rho = \left[\frac{VS}{\bar{A}} \right]$$

The delay distribution $P(t)$ is the probability of not having found a taxi at time t,

$$P(t) = \frac{B(t)}{\bar{A}} = e^{-\rho t}$$

with expected value $T = \int t \cdot P(t) dt = \frac{1}{\rho} = \frac{\bar{A}}{VS}$

118

where $$\frac{\partial T}{\partial V} < 0 \qquad\qquad\qquad \frac{\partial T}{\partial S} < 0.$$

Market Equilibrium

With free entry, and an exogeneously set price level $P = P^*$, equilibrium occurs when the marginal revenue obtained by the last unit of taxi service just covers its cost. We assume that all taxis will achieve over time the same occupancy ratio, $Q/(Q + V)$. Hence, equilibrium occurs at

$$P\left[\frac{Q}{Q + V}\right] = c$$

or,

$$PQ = c(Q + V).$$

The model can then be summarised:

$$Q = f(P, T) \qquad , \frac{\partial Q}{\partial P} < 0, \frac{\partial Q}{\partial T} < 0$$

$$T = g(V) \qquad\quad , \frac{\partial T}{\partial V} < 0$$

$$PQ = c(Q + V)$$

$$P = P^*$$

There exists, then, a set of prices P_f^*, which we will call the "feasible price set", for which the system of equations has a solution. For any $P_i^* \ \varepsilon \ P_f^*$, the model defines an equilibrium vector (P_i, Q_i, V_i, T_i). The range of outcomes of the model as P_i^* is varied within the set P_f^* is of interest. Starting with $P_0 = \min (P_f^*)$, consider the effect of successively higher price levels on the equilibrium of the market. At P_0 the equilibrium occupancy ratio is high, implying that the number of vacant taxis is low and the waiting time high. Consider the alternative equilibrium generated by a price $P_1 = P_0 + \delta$. The equilibrium occupancy ratio would be reduced; this probably implies that V would be increased, thereby decreasing T. The level of passenger trips demanded would change as

$$dQ = \frac{\partial f}{\partial P}dP + \frac{\partial f}{\partial T}\frac{\partial T}{\partial P}\,dP$$

The first term describes the reduction of trips demanded induced by the change in price, while the second term indicates the change in trips demanded caused by the induced change in service quality. The latter term, under most conditions, is positive. One might reasonably expect, then, that as the price is varied the level of trips might increase, and reach a maximum at some P^* (where $\frac{\partial f}{\partial P} + \frac{\partial f}{\partial T}\ \frac{\partial T}{\partial P} = 0$),

119

beyond which the equilibrium level of Q would decline. If we assume specific forms of the demand function and the delay function, the equilibrium may be simulated. For example, let

$$Q = A + B(P + rT)^{\alpha}$$

$$T = KV^{-1}$$

The various equilibria generated in this specific model by varying P^* are portrayed in Figure 1.

A graphical description of the equilibrium is given in Figure 2. A family of demand curves is drawn for various fixed levels of T, and hence of V. The supply curve in this model is the average cost function, or $c + \dfrac{cV_i}{Q}$.[8] Typically, the supply curve crosses the demand curve at two points, thereby generating two sets of solutions to the system of equations that defines the market. We consider only the set with greater Q for any value of P; this clearly dominates the other possible set of solutions.

The conditions for maximisation of Q within this system are found by:

$$Q \cdot \frac{\dfrac{\partial f}{\partial T}}{-\dfrac{\partial f}{\partial P}} \cdot \frac{\partial T}{\partial V} = c$$

or,

$$Q = -c \; \frac{\dfrac{\partial f}{\partial P}}{\dfrac{\partial f}{\partial T} \dfrac{\partial T}{\partial V}}$$

This equilibrium condition may be interpreted as equating with the marginal cost of V the marginal value (impact on Q) of the reduced waiting time generated by a unit increase in V.

For the sake of discussion, let us assume that the demand function is of the type that generates equilibria as in Figure 1. What are the characteristics of these potential equilibria?

