25th Australasian Transport Research Forum
Incorporating the BTRE Transport Policy Colloquium
Canberra 2-4 October 2002

Congestion charging and the optimal provision of public infrastructure: theory and evidence

Truong P Truong and David A Hensher
University of New South Wales; University of Sydney

Abstract

The paper provides a theoretical framework for analysing the effects of public infrastructure provision on private sector productivity using the example of a transport network. Public infrastructure such as a transport network is assumed to be a (congested) public good. When the provision of this good is at the long run equilibrium level, consumers pay a price which reflects the (individually-determined) marginal productivity of the good and the supplier is also recovering all its opportunity costs. In the traditional literature on transport congestion (Walters, 1961; Mohring and Harwitz, 1962), the concept of infrastructure capacity is often defined in term of the maximum level of traffic *flow*, which is more of a usage concept rather than a 'capacity' concept. Congestion is then defined in terms of an increase in the marginal social cost of this traffic *flow* over and above the marginal private costs (measured in terms of the average travel time per trip distance). Defined in this way, optimal congestion tax is seen to be limited just to the case when travel demand and traffic density is still at a relatively low level where 'bottleneck' congestion has not occurred. This paper explores an alternative definition of infrastructure 'capacity' and 'congestion' level where it is related more to the level of traffic density and travel speed, rather than to traffic flow. Defining capacity and congestion in this way, the paper opens the way for redefining the concept of an optimal 'congestion tax' or traffic toll which can be used to apply to the case of heavily congested or bottleneck situation as well as to the traditional case of 'low congestion'. The paper illustrates this new concept of congestion toll with an empirical estimation to an actual road network.

Contact author

Truong P Truong
School of Economics
University of New South Wales
NSW 2052
t.truong@unsw.edu.au

David A Hensher
Institute of Transport Studies
University of Sydney
NSW 2006 Australia
Davidh@its.usyd.edu.au

GO BACK TO START SCREEN GO BACK TO INDEX

Introduction

The role that public infrastructure investment can play in increasing private sector productivity is a burgeoning area of research. Although there have been many studies which look at this issue (See Aschauer (1988, 1989a, 1989b), Berndt and Hansson (1981), Nadiri and Mamuneas (1991), Alesina *et al.* (1991), Dixon and McDonald (1991)), these studies often do not take into account the *public* (or semi-public) nature of infrastructure goods.

One of the difficulties encountered in analysing the effects of a *public* good investment on private sector productivity is that to determine the optimal level of provision of a public good, it is essential that we equate, not the individual, but the *sum*, of all these individualised marginal willingness-to-pay for (marginal productivities of) the good, with its marginal opportunity cost. The information on this is not readily observable due to the problem of preference non-revelation.

Fortunately, in the case of a 'congested' public good, such as a transport network, congestion can act as though a kind of 'implicit tax' which individual users of the public infrastructure have to incur and hence, at equilibrium these implicit congestion taxes can act as 'Lindahl prices' which will entice the individual users of the congested good to reveal their true preferences about their marginal willingness-to-pay for the public good. If we can estimate the level of the implicit taxes from the level of congestion and the aggregate level of demand associated with this level of congestion, then we can use these to estimate the (*aggregated*) Lindahl prices to determine the optimal level of provision of the public good. Congestion, in other words, can act as a kind of 'invisible hand' which helps to restore equilibrium in the case of a public good. We illustrate this with an empirical calculation for an actual road network.

Public Infrastructure as a Congested Public Good

Let G be the stock or *capacity* of a public infrastructure asset available for use and G_i be the 'effective' level of utilisation of this capacity by user *i.* If G is a pure public good, then by definition:

$$G_i = G, \quad i = 1,..., n \tag{1}$$

where n is the total number of users (that is, every user has unimpeded equal access). On the other hand, if G is a pure private good, then we have instead:

$$\sum_{i=1}^{n} G_i = G. \tag{2}$$

In the general case when *G* is an impure (partially congested) public good, we have:

$$G_i \le G, \text{but} \sum_{i=1}^{n} G_i \ge G. \tag{3}$$

As Oakland (1987) pointed out, congested public goods can be treated as though equivalent to a combination of congestion externalities and a pure public good which relates to total system capacity. Thus, if $f^i(.)$ stands for the production function of user *i*, then we have:

$$f^i = f^i(L_i, K_i, G_1, ..., G_n, G) \tag{4}$$

where L_i and K_i stand for the private (labour, capital)[1] inputs, G_j is the effective level of utilisation of public infrastructure by user *j,* and G is the total system capacity (eg lane kilometres). We have:

$$\P f^i / \P G_i > 0; \quad \P f^i / \P G_j < 0, \quad \text{for } i \ne j \tag{5}$$

which says that *j*'s utilisation of the system capacity would contribute to congestion and therefore would impact negatively on user *i*'s marginal productivity. An alternative specification for a congested public good is to assume that each effective utilisation of the good is given by the user's own utilisation rate and an overall level of congestion *q*:

$$f^i = f^i(L_i, K_i, G_i, q) \tag{6}$$

where

$$q = q(\sum_{i=1}^{n} G_i, G). \tag{7}$$

We refer to equation (7) as the congestion function. Pareto optimal allocation of resources in the case of a congested public good can now be found by solving the following optimisation problem:

$$\begin{aligned} &\text{Max } f^i(L_i, K_i, G_i, q) \\ &\text{s.t. } f^j(L_j, K_j, G_j, q) \ge X^j, \ \ j \ne i, \ \ j = 1, ..., n. \\ &G_i \le G, \quad i = 1, ..., n. \\ &F(\sum_{i=1}^{n} K_i, G) \le 0. \end{aligned} \tag{8}$$

[1] These inputs would include a driver's time and the capital cost of a car for a road use context.

where X^j is the output or activity level associated with user *j*, and $F(.)$ is the transformation function between total private capital investment ($\sum_{i=1}^{n} K_i$) and total public capital goods production.

The Lagrangian for this optimisation problem is:

$$L = \sum_{i=1}^{n} l^i [X^i - f^i(L_i, K_i, G_i, q)] + \sum_{i=1}^{n} a^i (G_i - G) + mF(\sum_{i=1}^{n} K_i, G) \qquad (9)$$

with $l^i = -1$ and $X^i = 0$ for the reference *i*. The efficiency conditions with respect to G_i are as follows (assuming that $G_i < G$, and therefore $a^i = 0$ for all *i*).

$$-l^i f^i_{K_i} + mF_P = 0, \qquad (10)$$

$$-l^i f^i_{G_i} + \sum_{j=1}^{n} l^j f_q^j q_1 = 0, \qquad (11)$$

$$-\sum_{j=1}^{n} l^j f_q^j q_2 + mF_G = 0, \qquad (12)$$

where

$$\begin{aligned} &f^i_x = \partial f^i / \partial x, \quad x = \{K_i, G_i, q\}, \\ &q_1 = \partial q / \partial \left(\sum_{i=1}^{N} G_i \right), \; q_2 = \partial q / \partial (G), \\ &F_P = \partial F / \partial \left(\sum_{i=1}^{N} K_i \right), \; F_G = \partial F / \partial G. \end{aligned} \qquad (13)$$

Equations (10) and (11) can be combined to give:

$$(f^i_{G_i} / f^i_{K_i}) + q_1 \sum_{j=1}^{n} (f_q^j / f^j_{K_j}) = 0, \qquad (14)$$

and equations (10) and (12) can be combined to give:

$$q_2 \sum_{j=1}^{n} (f_q^j / f^j_{K_j}) = (F_G / F_P), \qquad (15)$$

Equations (14) and (15) can also be combined to give:

$$(f^i_{G_i} / f^i_{K_i}) + (q_1 / q_2)(F_G / F_P) = 0. \qquad (16)$$

If increasing the aggregate utilisation rate and capacity by the same proportion would leave congestion unchanged (i.e. the function $q(.)$ is homogeneous of degree zero in its arguments), then we have:

$$q_1 \sum_{i=1}^{n} G_i + q_2 G = 0. \tag{17}$$

Multiplying (16) by G_i and summing over all *i*'s using (17), we have:

$$\sum_{i=1}^{n} (f_{G_i}^i / f_{K_i}^i) G_i = G(F_G / F_P) = GP_G, \tag{18}$$

where P_G = (F_G/F_P) is the shadow price of public capital in terms of private capital foregone.

Equation (18) is a special case of the Samuelson condition for the optimal provision of a congested public good[2]. The ratio $(f_{G_i}^i / f_{K_i}^i)$ stands for the marginal productivity of a public capital good relative to that of a private capital good for user *i*. If we adopt the benefit principle of taxation then each individual user should be charged an individualised (or Lindahl) price for the effective utilisation of public infrastructure as follows:

$$\begin{aligned} P_G^i &= (f_{G_i}^i / f_{K_i}^i) \\ T_G^i &= P_G^i G_i \end{aligned} \tag{19}$$

where P_G^i is the per unit price for public infrastructure capacity G and where T_G^i is the total contribution from user *i* towards the total public infrastructure capacity costs[3]. Using (16), we have:

$$\begin{aligned} P_G^i &= -(q_1/q_2)(F_G / F_P) = -(q_1/q_2) P_G \\ T_G^i &= -(q_1/q_2) G_i P_G \end{aligned} \tag{20}$$

Application to a Transport Network

To operationalise the model, we need an empirical specification of the congestion function (7). One specification of this function is the form used in many public infrastructure studies such as the one proposed by Shah (1992):

---

[2] See Oakland (1987, p. 501).

