Transportmetrica, Vol. 1, No. 2 (2005), 81-117

ECONOMIC FUNDAMENTALS OF ROAD PRICING: A DIAGRAMMATIC ANALYSIS, PART I – FUNDAMENTALS[1]

TIMOTHY D. HAU[2]

Received 2 April 2004; received in revised form 28 February 2005; accepted 3 March 2005

Parts I and II together presents a conceptual framework for road pricing based on a rigorous diagrammatic -- but non-mathematical -- framework derived from first (economic) principles. It throws light on congestion pricing systems and issues surrounding short-run and long-run marginal cost pricing, scale economies and diseconomies, indivisibilities, road durability, the peak-load problem in urban transport and the financial viability of the public provision of road services. The paper integrates the ideas of Mohring, Strotz, Vickrey, Walters, Keeler, Small, Winston and Newbery into a single analytical framework.

Analysis of the traditional road pricing arguments demonstrates why congestion pricing as practiced in the past has understandably encountered obstacles to implementation. This is partly because both types of road users, the tolled and the tolled off (those who avoid a road in order to shun the toll), are shown to be worse off -- with the exception of the government -- under a constant value of time. Even if differences in time valuation are taken into account, it is still essentially the case that primarily those with very high time values are made better off. Unless congestion toll revenues are earmarked and travellers perceive that the money is channeled back in the form of reduced taxes, lower user charges or improved transport services, neither the priced nor the priced off would support road pricing. It is only in the case of hypercongestion can congestion pricing be shown to make everyone better off.

In the absence of scale economies or diseconomies, the level of economic profits, i.e., toll revenue collections less the fixed and non-use related costs of a road, serves as a surrogate market mechanism indicating that a road ought to be expanded or downsized. The decision to let roads deteriorate over time in and of itself is an act of disinvestment. Further, it is shown graphically that if a road authority were to efficiently charge for congestion, it is possible to make money on the road. Profitable roads arise in urban areas in the long run because land rents are high and congestion tolls reflect the high opportunity costs. Yet, efficient pricing in the presence of both indivisibilities and diseconomies of scale in urban roads may curtail the extent of profitable undertakings, whereas pursuing marginal cost pricing under the restrictive conditions of both indivisibilities and scale economies in rural roads could result in profits in the short run. Economic efficiency would be enhanced if optimal pricing were pursued in the short run and optimal investment in capacity were pursued in the long run. The rule is therefore to implement short-run marginal cost pricing while varying road capacity over the long run.

Extensions by Newbery, Small and Winston have enriched the basic Mohring model that this paper develops diagrammatically by incorporating the fact that heavy vehicles are responsible for extensive road damages. Charging for both the external and variable cost of road damages on a vehicle weight per axle basis can help cover the deficit that may arise from congestion tolling. Even if a road network is broadly characterized by increasing returns to scale in road construction's use and strengthening, the deficit could be closed by diseconomies of scope. The existence of scope diseconomies in highways means that a road network that accommodates both loading and traffic volume, as found universally, costs more than the sum of an autos-only and a (smaller) trucks-only road system. Hence, the surplus associated with diseconomies of scope offsets the potential loss associated with scale-specific economies. The viability of a dedicated road or transport fund is enhanced by the fact that the road pavement is charged in two dimensions: once when traffic flow creates congestion and another when traffic loadings cause road damages.

KEYWORDS: Transportation economics, road pricing, congestion charging, congestion tolling, highway investment, diagrammatic analysis

---

[1] This paper is based on Timothy D. Hau (1992a) Economic Fundamentals of Road Pricing: A Diagrammatic Approach, World Bank Policy Research Working Paper Series, WPS 1070, World Bank, Washington D.C., pp. 1-96. The first part of this work appeared as Timothy D. Hau (1998) Congestion pricing and road investment, In Kenneth J. Button and Erik T. Verhoef (eds.) Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility, Chapter 3, Edward Elgar, Cheltenham, U.K. and Northampton, M.A., U.S.A., pp. 39-78.

[2] School of Economics and Finance, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, P.R. China. Corresponding author (E-mail: timhau@hku.hk).