A point of immediate interest is P_Q, where Q is maximised. Is P_Q the price that

---

[8] The marginal cost is equal to c for any given level of V, and hence for any given quality level. Moreover, average cost always exceeds marginal costs; setting price equal to marginal cost would incur a deficit equal to cV. An alternative interpretation of the cost function has been used by Herbert Mohring: he defines total costs inclusive of the delay costs of the travellers. Hence, marginal cost could be interpreted here as the marginal external delay costs imposed on other passengers by the addition of one more passenger, holding the number of taxi-hours constant. This interpretation is analogous to a short-run marginal cost, while the marginal cost defined above is analogous to a long-run marginal cost. See Herbert Mohring, "Optimization and Scale Economics in Urban Bus Transportation", in *American Economic Review* (forthcoming).

120
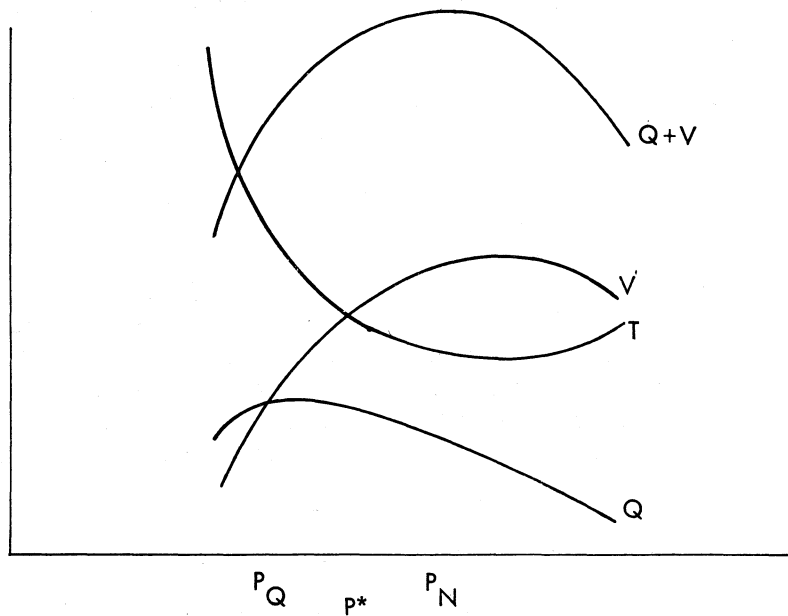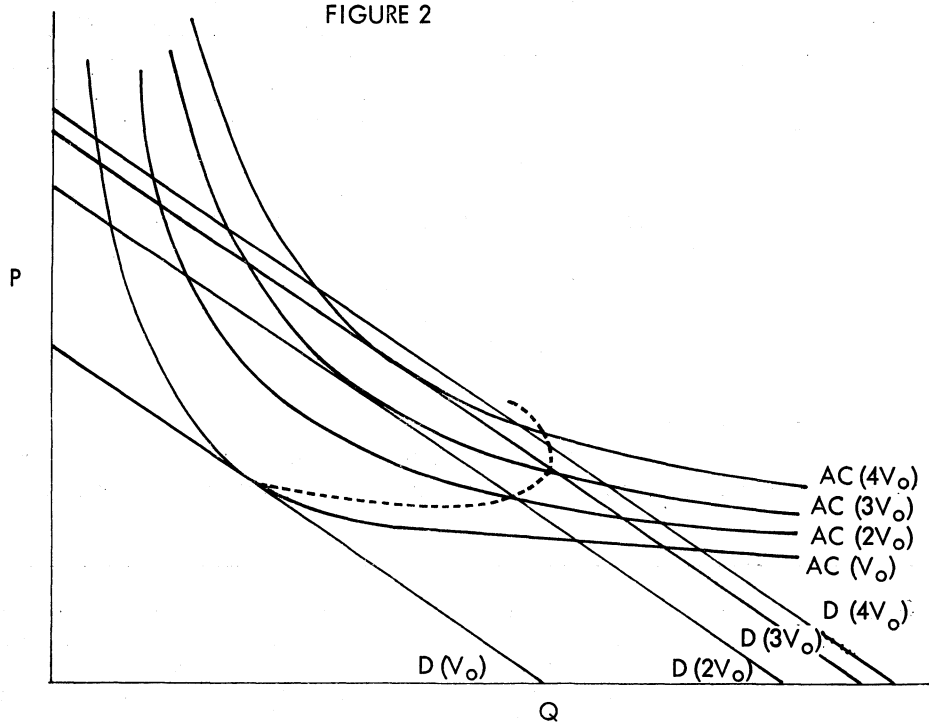
FIGURE 1   FEASIBLE EQUILIBRIA