[3] In practice when a price is charged for the use of a public infrastructure, this would consist of both a 'usage' component to cover the short run operating and maintenance costs and a 'capacity' (or capital) component to cover the rental price of infrastructure capital. Here we are concerned only with the capacity component.

$$G_i = G.(I_i)^q \qquad (21)$$

Here, I_i stands for an 'index of use' (of the road capacity) by user *i*, and *q* is a 'parameter indicating the degree of publicness of public infrastructure'[4]. In the case of a road network, *q* can be used to indicate the 'degree of congestion' on the road[5]. *Capacity* utilisation level G_i can be indicated by the *speed* of user *i*'s vehicle. Thus, if there is no congestion ($q = 0$), every vehicle can travel at the maximum speed G allowed by the capacity[6] of the road. When there is some congestion on the road ($q > 0$), then speed will depend on the index of use as well as the degree of congestion. For example, if there are n users of the road (n can be related to the traffic volume, or density, on a particular road segment) and assuming that the index of use is the same for all vehicles, then we can construct an index of use $I_i = (1/n)$ for all *i*'s. In this case, the (relative) index of capacity utilisation for each vehicle (user) will be $G_i/G = (1/n)^q$ which implies $(G/n) < G_i < G$ for $(0 < q < 1)$. The situation when $q = 1$ and $G_i = (G/n)$ can be described as 'full' or 'complete' congestion. In this limiting situation, any percentage increase (or decrease) in the traffic density will be matched exactly by the same percentage decrease (or increase) in the average speed of all remaining vehicles, and this can be explained as follows. When congestion is 'full', road space is considered as though a pure private good, and is completely 'rival'. This means one vehicle's 'consumption' of road space must be at the full expense of another vehicle's, and this follows from the condition that the total level of consumption of road space by all vehicles must remain a constant. Since 'speed' is the rate of consumption of road space per unit of time, we have the product of average speed and the number of vehicles equal to the total level of consumption of road space per unit of time. Since the latter is a constant, any percentage increase/decrease in the number of vehicles must be matched exactly by the same percentage decrease/increase in the average speed.

There are advantages in defining capacity (and capacity utilisation) of the road in terms of speed rather than in terms of traffic 'flow' [7] (vehicles per lane per hour), or in terms of traffic 'density' (vehicles per lane per km). If we use an

---

[4] Shah (1992, p. 29).

[5] Shah referred to *q* as the degree of 'publicness' of the infrastructure good, but in fact, it can be seen that the greater the value of *q*, the *smaller* would be the value of G_i. Hence, it would be more appropriate to refer to *q* as the degree of *non*-publicness (or 'rivalness' in consumption) of the public infrastructure good, and in the case of a road network, the degree of 'rivalness in consumption' is in fact the degree of 'congestion' in the network.

[6] In practice, the maximum speed G is determined not only by the physical capacity of the road, but also by traffic regulation and safety standards.

[7] In the literature on congestion modelling, there has been a heated debate about whether it makes sense to define supply and demand for travel in terms of traffic *flow*. One difficulty is the fact that a single level of traffic flow can correspond to two different levels of traffic *congestion* (i.e. there is a backward-bending supply curve for traffic flow – Lindsey and Verhoef, (2000)). Traffic flow, therefore, is not an unambiguous measure of traffic congestion. Speed, on the other hand is related uniquely to congestion, and hence can be used as an indicator of capacity utilisation level (which is also a measure of congestion).

analogy with electricity generation, 'capacity' in electricity generation is defined as the *rate* of electricity supplied per unit of time (kW), while 'usage' is defined as the total *quantity* of electricity consumed over a period of time. Usage is thus equal to 'capacity * time period' (kWh). Similarly, for a road network, capacity can be defined in terms of the maximum rate of 'supply' of road space per unit of time to each vehicle (speed), while road *usage* can be defined as the amount of road space 'consumed' by each vehicle over a period of time, i.e. usage (distance travelled by a particular vehicle) = capacity (speed) * time period.

There is also an alternative way of looking at equation (21). Instead of looking at *speed* (which represents capacity), we can now look at its inverse, which is the average *travel time* on the road, as follows:

$$t_i = t_0.(n)^q \qquad (22)$$

Here, t_i = (1/G_i) is the average or 'effective' travel time (per kilometre) on the road when there is congestion (at level q and with n users), and t_0 = (1/G) is the free-flow or zero-congestion travel time. From equation (22), we can derive:

$$e_n^t = \left.\frac{d\ln(t_i)}{d\ln(n)}\right|_{t_0=(1/G)=constant} = q \qquad (23)$$

where $d\ln(x)$ is used to denote the change in the logarithm of variable x, or the percentage change in variable x, where $x = \{t_i, n\}$, and (e_n^t) is therefore the elasticity of travel time with respect to an increase in the number of users. Equation (23) says that this elasticity is given exactly by the congestion measure. This gives us another useful interpretation of the congestion measure q.

Equation (21) can now be used to define (or 'calibrate') the level of congestion q in terms of the relative rate of capacity utilisation (G_i/G)[8] and the index of use (1/n). To illustrate this, consider a hypothetical (maximum) speed-versus-traffic density relationship as shown in Figure 1[9]. From this speed-density diagram,

[8] Or (t_0/t_i).

[9] See, for example, Lindsey and Verhoef (2000) figure 1, page 355. Note that we are considering only the *maximum achievable speed* for any given traffic density because in reality, the actual speed can be influenced by many factors (bad weather, accidents, vehicle breakdowns, driver's behaviour, etc.) rather than just the 'normal' congestion factor (i.e. traffic density). Therefore, to isolate this particular relationship between speed and traffic density which reflects congestion, we must use the maximum achievable speed to filter out these other factors (see the section on "Empirical Application to an Actual Road Network' below). The 'speed-density' diagram which we use is not simply a characterisation of 'demand', but rather a picture of how demand (traffic density level) and supply (network capacity) interact. For example, if we keep the supply curve (capacity of the road network) fixed and allow the traffic demand (traffic density level) to increase, then the equilibrium 'price' of the traffic flow will also increase, and this is reflected in the increased level of congestion. This will then act to decrease the level of "effective" capacity utilisation even if the actual level of capacity remains

choose the 'reference' density level n_0 such that at that density, we regard congestion as still being zero. In reality, the definition of n_0 is conditional on the definition of the maximum free-flow speed G_0, and since the latter is dependent, not only on the *physical* capacity of the road, but also on traffic regulations, on drivers' behaviour, etc., the determination of G_0 and n_0, therefore, cannot be left entirely dependent on the physical capacity of the road. To determine the level of G_0, thus, we need to consider, not only the existing legal speed limit on a particular road, but also the existing physical characteristics. Once the level of G_0 is determined, however, we can use this to define n_0: it is the level of traffic density such that, *above* this level, the maximum achievable speed cannot be greater than G_0. Below this traffic density level n_0, of course, there may be *observed* achievable speeds which are greater than n_0 (e.g. drivers exceeding the speed limit, and/or driving at an unsafe speed). However, for the purpose of defining 'congestion', we will ignore these 'outliers'. The congestion level for traffic density less than or equal to n_0, thus, will by definition be zero. Congestion will start to increase from zero to a positive value when traffic density n exceeds n_0 because the maximum achievable speed will start to decline below G_0 (see Figure 1). We can now also re-write equations (21) and (22) in an alternative ('indexed', or 'relative' form) as follows:

$$G_i = G_0.(n_0 / n)^q \tag{24}$$

$$t_i = t_0.(n / n_0)^q \tag{25}$$

where n_0 is the maximum 'free-flow density' of the road and can also be used as an alternative definition of the 'capacity' of the road[10].

From equation (25), we can also rewrite equation (23) in an alternative form as follows:

$$e_n^t = \left.\frac{d\ln(t_i)}{d\ln(n)}\right|_{G_0, n_0=\text{constant}} = q \tag{26}$$

Equation (26) says that keeping 'capacity' G constant (as is implied by equation (23) will require that *both* G_0 and n_0 must remain constant. In some cases, G_0 can change without any change in the *physical* capacity of the road (e.g. a change in the legal speed limit without any change in the physical conditions of

---

the same. Traffic density therefore is used to represent, not a demand for 'traffic' or trips as such, but rather a demand for system capacity. This represents a departure from conventional approach where the focus is on 'traffic' or 'trips' rather than on capacity of the road.

[10] n_0 is to be 'calibrated' from observed data. Comparing equation (25) with an alternative formulation of the (congested) travel time function, such as that used by the Bureau of Public Roads: $t = t_0\,[1+0.15(Q/C)^4]$, where Q is the traffic volume (vehicles/hour), and C is 'capacity' (also vehicles/hour travelling between two nodes of a link) (see, Ozbay *et al.* (2001), p. 81), we note that the ratio (Q/C) plays a role similar to that of the ratio (n/n_0) in equation (25).

the road). In which case, n_0 will need to be re-calibrated be consistent with the new definition of G_0. In other cases, it is quite possible that the legal speed limit G_0 will remain the same, despite an improvement in the physical capacity of the road. In this case, the speed-density curve will have shifted to the right and this implies a new level of n_0. The new level (n^*_0, as shown in Figure 2) can now be regarded as the new (and improved) capacity of the road. If, on the other hand, we want to keep the definition of n_0 constant, then the change in physical capacity of the road must be reflected as a change in the free-flow speed (from G_0 to G^*_0.as shown in Figure 2). In practice, of course, we can also represent the change in the physical condition of the road as *both* a change in G_0 as well as a change in n_0.