1. INTRODUCTION

In major conurbations of both developed and developing countries, congestion is incessant during the peak periods and oftentimes the interpeak.[3] Yet the traditional methods of effectively curtailing congestion are few, and their usefulness limited. On the supply side, the expansion and improvement of roads is restricted by increasingly tight fiscal and physical constraints. On the demand side, however, the problem can be addressed by pricing or regulation. This paper uses the pricing (or market-based) approach to grapple with congestion because of its inherent flexibility and power of discrimination.[4] The aim of road pricing is to internalize the externalities generated by road use. My focus here is on removing the external effects caused by motorists by charging for congestion and road damage externalities. Congestion is recognized as an important type of externality from vehicle usage in both developed and developing countries in that it represents a large share of total estimated road use costs (Newbery, 1988a, b, 1990; Newbery et al., 1988).[5]

This research begins with a series papers on road pricing in theory and practice. The first two papers present a conceptual framework for road pricing. It gives an interpretative abridgment of the literature on the theory of optimal pricing, investment and durability of roads based on the works of Herbert Mohring, Robert Strotz, William Vickrey, Alan Walters, Theodore Keeler, Kenneth Small, Clifford Winston and David Newbery.[6] It aims to integrate their ideas and principles into a single analytical framework. I am convinced that the presentation of a rigorous and unified diagrammatic -- but non-mathematical -- framework derived from first economic principles casts important light on congestion pricing systems and on issues surrounding short-run and long-run marginal cost pricing, scale economies and diseconomies, indivisibilities, the peak-load problem in urban transport, optimal road durability, financial viability and cost recovery in the public provision of road services. While there are several automatic road user charging and electronic toll collection systems in use in parts of Norway, Italy, France, and the United States, as well as bills in parliament to implement various forms of road pricing in Santiago and Stockholm, one congestion pricing system is currently operating (Hau, 2006). That is Singapore's Area Licensing Scheme, which is now in the process of being converted into the Electronic Road Pricing System, to be operational in

---

[3] By the turn of the century, more than four-fifths of the world's most heavily-populated agglomerations are projected to be in the developing world (World Bank (1991), pp. 3 and pp. 22).

[4] The regulatory approach on quantity -- the so-called command and control measures -- suffers from its inability to provide correct market signals to induce the most efficient trips to be undertaken. In contrast to pricing incentives, it generates virtually no revenues for the public sector (see Pozdena et al. (1990)).

[5] The total cost of undertaking a trip -- both 'private' and 'social' -- involves: 1) congestion (which is borne by road users); 2) pavement wear (which is covered by the road authority); 3) air and noise pollution; and 4) costs of accidents (both of which are borne by society at large); in addition to 5) vehicle operating costs. 'Social costs' (i.e., both internal and external costs) here are distinguished from 'private costs' (i.e., internal costs), the latter of which are self-financed. Those traveller-borne internal costs include: 1) vehicle operating costs as well as 2) time costs and delay in congested traffic (Newbery's definition of road use costs refers only to the external costs.) This paper deals mainly with the marginal cost pricing of congestion and pavement wear charges as opposed to marginal *social* cost pricing, which is defined to include -- in addition to private costs -- all the *external* costs of congestion, air pollution, noise pollution, accidents, road damages and externalities. Hence, strictly speaking, 'congestion pricing' refers solely to the pricing of congestion externality whereas 'road pricing' refers more generally to the pricing of all externalities from mobile sources.

[6] I owe much intellectual debt to these authors and I am very grateful for the opportunity of discussing my work with several of them.

1995 (see Watson and Holland (1976, 1978), Behbehani et al. (1984)). Even so, the charging of vehicles by daylight hours in Trondheim, Norway (with further differential pricing for electronic tag users) since 1991 could be regarded as a rudimentary form of congestion pricing.

Technological breakthroughs in automatic road user charging have brought electronic road pricing much closer to reality. Thus, Hau (2006) in this research presents a taxonomy of alternative technologies of congestion pricing. It drops the crucial assumption that the implementation of short-run marginal cost pricing is costless (Hau, 1990). In particular, it compares and contrasts the methods of manual versus electronic charging schemes. Electronic approaches are analyzed progressively by discussing first, the increasingly popular electronic toll collection mechanisms which incorporate technological advances in automatic vehicle identification -- commonly known as AVI -- as well as smart cards. The appropriateness and applicability of electronic toll collection mechanisms to electronic road pricing are then dealt with. The companion paper also analyzes the relative cost-effectiveness of each technology and performs benefit-cost analysis where data permits. The implications of using each of these technologies for relieving congestion are discussed and policy recommendations are drawn.

2. CONCEPTUAL GUIDELINES

Rising real incomes result in increased aspirations for the ownership of private automobiles. Barring major restraint measures, an increasing number of motor vehicles means that travel demand swells concomitantly. Because municipalities are finding it increasingly difficult to finance new road construction and improvements, the rate of growth of travel demand outstrips the growth of road capacity (Hau, 1988). Transport planners, with limited options available, find it very difficult to combat effectively mounting traffic problems in the face of increasing urbanization. The resulting traffic explosion is an illustration of Parkinson's Law or Downs' law of peak-hour expressway congestion, in which commuter traffic ascends rapidly to the level of new capacity in urban areas (Downs, 1962; Mohring, 1965). Traffic engineers have long been familiar with this "fundamental law of highway congestion" in which latent demand expands to fill the gap created whenever highway capacity is improved (Small et al., 1989, pp. 85). In this section, I set up the conceptual guidelines that allow authorities to curtail traffic congestion in an efficacious manner at the same time as satisfying the World Bank's general guidelines for public sector projects and urban transport policy (World Bank, 1977, 1986; Julius and Alicbusan, 1989).