P_Q $P*$ P_N

FIGURE 2

AC $(4V_o)$
AC $(3V_o)$
AC $(2V_o)$
AC (V_o)

D $(4V_o)$
D $(3V_o)$
D (V_o) D $(2V_o)$

Q

121

would be reached by an unregulated market? How does it relate to the price that would be set by a monopolist in this market? Does it represent, in any sense, an efficient point under the constraint that revenues cover costs?

The description of the equilibrium of this market without regulation (i.e., with P determined endogenously) requires an additional behavioural assumption, to replace $P = P*$.[9] It is conceivable (perhaps probable) that the unregulated equilibrium is indeterminate, taking any of the solution values shown in Figure 1. If there is a P at which the competitive system is in equilibrium, there must be no incentive for an individual firm (or driver of his own taxi) to change that rate, or for the number of taxis in service to change.

Consider the situation facing the individual taxi driver, if the prevailing price P_0 is less than P_N (Figure 1). The individual driver, acting alone, could raise his price to $P_0 + \delta$, and increase his income. This can be seen to follow if δ is less than the expected cost to the customer of his additional wait if he were to reject this higher priced taxi, and if the price elasticity of demand at P_0 is less than 1. If this is perceived by one driver, it should be perceived by all; a progression of price increases might result, each followed by an influx of additional taxis [as $P \left(\dfrac{Q}{Q + V} \right) > c$].

The total revenue to the industry continues to increase until it reaches P_n, where $\dfrac{d}{dP}$ (Total Revenue) $= 0$. It is at this point that the individual incentive to increase prices disappears. The depressing force on prices for $P > P_n$ is more indeterminate. It is not feasible for the individual taxi driver to exploit the price elasticity with a unilateral price reduction. The price the customer expects to pay is the prevailing price; hence a low-price taxi acting alone would have difficulty in communicating with potential customers and inducing additional trade with price reduction. However, the existence of taxi associations or taxi fleets could allow effective communication, making possible downward price competition. The point of maximum total revenue to the industry, which occurs at the point where the number of taxi hours in service $(Q + V)$ is maximised, might then be described as a possible point of stable equilibrium in an unregulated industry. This point is identified in figure 1 as N max, occurring at P_N.[10]

It can be shown, on the other hand, that the profit-maximising price to the monopolist with the same cost function, and facing the same demand function, would generate a level of trips, Q, less than Q max.

Welfare Characteristics

As pointed out above, given the decreasing· cost characteristic of this market, setting the price $= c$ and choosing an optimum V_0 would generate an efficient but

---

[9]Alternatively, it requires a completely different description, in which a market is made at each contact between customer and taxi. The price in each market would be subject to considerable variation, depending on bargaining skills, immediately available alternatives, etc.

[10]The additional assumption made by Orr (describing equilibrium under entry constraints) was that $Q/(Q + V) = N(P)$, where $N(P)$ is an unspecified "normal" utilisation ratio.

122

unfeasible (deficit) equilibrium.[11] Given the constraint that revenues cover costs, what would determine an optimal price in this market? If one defines social welfare as the summation of net benefits as given by the aggregate consumer surplus (since total costs are held equal to total revenues), one can specify a point of constrained welfare maximisation for the system. This can be derived by maximising W, where

$$W = \int P(Q, T)dQ - c(Q + V),$$

subject to

$$PQ = c(Q + V)$$

$$T = T(V)$$

and where $P(Q, T)$ is the inverse of the function $Q = f(P, T)$.

For a specific form of the function f, one may then derive a solution for Max W (if it exists), (P_W, Q_W, V_W, T_W).