Figure 1. Speed versus density for a hypothetical road segment

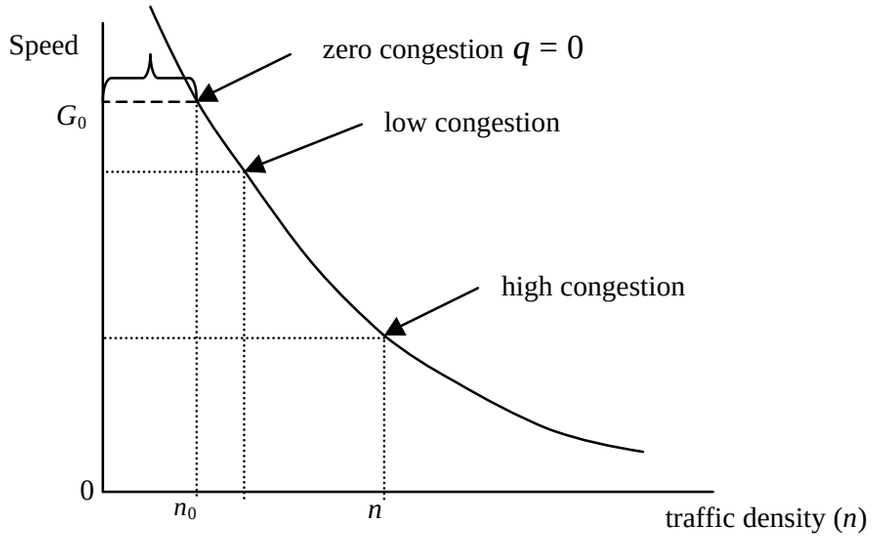
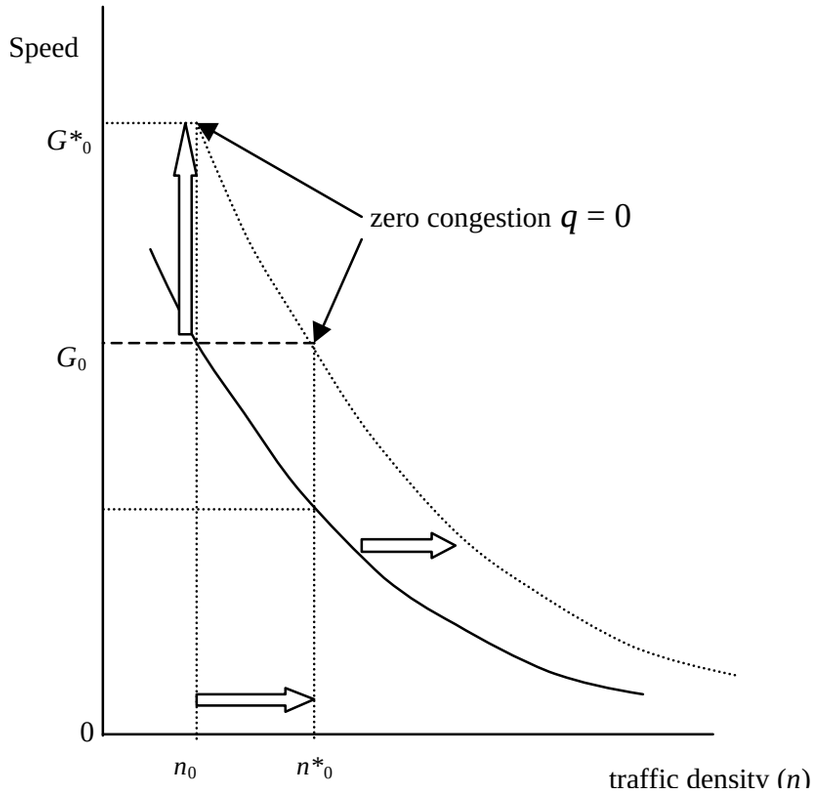


Figure 2. Effect of an increase in the *physical* capacity of the road



If we substitute I_i = (1/*n*) for all *i*'s into equation (21), summing over all *i's*, taking the logarithm of both sides and then re-arranging terms, we can re-write equation (21) in an alternative form which shows the congestion level as a function of total capacity and total capacity utilisation as follows:

$$q = 1 - \frac{1}{\ln n}[\ln(\sum_{i=1}^{n} G_i) - \ln G)] \tag{27}$$

or alternatively:

$$n^{1-q} = (\sum_{i=1}^{n} G_i)/G \tag{28}$$

Using (27), we can derive:

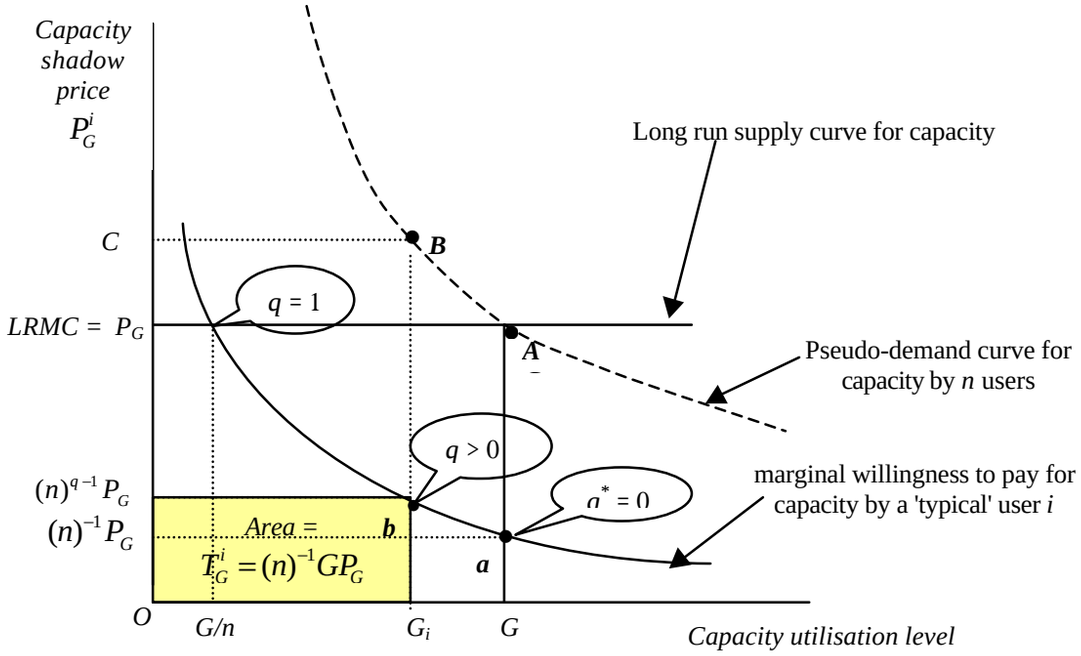
$$\begin{aligned}(q_1/q_2) &= [\partial q/\partial(\sum_{i=1}^{n} G_i)]/[\partial q/\partial(G)] \\ &= -G/(\sum_{i=1}^{n} G_i) \\ &= -(n)^{q-1}\end{aligned} \tag{29}$$

Substituting this into (20), we obtain:

$$\begin{aligned}P_G^i &= (n)^{q-1} P_G \\ T_G^i &= (n)^{-1} G P_G\end{aligned} \tag{30}$$

Equation (30) shows the rate of 'trade-off' between an increase/decrease in the total level of capacity utilisation by all users and the system capacity (G) to keep the level of congestion constant. This represents the 'shadow price' or opportunity cost of (total) private capacity utilisation in terms of the public good (system capacity) forgone. When congestion is 'full' (i.e. q = 1 and capacity becomes a pure private good) the rate of trade off is one-for-one, and the opportunity cost of (private) capacity utilisation is equal to 100% of the shadow price of system capacity, i.e P_G^i=P_G as is expected. When the congestion level is zero, however, and system capacity is a pure public good, all users can share the system capacity without diminishing the level of utilisation of one another, and the opportunity cost of system capacity to each user is equal to (1/*n*)P_G. In general, when 0<q <1, the opportunity cost of G_i should be somewhere between these two extreme cases and is given by equation (30). This can also be illustrated as in Figure 3.

Figure 3. Capacity Utilisation Level and Capacity Shadow Price



In Figure 3, we assume that the long run marginal cost (*LRMC*) of system capacity is a constant[11] at P_G. The long run supply curve for capacity is then a horizontal straight line at *LRMC* = P_G. Assume also that there are n users of the road in the long run. From the equation for P_G^i (which shows the individual marginal willingness-to-pay for effective capacity utilisation by a 'typical'[12] user *i*), we can 'scale this up' to get a (pseudo) demand curve for capacity by all indexed users of $(n\,P_G^i) = (n)^q\,P_G$. Given these demand and supply conditions, equilibrium is then established in the long run when capacity is settled at G and congestion level is settled at $q^*=0$[13]. Each user then pays a total capacity charge of $T_G^i=(1/n)(GP_G)$ and the supplier recovers fully the total capacity cost of (GP_G). This is the long run equilibrium situation. From this analysis, it can be seen that the *total* level of contribution by each user *i* to system capacity costs, T_G^i (the shaded area in Figure 3)[14], is independent of the level of congestion and

[11] This is not essential, but made for simplicity.