In a nutshell, the essential principles include:

1) implementing short-run marginal cost pricing (short-run efficiency) to generate maximum net benefits for society: *efficiency pricing*;
2) undertaking investment in infrastructure whenever the additional benefits exceed the costs (long-run efficiency) to society of doing so: *economic viability*;
3) investing in transport services when benefits exceed costs; promoting public transport services especially when revenues exceed costs: *financial viability*;
4) maintaining 'fairness' among beneficiaries, for example, via benefit taxation -- *equity* -- where possible; and
5) using pricing and cost recovery policies to improve the efficiency of managing the public sector -- *cost-effectiveness* and *managerial efficiency* -- if possible.

3. CAUSE OF CONGESTION: PUBLIC/PRIVATE OWNERSHIP OF ROADS

Roads which are infrequently utilized possess the characteristic of nonrival consumption among users and are traditional examples of public goods. Joint consumption means that roads yield services that are simultaneously enjoyed by more than one user, without substantial detriment to the satisfaction of others. If roads are totally nonrivalrous, then neoclassical economic principles dictate that roads ought to be provided for by the public sector and financed from general revenue taxation (and perhaps land value taxation), fully taking into account the social opportunity cost of public funds. On the other hand, roads which are heavily utilized have the nature of rival consumption among users and are called congested public goods. These variable-use congested public facilities then approximate to varying degrees the rival nature of private goods. Private goods are of course provided contingent upon payment, excluding those who are not willing to pay for them. Alas, with free access to roads, people are not barred from the use of scarce services, resulting in overuse. Hence market failure due to nonexcludability calls for governmental intervention in the form of better designed road user charges and motivating charges which correct for externalities. (The failure of the voluntary pricing mechanism due to nonexcludability is referred to as the 'free-rider problem' in the public finance literature (Boadway and Wildasin, 1984, Chapter 3)). Note that it is the traditional inability to exclude motorists from the use of crowded streets that is *the* cause of market failure. However, when refinements in automatic vehicle identification technology become even more cost-effective, automatic road user charging means that market failure is somewhat overcome. This is because prices are then able to reflect both the intensity of demand and the true economic cost of road use. The problems of nonexcludability and preference revelation would then be virtually eliminated and the advantages of price incentives reaped. Thus the standard public finance text (see, for example, Musgrave and Musgrave (1989, Chapter 4)) argues that both the nonrival consumption characteristic of public goods and the nonexcludability -- or rather, costly excludability -- of congested public goods are causes of market failure, calling for government intervention. Hence, roads which possess the attribute of congested public goods, and thus have a partially rival consumption characteristic, ought to be treated by the relevant governmental authorities as mixed or impure public goods, if not private or club goods (Buchanan, 1965).

The foregoing discussion leads naturally to the transport economist's definition of congestion. That is, as more and more vehicles join a traffic stream, the travel times of all motorists making trips are raised, resulting in delay to all. In essence, the congestion phenomenon is one of excess demand, given a fixed road network.

Yet the fundamental reason why congestion occurs so ubiquitously is because property rights are not clearly delineated, yielding market failure. If roads are both privately owned and competition prevails in the provision of roads, usage would be (Pareto) optimal in the absence of scale economies (Knight's conjecture (1924)).[7] If roads are not

[7] We know that competition results in economic efficiency. In the context of roads, owners maximize the return to their land by setting the price equal to the short-run marginal cost of production, where the difference between the short-run marginal cost and the average variable cost represents the quasi-rent to the fixed factor of production. (The notion of quasi-rent is discussed in footnote 31.) This is the analog to society mimicking the decentralized competitive landowner's rent maximizing behavior by charging the optimal (Pigouvian) toll -- the divergence between marginal cost and average variable cost (see discussion in Section 9 and the proof of this conjecture in Vickrey (1968)). Knight further argues that the case of increasing returns is dealt with by the market allowing only

privately owned, access to and usage of roads are effectively free to the traveler, resulting in excessive use of those roads where traffic is heavy enough to produce significant adverse interaction among vehicles. (The overcrowding occurs *despite* the fact that motorists already have to pay fuel taxes and registration/license fees and purchase taxes for owning a vehicle, which can be regarded as entry fees only.)[8] On the other hand, it is unlikely that private ownership of roads would yield a perfectly competitive market structure because alternative routes are in general not perfect substitutes. Hence, those road owners might exploit their locational attributes and monopolistic prerogatives by raising charges above marginal cost. In addition, incentives would arise for collusion. Thus competition would at best be imperfect in the case of road provision where lumpiness is found. The relevant comparison, based on efficiency consideration, is the welfare cost associated with government intervention vis-a-vis the dead-weight loss associated with the monopolistic or oligopolistic pricing practices of private ownership. In addition, it seems reasonable to argue that there are other grounds -- such as equity -- for regarding landowners' super-normal profits, or collusive tendencies, as undesirable.