Under some conditions this point coincides with the point where Q is maximised. It is the same, for example, for demand functions linear with respect to P and T. If the function is of the type of constant elasticities, $Q = A P^\alpha T^\beta$, P_W is less than P_Q if $\beta < -1$; P_W is greater than P_Q if $\beta > -1$; while equilibrium is not defined for $\beta = -1$.

If the demand function can be simplified so that it is given as

$$Q = f(P_t)$$

where

$$P_t = g(P, t),$$

then the price that maximises Q also maximises W. The proxy variable, P_t, may be interpreted as the total trip cost, being the sum of the price and the implicit cost of the waiting time. For example, if $P_t = P + rT$, then Q and W are maximised where P_t reaches a minimum.[12] On Figure 3 we have plotted P_t derived from equilibria generated by changing P. If demand is described in this way, it is intuitively clear that if every customer has the same value of time, r, net benefits are maximised at the equilibrium which minimises total trip cost, P_t. In a sense, then, we might describe the point of use maximisation, P_Q, as an "egalitarian optimum", indicating the implicit assumption of equal time value for the various customers. But not all customers have a value of time equal to the group average. Suppose the demand function of each individual could be represented as

---

[11] If external (congestion) costs were imposed on all traffic, the efficient fare would be equal to the marginal cost, c, plus the efficient congestion fee. This fare would cover the industry's costs if the implicit toll revenues exceeded the fixed cost of vacant taxis, cV_0. It would be inefficient, however, to impose such an implicit toll on taxis if it were not uniformly applied to other traffic.

[12] Let $P_t = h(Q)$ be the inverse of the function $W = f(P_t)$. Then it can be shown that $W = \int h(Q)dQ - P_t Q$ is equivalent to the net benefits as defined above. For the maximisation of W with respect to P, set

$$\frac{dW}{dP} = \frac{dW}{dQ}\frac{dQ}{dP} = 0$$

since Q is maximised where $\dfrac{dQ}{dP} = 0, \dfrac{dW}{dP} = 0.$
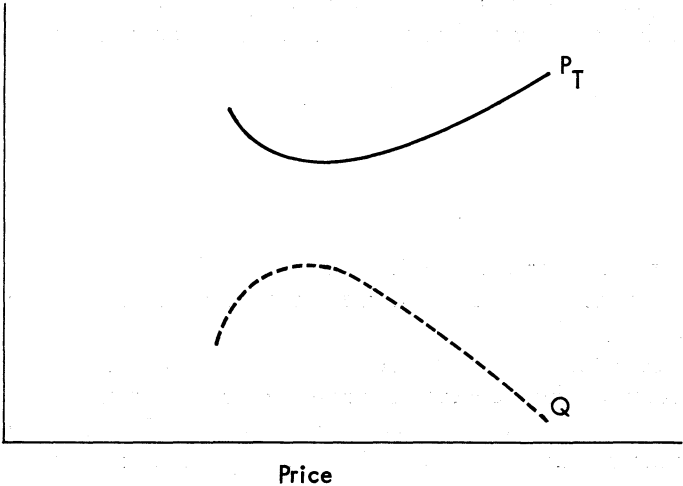
123