[12] If users are different in their marginal willingness-to-pay for capacity, then we use the marginal willingness-to-pay for capacity of the *marginal* user (assuming uniform pricing for all users).

[13] This assumption is made here for simplicity and can be relaxed. For example, if we assume that $q^*> 0$ in the long run, then long run capacity will settle at a level $G^* < G$.

[14] Which strictly relates to G_i and not G but as a hyperbola curve, the two rectangles positioned at "*a*" and at "*b*" are equal

therefore, the total contribution by *all* users to system capacity costs is also independent of the level of congestion, (area $OCBG_i$ = area OP_GAG in Figure 3) if the equilibrium condition (30) is to be followed. This implies that at equilibrium the supplier can always recover fully the total system capacity costs, and each individual (marginal) user is also paying exactly for the total *effective* level of capacity utilisation (rather than the actual *supplied* level). That is, T_G^i (the shaded are in Figure 3) corresponds exactly to the total willingness to pay by each (marginal) user for capacity up to the effective level of capacity, G_i, rather than the total supplied level G. This makes sense because, in reality, the level of G_i is unknown to the supplier, and therefore, capacity is often charged by the fixed total amount T_G^i rather than by the 'per unit' shadow price P_G^i. Given that each user is going to be charged this fixed total amount (the toll charge) *irrespective of the level of congestion*, each will consider matching this total charge with the total willingness-to-pay for effective capacity. Those users who can will remain on the road, while those who cannot will try to switch to non-tolled routes (or change travel mode, etc.). This means, at equilibrium, the total willingness-to-pay for effective capacity of the marginal user will be just equal to the total charge, and the supplier's total revenue is also equal exactly to the total capacity costs (area $OCBG_i$). This is the equilibrium condition implied by equation (30). In the next section, we will examine situations in the short run where deviations from this long run equilibrium may occur.

Short Run Effective Utilisation of a Congested Public Good

Up to now we have implicitly assumed that traffic demand level (n) and especially capacity supply (G) are at their 'long run equilibrium' level (this is implied by equation (18) from which equation (30) was derived). We now consider the situation when their short run levels may deviate from these long run equilibrium levels. This will enable us to establish the short run effective capacity utilisation charge that is consistent with the revenue that has to be raised from users to recover the short run full cost of infrastructure (including maintenance) and perhaps also to raise revenue for long term investment in capacity (the latter being the Fully Allocated Cost Method). This is commonly the case with toll roads that are part of a transport network and which are unlikely to be at their optimal level (as determined by the shadow price of capital throughout the economy) in the short run.

In the 'short run', assume that total private capital $\left(\sum_{i=1}^{n} K_i\right)$ and public infrastructure capacity G are fixed at levels which may be different from their long run equilibrium values. Furthermore, we can also assume that although *collectively* all users can influence the level of congestion q (see equation (7)), individually, the effect of a single user's action on q can be considered to be negligible. This means the values of q, q_1 and q_2 can also be considered as 'given' from an individual user's point of view.

Consider the long run equilibrium situation as described in Figure 3 of the last section again. Assume now that there are two different types of short run disequilibrium situations:

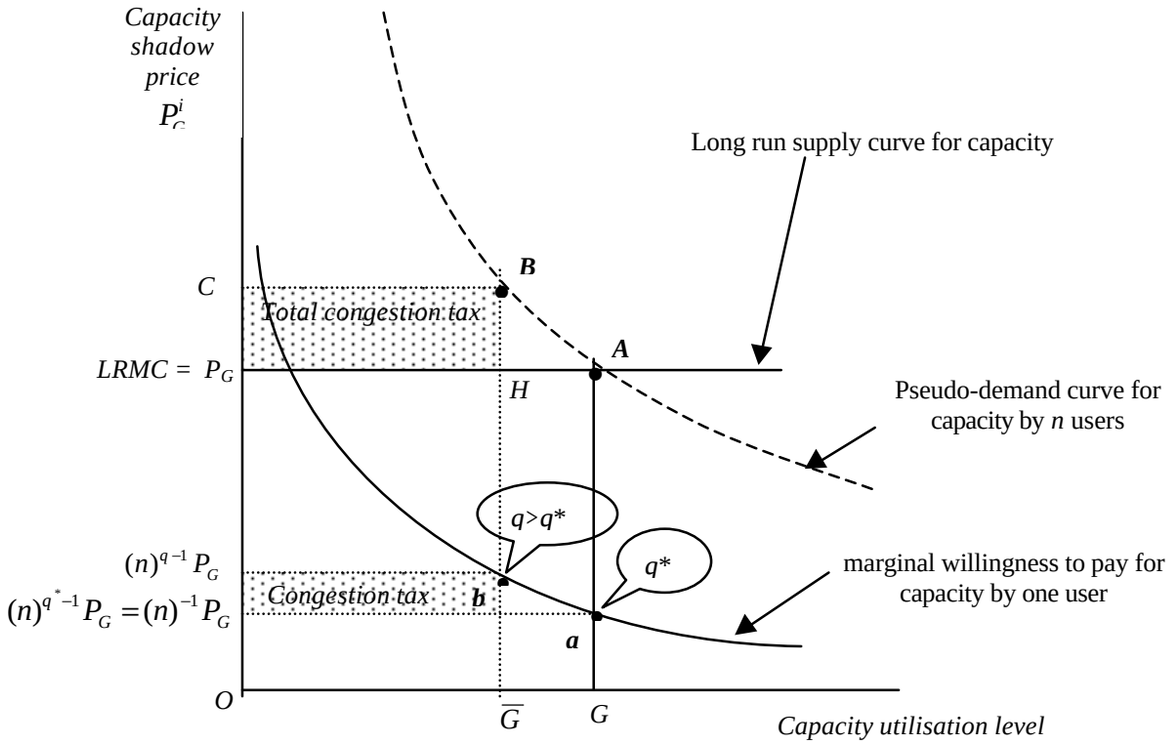
(i) Short run actual capacity $\overline{G}$ is less than long run equilibrium level G, but indexed traffic density level remains at n.

(ii) Short run traffic density level is now at $m > n$, but road capacity remains at G.

In the first situation (Figure 4), congestion will have to increase, from the 'long run equilibrium' level at point a (assumed to be $q^* = 0$) to a short run equilibrium level of $q > q^*$ at point b. At point b, congestion acts as a kind of 'invisible tax' on the user to increase the marginal cost of capacity utilisation from $(n^{q^*-1}P_G)$ to $(n^{q-1}P_G)$, and reduce the level of capacity utilisation from (G) to $(\overline{G})$. The total (implicit) congestion tax paid by each user is then given by: $(n^{q-1} - n^{q^*-1})\overline{G}\,P_G$ (the bottom shaded area in Figure 4). If the supplier now also acts as though collecting this implicit tax (by levying a capacity charge based on the long run equilibrium level of capacity (G) even though now, the short run *effective* level of utilisation by each user is only $\overline{G}$), then the supplier will accumulate a revenue surplus of $(n^q - n^{q^*})\overline{G}\,P_G$ (the top shaded area $CBHP_G$ in Figure 4)[15]. Note that, this surplus is exactly equal to the difference between the actual capacity charges levied on all users of $(n^q\overline{G}\,P_G)$ and the actual capacity costs of $n^{q^*}\overline{G}\,P_G = \overline{G}\,P_G$. This surplus, therefore, can be used to expand capacity, say from $(\overline{G})$ to (G) if congestion is to be reduced from q to q^* in the long run (assuming that traffic density level is to remain at n).

[15] Levying a capacity charge based on long run level of capacity G rather than the short run level implies an individual capacity charge of $n^{q^*-1}GP_G$. From equation (30), this is also equal to $n^{q-1}\overline{G}\,P_G$.

Figure 4. Short Run Capacity Supply constraint

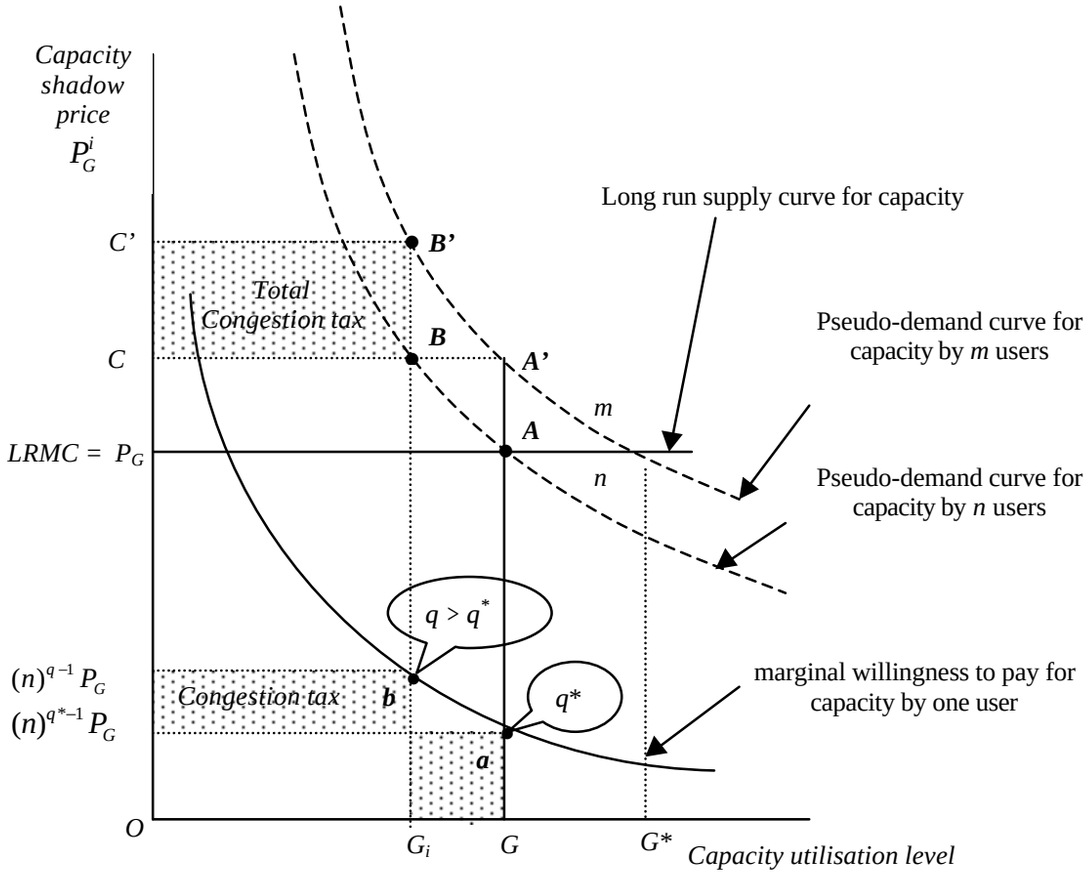


In the second situation where traffic density increases and capacity remains fixed (Figure 5), congestion will also increase, from the 'long run equilibrium' level of $q^* = 0$ at point a to a short run equilibrium level at point b where $q>q^*$. Again, we see that short run congestion acts as a kind of 'invisible tax' on each user to increase the marginal cost of capacity utilisation from $n^{q^*-1}P_G$ to $n^{q-1}P_G$, and thereby reduces the level of capacity utilisation from G to $G_i = G(n)^{-q}$. The total implicit 'congestion tax' paid by each user in this case is thus equal to: $(n^{q-1} - n^{q^*-1})(G_iP_G)$ (the left-bottom shaded area in Figure 5). This is the same as in case (i) of Figure 4, except that here, actual capacity supply is at G rather than at G_i, and therefore, actual capacity cost is GP_G rather than G_iP_G. If the supplier collects this (implicit) congestion tax from n users only, then this is just sufficient to pay for the actual 'extra' capacity supplied $(G - G_i)$ even though this is not effectively utilised by the users due to congestion (the right-bottom shaded area in Figure 5). However, since there are $m > n$ users, the extra amount of implicit congestion taxes collected from these extra number of users, i.e. $(m - n)$ $[(n^{q-1} - n^{q^*-1})(G_iP_G)]$ (the top shaded area in Figure 5) will represent the amount of revenue surplus collected by the supplier over and above the actual capacity costs. Again, this surplus can be used to expand capacity beyond G, say to G^*, if demand is assumed to remain at m in the long

run. However, if the increased traffic density level is assumed to be only a short run phenomenon, then the extra revenue collected from these additional traffic can be seen as a short run 'congestion tax' device to 'ration' demand back to the long run equilibrium level n.

In reality, there can be a mixture of cases (i) and (ii), and therefore, the implicit or invisible 'congestion taxes' levied on the road users when congestion occurs can act both as a short run price rationing device, and/or a long term means for raising revenue to expand capacity.

Figure 5. Short run disequilibrium traffic demand



Empirical Application to an Actual Road Network

The first task in empirically identifying appropriate short run user charges is to determine the level of congestion. We have collected a sample of 3,730 road segments (or links) of various types in the Sydney Metropolitan Area for 2001[16]. For each link, we obtained information on the link type (arterial, highway, expressway, freeway, etc.), link length (kms), number of lanes, vehicle density (vehicles per lane per km), travel time and speed, for different time periods of day (AM, Mid-day, PM) and night time (Nite). We have selected freeway conditions only since they relate most appropriately to tollroad settings, the focus of this paper. From this data we first plot the information on vehicle speed versus vehicle density for various times of day. This is shown in Figure 6. From this speed-density scatter diagram, we observe that there is a definite (negative) relationship between the *maximum* speed achievable at any level of traffic density and the actual traffic density as hypothesised earlier in Figure 1. Such a relationship, however, will start only when the traffic density level reaches a certain 'minimum' level (called n_0 in Figure 1). In practice, we can let this level n_0 to be determined econometrically by regressing the values of $\ln(G_0/G_n)$ against the values of $\ln(n)$ and find the value of $\ln(n_0)$ when $\ln(G_0/G_n)$ reaches the value of zero (for any assumed value of G_0). If the level of G_0 is chosen too high, then there will not be a sufficient number of observations on the values of $\ln(G_0/G_n)$ which are close to zero to estimate the value of $\ln(n_0)$ accurately. Therefore, in practice, we will first choose a level of G_0 which is sufficiently low (say, 80 km/h) to allow for a sufficient number of observations of $\ln(G_0/G_n)$ which are close to zero. Then, having estimated the relationship between $\ln(G_0/G_n)$ and $\ln(n)$ based on the assumed value of G_0 which may be lower than the actual speed limit on the road (say G^*), we can now take into account this fact by adding to the *estimated* value of $\ln(G_0/G_n)$ a constant term $\ln(G^*/G_0)$. This is because[17] $\ln(G^*/G_n)$. = $\ln(G^*/G_0)$. + $\ln(G_0/G_n)$. Note, however, that this will also involve a re-estimation of the value of $\ln(n_0)$ since the newly estimated function $\ln(G^*/G_0)$ will cut the horizontal axis at a different value of $\ln(n_0)$. This is illustrated in Figure 7. Here, the estimated value of $\ln(n_0)$ is 2.59 for G_0 = 80 km/h, but changes to 0.5 for G_0 = 105 km/h. Note that without using this indirect method, it may not be possible to estimate the value for $\ln(n_0)$ for G_0 = 105 km/h because there will be no actual observations of $\ln(G^*/G_n)$ which are close to zero when G_0 is set to 105 km/h.

In Figure 8, we plot the empirically estimated relationship between $\ln(G_0/G_n)$ and $\ln(n_0)$ (using equation (24)) and compare this with the Bureau of Public Roads (BPR) formula: $\ln(G_0/G_n) = \ln(t_n/t_0) = [1+0.15(n/n_0)^{\alpha}]$, where α is given the values of α =1 and α = 4. It can be seen that when α =1, the BPR formula

[16] Data was purchased from the Transport Data Centre (within the New South Wales Department of Transport).

[17] In practice, if there are sufficient number of observations of $\ln(G^*/G_n)$ close to zero (i.e. sufficient number of G_n close to G^*), then we may simply choose G_0 as being equal to G^* so that there will not need to be any adjustment to the estimated value of $\ln(G_0/G_n)$.

comes close to the observed values and our predicted values. However, when $a = 4$ (the value of a in the original BPR formula), the BPR formula gives values which are significantly different from both the observed values and our predicted values. This illustrates the advantage of using an equation such as (24) where the flexible 'congestion' parameter α can be used to fit in with a particular type of road and traffic condition, whereas the BPR formula cannot.

From the empirically estimated relationship between $\ln(G_0/G_n)$ and $\ln(n)$ (as shown in Figure 7), we can now estimate the values of q for different levels of traffic density n using the formula $q = [\ln(G_0/G_n)]/[\ln(n/n_0)]$, and assuming a particular level for G_0. This is plotted in Figure 9, and from there we can see that, except for the random variation in the estimated values of q for low traffic densities (due perhaps to the lack of accurate empirical observations on the values of G_n for low values of n), the relationship between q and the traffic density level n is quite stable when n gets large (and this is also where the main interest in congestion measurement lies). Note also that the higher the assumed level of G_0 for any given set of observed speed G_n, the higher will be the estimated congestion level. In other words, 'congestion' is a relative concept. Its definition and measurement depends on the (arbitrary) setting of the 'zero congestion' (or maximum free speed) reference level.