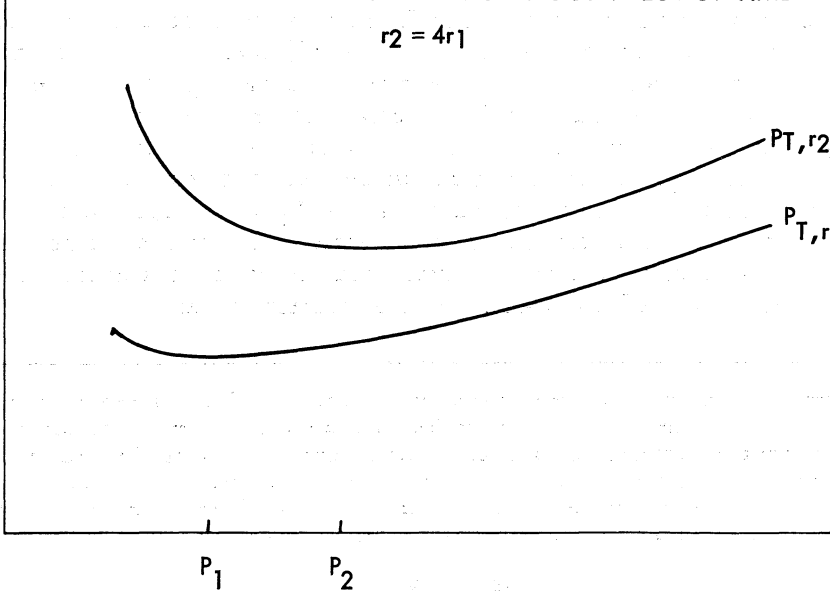FIGURE 3   TOTAL TRIP COST: HOMOGENEOUS VALUE OF TIME

Price

FIGURE 4   TOTAL TRIP COST: HETEROGENEOUS VALUE OF TIME

$r_2 = 4r_1$

124

$$Q_i = f_i (P_{t,i})$$

where

$$P_{t,i} = P + r_i T$$

and

$$Q = \Sigma \, Q_i = \Sigma f_i \, (P_{t,i}).$$

Then we can observe the effect of the selection of a price level P^* on each individual's welfare by observing the effect on $P_{t,i}$. An example is portrayed in Figure 4, in which we have assumed that market demand Q comes from only two groups, each having homogeneous time value r_1 and r_2. As might be expected, the existence of two or more groups with different time values creates a conflict situation. As price increases from c to P_1, $P_{1,t}$ and $P_{2,t}$ both decrease. In the interval $[P_1, P_2]$, however, $P_{1,t}$ increases while $P_{2,t}$ decreases. Beyond P_2, P_t increases for both groups. We could then limit the "efficient" prices to the set $[P_1, P_2]$. Within that set can be defined P_Q, where $(Q_1 + Q_2)$ is maximised. One can also define a point P_W, where the sum of individual net benefits is maximised (subject to the constraints of this system), assuming that the interpersonal addition implied is meaningful. In general $P_W \neq P_Q$.[13] More important, the hypothetical "competitive" price (which maximises $Q + V$) may be greater than P_2, in which case it is clearly inefficient.

The policy dilemma of defining a unique price and service quality remains. We would expect that the price should be chosen from the set (P_L, P_H), where P_L reflects cost minimisation to the customers with least time value, and P_H reflects costs to customers with the highest time value. Of the points within that set, perhaps P_Q has the greatest pragmatic appeal. A rationale for any single price would involve a balancing or weighting of the implicit preferences of the various customers. A choice of the point P_Q has the virtues of achieving this with an easily conceived and observed criterion, which does not involve imputations of the users' individual demand functions.

Scale Effects

The system of equations which defines this market may have no solution if, for example, demand is inadequate to support cruising taxis at a feasible level of time delay. Suppose we express the demand function as