Figure 6: Speed versus Traffic Density on Freeway

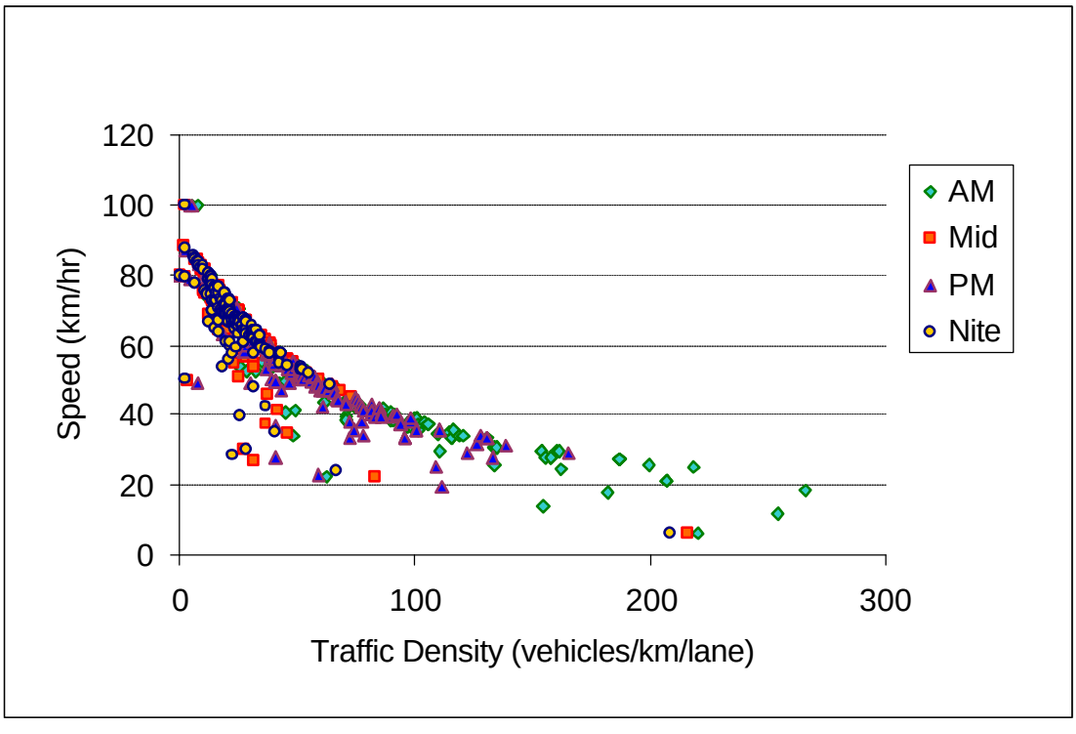


Figure 7: Estimated relationship between (the minimum value of) $\ln(G_0/G_n)$ and $\ln(n)$ for different assumed values of G_0.

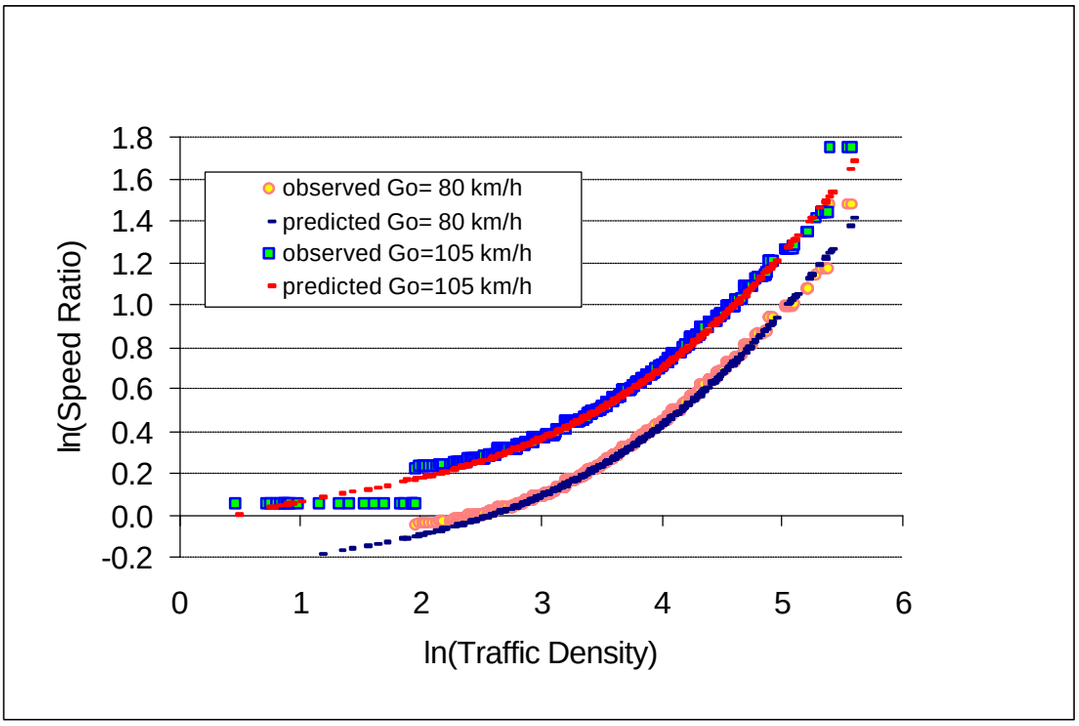
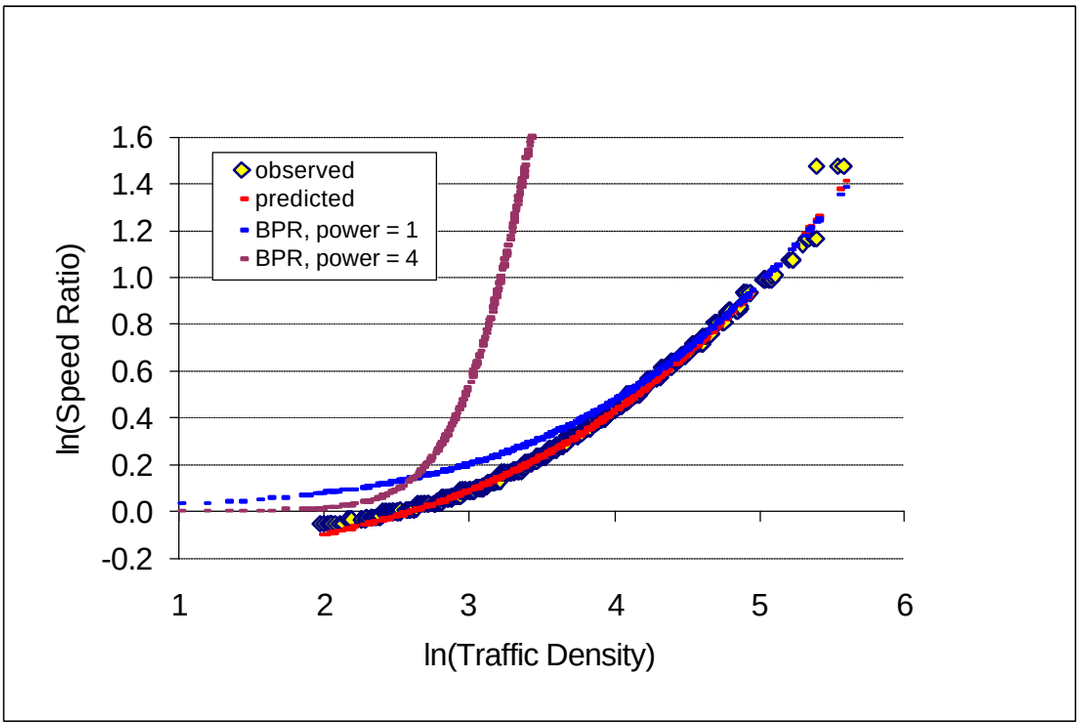
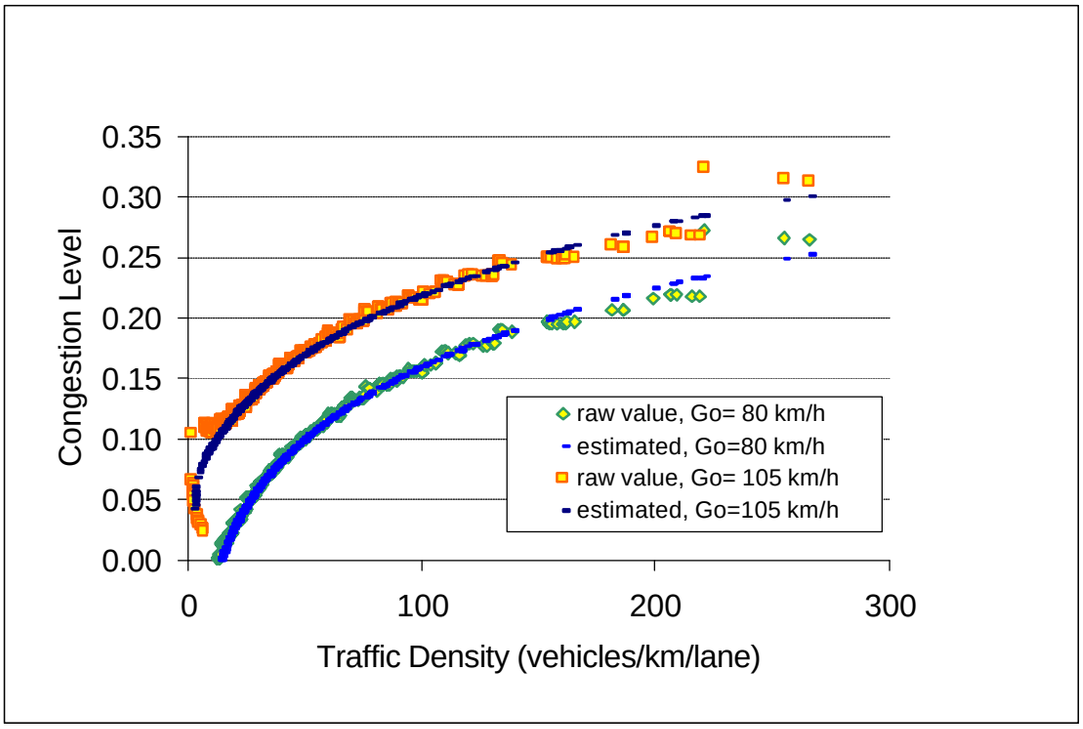


Figure 8: Comparing the predicted relationship with that implied by the Bureau of Public Road (BPR) formula (assuming G_0 is equal to



80 km/h).