$$Q = \lambda f(P, T),$$

---

[13]To find P_W, let $P_1 = f_1 (Q_1)$

$$P_2 = f_2 (Q_2)$$

Then $W = \int f_1 (Q_1) dQ_1 - Q_1 f_1(Q_1) + \int f_2 (Q_2) dW_2 - Q_2 f_2(Q_2)$
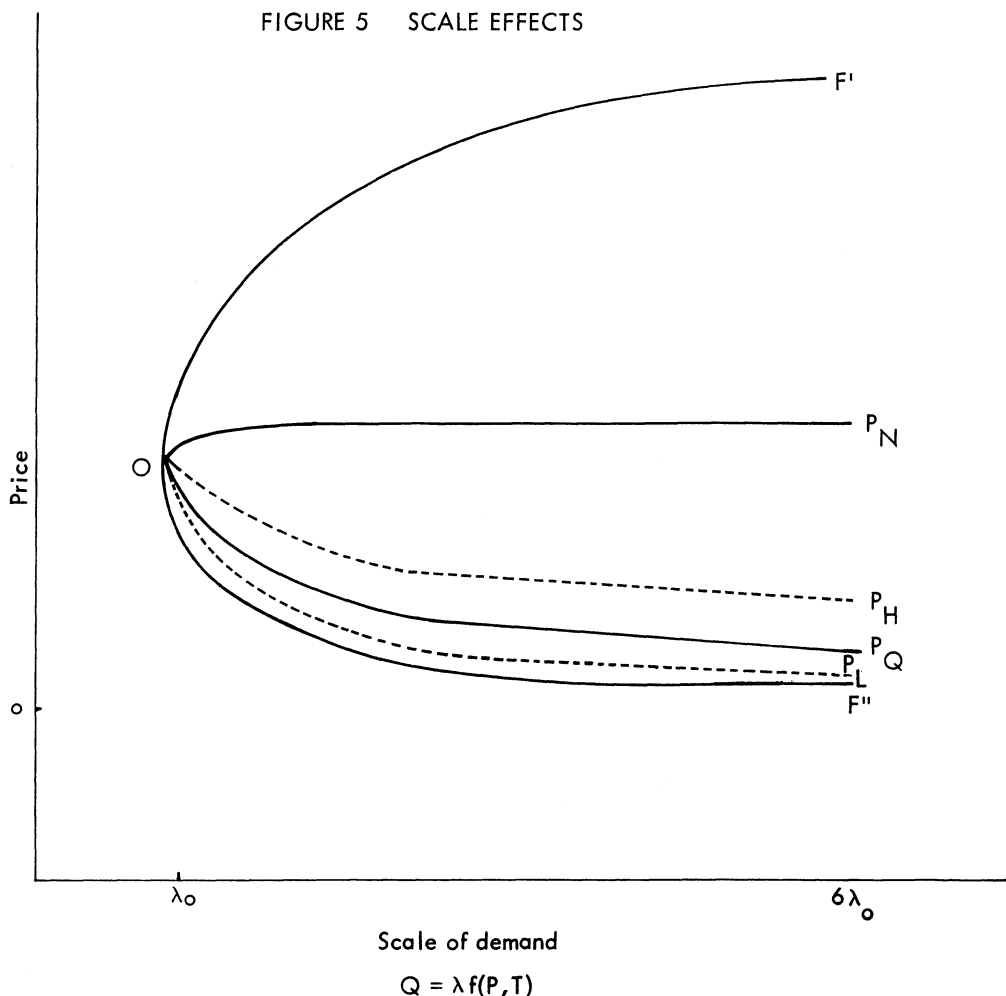
W max occurs where:

$$\frac{dW}{dP} = Q_1 \frac{dP_1}{dP} + Q_2 \frac{dP_2}{dP} = 0$$

Q max occurs where:

$$\frac{dQ}{dP} = \frac{\partial Q}{\partial P_1} \frac{dP_1}{dP} + \frac{\partial Q}{\partial P_2} \frac{dP_2}{dP} = 0$$

125

FIGURE 5   SCALE EFFECTS

Scale of demand

$$Q = \lambda f(P, T)$$

where λ is a constant. Then there is some $\lambda = \lambda_0$, a critical point at which the system has a solution, and the feasible set is not empty. As we increase λ, we increase the density of the trip demand and increase the range of the feasible set of solutions. In Figure 5 we have illustrated the effect on various solution values discussed above as the density or scale of demand is changed. At $\lambda = \lambda_0$, the feasible set has only one value of P, labelled 0 on the graph. As λ increases, the feasible set is bounded by $F'F''$. Similarly, we have plotted the levels of the regulated price which would yield: maximum total revenues (and maximum number of taxis in service), P_N; minimum total cost for customers with high time value r_2, P_H; minimum total cost for customers with low time value r_1, P_L; and maximum number of trips, P_Q. P_H, P_Q, and P_L all approach the marginal cost, c, as λ grows large. But within the range of λ likely to be encountered P_H would remain significantly higher than P_L, so the conflict would continue to exist.

126

CONCLUSION

If quality differentiation is constrained, a unique optimum price and quality level in this or similar markets is not defined in the absence of interpersonal summation of consumers' benefits. The differential impacts of regulatory price strategy can be conveniently observed, in many transport markets, in terms of the effect on the cost of waiting or delay. There is reason to believe, moreover, that the price generated by the "competitive equilibrium" may be clearly inefficient, being higher than the set of prices that could be defined as efficient, given the constraints of this market. This set contains within it a price at which total use of the service is maximised, which may serve as a useful pragmatic policy goal.

University of North Carolina at Chapel Hill

127