Figure 9: Congestion Level versus Traffic Density on Freeway for different assumed levels of free-flow speed G_0.



Implicit or Invisible Congestion Taxes

If equation (30) is used to devise a system of capacity charges or tolls for road usage, then in the long run, not only are the users paying exactly for the total amount of capacity they can effectively make use of, but the supplier is also recovering all capacity costs (GP_G), given any assumed 'long run equilibrium' level of congestion q^*. Suppose, however, that in the short run, the actual congestion level is different from q^*, say $q > q^*$, then the user is implicitly paying *more* for capacity (because the effective level of capacity utilisation has been reduced below the long run equilibrium level). Congestion, thus, acts like an implicit or 'invisible tax' on the user, and the supplier, if allowed to collect all these implicit taxes, will be running an 'invisible' surplus. The estimation of this surplus depends on the assumption one makes about the state of short run disequilibrium situation (as explained by cases (i) and (ii) of Figures 4 and 5).

If the short run disequilibrium situation is due to a shortfall of actual capacity relative to (long run) traffic demand estimation (case (i) of Figure 4), then the 'surplus' can be estimated according to the formula: $[(n^q - n^{q^*})\,\overline{G}\,P_G]$, where q and $\overline{G}$ are the current congestion level and (assumed) maximum free flow

speed respectively, and q^* is the desired or 'long run' congestion level when 'capacity' is to be increased to G[18]. To illustrate this, assume that current 'capacity' is $\overline{G}$ =80km/h, and current traffic density is n =100 (vehicles/km/lane). From Figure 9, we can see that at this density, the congestion level will be q = 0.156 for a maximum achievable speed of G_i = 39 km/h[19]. Now, suppose we set a target of q^*=0 in the long run. To achieve this target, the (physical) capacity of the road must be expanded, and hence the speed-density curve will shift to the right, to such an extent that, at density level n, maximum achievable speed will now be equal to $\overline{G}$. The new 'capacity' of the road can now be measured *either* by an increase in the free-flow speed from $\overline{G}$ to (a maximum level of) G^*, *or* an increase in free-flow density from n_0 to n, or perhaps a mixture of the two, i.e. $\overline{G}$ to G_1, and n_0 to n_1 (see Figure 10). The implicit congestion tax which is being currently collected and which can be used for the purpose of this capacity expansion is given by the formula: $[(n^q - n^{q^*})](\overline{G}\, P_G)$. From the values obtained from Figure 9, this is equal to $[(n^q - n^{q^*})]= [(100^{0.16} - 1)]$=1.09 times the current capacity costs ($\overline{G}\, P_G$). What this implies is that if the current toll charges are just sufficient to recover all capacity costs in a long run equilibrium situation (i.e. when congestion is just about equal to zero, q^*=0, with a free-flow speed of 80 km/h) and if current congestion level is 0.156 (representing an achievable speed of around 36 km/h), then the user is implicitly paying some 'congestion tax' for this delay. If we measure this delay in terms of the extra travelling time incurred, then the percentage increase in average travelling time is [(1/36)-(1/80)]/(1/80) = 22 per cent. The value of $[(n^q - n^{q^*})]=[(100^{0.16} - 1)]$=1.09 implies that the current users of the road are willing to pay an extra 109 percent of the current toll (i.e. a little more than double the current toll) to increase speed from 36 km/h to 80 km/h (or reduce the congestion level from 0.156 to 0). Figure 11 plots the values of $[(n^q - n^{q^*})]$ for various levels of traffic density n and their corresponding congestion level q (as shown in Figure 9). These values can be referred to as implicit congestion taxes which act as an 'invisible hand' in guiding users and the supplier of the congested public infrastructure facility towards a situation of long run optimal use. The (implicit) taxes reveal the users' extra willingness-to-pay[20] for additional capacity that can reduce congestion from the current level q to a desired level q^* assumed to be zero. Given these signals, the suppliers can choose to do either (a) nothing, which implies leaving the potential demand for more capacity unsatisfied, and congestion unaltered, or (b) to invest in extra capacity using the estimated value of the 'implicit taxes' as a guide towards the optimal level of capacity to be installed in the long run. The supplier can choose to do (b) by increasing the existing level of capacity charges and using

---

[18] Note that in practice, capacity expansion can be described in terms of either an increase in maximum free-flow speed G_0 (keeping the free-flow traffic density n_0 constant) or an increase in the free-flow traffic density n_0 (keeping the maximum free-flow speed G_0 constant). We are considering here only the first case.

[19] With a traffic density level of n = 100, we have: $\ln(100)$=4.605. We derive $\ln$(speed ratio)= 0.72, or speed ratio = 2.05, which gives actual speed = 80/2.05 = 39 km/h.

[20] The existing toll levels in Sydney typically vary from $2.20 to $3.50.

the extra revenue to expand capacity[21], and so long as the implicit tax remains positive this will imply the existing capacity is less than optimal[22].

Next, we consider a different situation when the current level of congestion is assumed to be due, not to a shortfall of capacity, but to a temporary increase in the level of traffic demand over and above an assumed long run equilibrium level, the latter is determined exogenously and independently of the model. This is 'case (ii)' as referred to in the previous section and described in Figure 5. In this case, let m be the current level of traffic density and n_0 be the desired long run equilibrium level (see Figure 12). The analysis of the previous section suggested that the implicit congestion tax being paid by current users of the road is equal to $(m-n_0)\,[(n_0^{q-1}-n_0^{q^*-1})G_iP_G] = (m-n_0)\,[(n_0^{q-1}-n_0^{-1})(n_0/m)^{\hat{q}}](GP_G)$, where (GP_G) is total capacity costs. Figure 13 plots the values of $(m-n_0)[(n_0^{q-1}-n_0^{-1})(n_0/m)^{\hat{q}}$ against the values of m (and its corresponding value of q) for a given level of n_0 (using the information provided by Figure 9 and assuming that free flow speed is 80km/h). For example, if m = 130 (vehicles/km/lane) and n_0 = 13.3 (vehicles/km/lane) (i.e. about one-tenth of the current level) then from Figure 13, it can be said that the current users are being charged an implicit congestion tax equivalent to 350% of the current capacity costs. This implicit congestion tax is also a measure of the total willingness-to-pay by current users to reduce the current level of congestion from q = 0.184 to zero, and therefore, to increase the speed from 33.5 km/h to 80km/h. In this case, however, the supplier may choose not to expand capacity but rather to use the existing congestion level as an implicit tax to 'ration' demand to the most efficient users.

[21] Assuming that the level of demand n is not significantly affected by the extra charge, otherwise, we will consider the existing level of demand n as a mixture of long-run equilibrium demand (which is to be determined exogenously by factors such as population density, activity distribution, etc) and short run fluctuation in demand (which is case (ii) considered below).

[22] In this paper, we assume the long-run marginal cost of supply of public infrastructure is constant at the shadow price level P_G. This assumption, however, can be varied without altering the basic arguments of the analysis.

Figure 10. Implicit Congestion Tax for case (i): Short Run Capacity Constraint

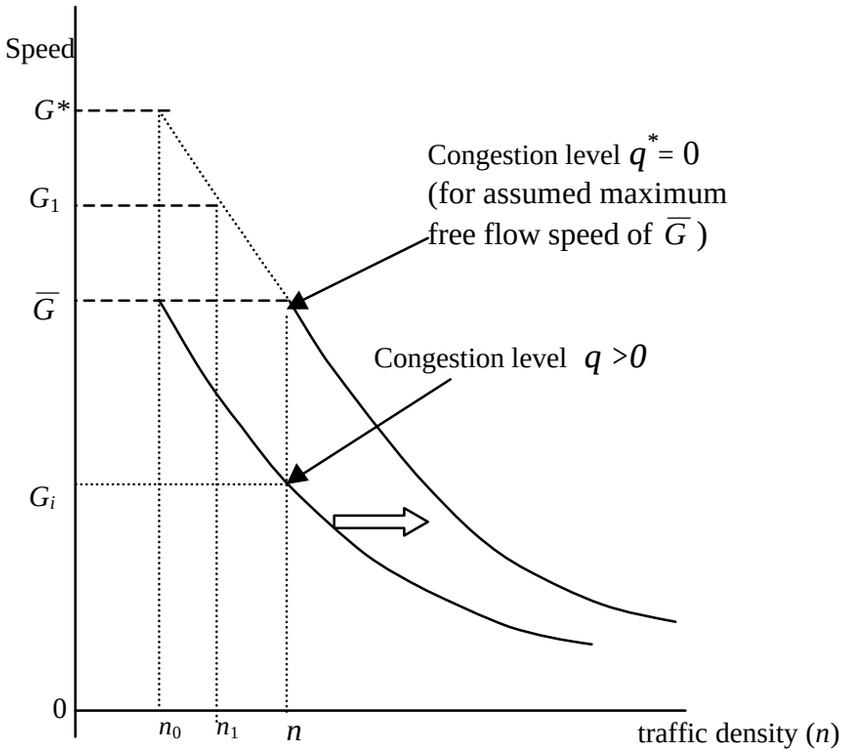


Figure 11: Implicit Congestion Tax at different traffic density levels, assuming that long run congestion level is zero and maximum free flow speed is 80 km/h.

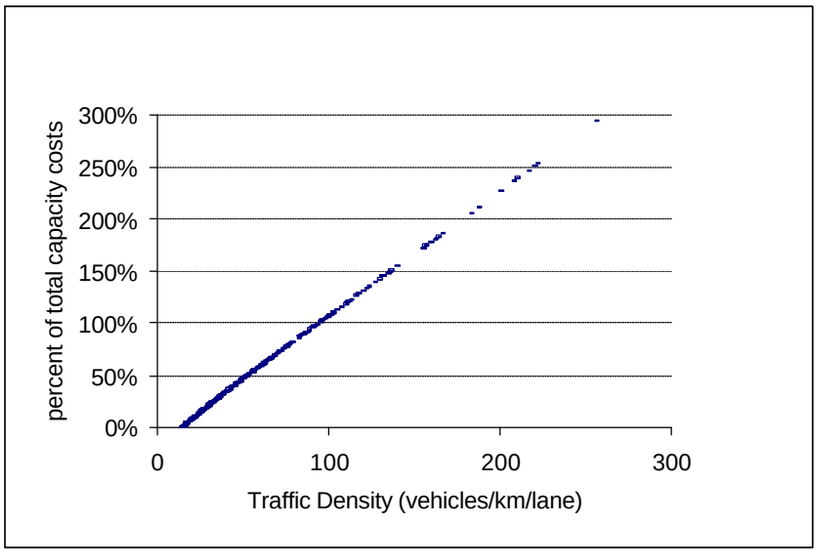


Figure 12. Implicit Congestion Tax for case (ii): Short Run Excess Demand

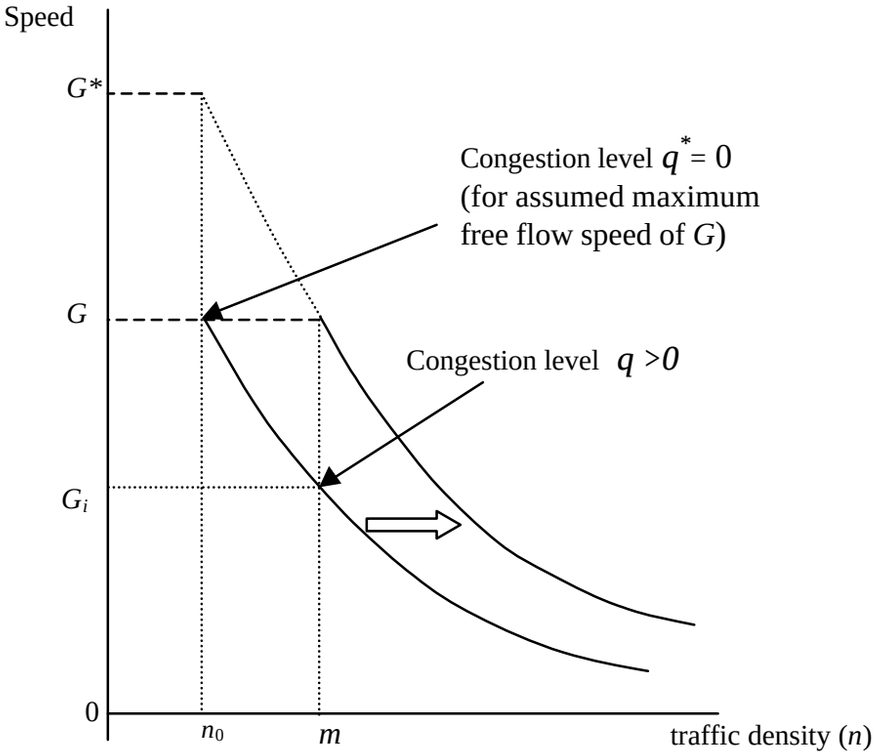
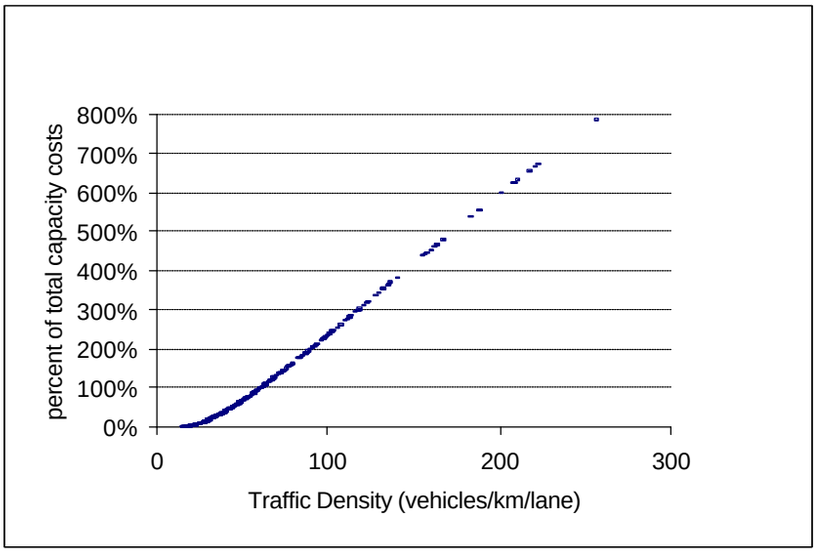


Figure 13: Implicit Congestion Tax at different traffic density levels (m), assuming that long run congestion level is zero (long run traffic density level is n_0) and maximum free flow speed is 80 km/h



Conclusions

In this paper, we have set out to explore the role of public infrastructure investment in private sector productivity by constructing a model of private sector activity using public infrastructure as a kind of congested public good input. The paper establishes the conditions under which optimal provision of public infrastructure can be said to have been reached. This is when each individual user pays an ‘effective’ (but unobserved) price or capacity charge for the use of the public infrastructure which is just sufficient to cover its marginal productivity, and the aggregate of all these effective charges is also equal to the supply cost of the infrastructure. When this condition is satisfied, the ‘congestion’ level can be assumed to be at a long run equilibrium level (which can then be ‘calibrated’ to a ‘reference’ level such as zero). We then proceed to establish a relationship between the actual levels of public infrastructure demand (as indicated by the level of traffic density), supply (as indicated by the level of free flow speed), and the level of congestion. Congestion therefore acts as a kind of invisible hand or price signal to guide demand and supply for the infrastructure towards the equilibrium level. We illustrated how these signals can be used to guide investment in infrastructure capacity towards a long run equilibrium level. In the context of private sector participation in the optimal provision of toll roads, our model can provide useful guidelines towards the determination of an ‘optimal’ set of shadow toll prices for these privately funded public infrastructure goods.

References

Alesina, A., D.W. Gruen and M. T. Jones (1991) Fiscal Adjustment, the Real Exchange Rate and Australia's External Imbalance, *The Australian Economic Review*, 3rd quarter.

Aschauer, D. A. (1988) The Equilibrium Approach to Fiscal Policy, *Journal of Money, Credit, and Banking* 20, 41-62.

Aschauer, D. A. (1989a) Is Public Expenditure Productive?, *Journal of Monetary Economics* 23, 177-200.

Aschauer, D. A. (1989b) Does Public Capital Crowd Out Private Capital?, *Journal of Monetary Economics* 24, 171-188.

Berndt, E. R. and B. Hansson (1991) Measuring the Contribution of Public Infrastructure Capital in Sweden, *National Bureau of Economic Research (N.B.E.R.) Working Paper 3842*, September.

Dixon, P. B. and D. McDonald (1991) Labour productivity in Australia 1970-71 to 1989-90, in: Economic Planning Advisory Council (EPAC), *Background Paper, No. 8,* AGPS, Canberra, January.

Nadiri, M. I. and T. P. Mamuneas (1991) The Effects of Public Infrastructure and R&D Capital on the Cost Structure and Performance of U.S. Manufacturing Industries, *N.B.E.R. Working Paper 3887*, September.

Lindsey, R. and E. Verhoef (2000) Congestion Modelling, in Hensher, D.A. and Button, K. (eds) Transport Modelling, Handbooks in Transport Vol 1, Pergamon Press, Oxford, Ch 21, 353-374.

Oakland, William H. (1987) Theory of Public Goods, Chapter 1 in: A. J. Auerbach and M. Feldstein, eds., *Handbook of Public Economics*, Volume 2 (North Holland, Amsterdam).

Ozbay, K., Bartin, B., and J. Berechman (2001), “Estimation and Evaluation of Full Marginal Costs of Highway Transportation in New Jersey”, *Journal of Transportation and Statistics,* 81-103.

Shah, A. (1992) Dynamics of Public Infrastructure, Industrial Productivity and Profitability, *Review of Economics and Statistics* 74, 28-36.

Truong, T.P. and D.A. Hensher (2002) Congestion As The Invisible Hand In The Optimal Provision Of Public Infrastructure Goods, Institute of Transport Studies, The University of Sydney, May.