In practice, virtually all roads belong to the government and not to private owners. Because of free-access roads, one should not be surprised to be confronted with the pervasiveness of congestion. This is the common property resource problem, which yields an externality.[9] One might ask whether or not some congestion-prone roads ought to be privatized to internalize the externality. The key factor is whether (sufficient) competition could be made to prevail to ensure, in the absence of regulatory restraints, that these roads are not overpriced. A testable hypothesis in contestable markets theory is to check whether the condition of potential competition and not actual competition is satisfied (Baumol et al., 1982, Chapter 1). One surmises that it may be difficult to assume that it would be effective in practice, especially in the case of roads where irreversibility and lumpiness exists (Baumol and Lee, 1991). Also, noncomparability may occur because a competitor would find it difficult to duplicate a road on the same alignment and would likely settle for inferior alignment. Economic efficiency could be

---

one agent to remain behind to exhaust scale economies (see Walters' (1954) interpretation of Knight's conjecture (1924), which does not accord with my interpretation here). I argue that the government is left with the role of being the agent that charges the congestion toll, in order to prevent monopoly pricing. After all, it is unclear *a priori* that the informational requirements and transaction costs of public ownership would exceed the social costs (such as rent-seeking) of regulating private ownership. In other words, it is debatable whether government failure is necessarily greater than market failure, especially in developing countries. For instance, the government's power of eminent domain over public projects may reduce its land acquisition and road construction costs whereas the common property nature of roads would encourage rent dissipation.

[8] The fuel taxes which motorists pay are generally used to contribute to: 1) the maintenance and operating (and in part the investment) costs of roads, 2) debt service on highway bonds, and/or 3) general tax revenues. (Independent of how fuel tax revenues are disposed of, the main problem of the uniform fuel tax is its inability to vary by time and place of use (see Hau (2006)).) The combination of a fuel tax, first registration taxes and annual license fees can be regarded as, to a rough order of approximation, a two-part or multi-part tariff.

[9] As long as property rights are well-defined, exchangeable and enforceable, private bargaining would internalize all 'externalities' and yield an efficient allocation of resources independent of who is held liable for creating external effects in the first place (Coase, 1960, 1988). Since Coase's so-called theorem holds only if the costs of transactions are nil and income effects are negligible, it applies mainly to the case of small numbers. With congestion and mobile-source externalities, I argue that the large numbers of people affected clearly result in large transactions costs, and hence market failure, beckoning government intervention. Further, Cooter (1982) argues that Coase does not distinguish between zero transactions costs and zero bargaining costs, and that the generality of the "Coase Theorem" is overstated.

enhanced in some cases if there were a *mixture* of both private and public roads, since the existence of publicly-provided *free*ways might limit the degree to which those who use private expressways would be charged monopoly prices. Similarly, private toll roads could serve to over-zealous governments from overcharging on public toll roads.[10]

In short, the congested public good nature of roads suggests that these facilities should be treated more like club goods or private goods, yet the status quo appears to suggest that public ownership of roads is institutionally preferable. If so, municipalities ought to simulate the workings of a competitive, private firm and industry by setting congestion tolls on public roads to internalize the congestion externality. This thereby deals with the problems of nonexcludability and preference revelation explored earlier. If a private ownership arrangement is deemed beneficial at times, the government could exercise its power of eminent domain in reverse by altering the structure of property rights of the relevant facilities from public to private ones. This could be done, for example, by auctioning off government land to the highest bidder (and simultaneously reducing general taxes and/or providing more public goods), or, alternatively, by designating the competitive provision of highway services to corporations or private contractors via "franchise bidding" (Mills, 1981). I have argued, on both efficiency and equity grounds -- following our conceptual guidelines -- that the facility's ownership arrangement ought to reside with the government, with the proviso of setting cost-effective congestion tolls.[11] I next show how a congestion toll can be derived from a transportation engineering speed-flow curve. In the process, the equivalence of efficient congestion tolling and the short-run marginal cost pricing of vehicle flow is established.

4. FOUNDATIONS OF ROAD CONGESTION -- THE CLASSICAL CASE (IN THE SHORT RUN)

Road congestion is well founded in the economics and transportation engineering literature (Dupuit, 1844, 1849; Pigou, 1920; Walters, 1961a,b).[12] It begins by considering a representative driver cruising under low traffic conditions along a given stretch of urban road with fixed beginning and end points. The representative driver would be able to achieve a mean free speed that balances the benefits to him of a faster trip against the costs to him of higher energy requirements and a greater risk of an accident. *Ceteris paribus*, as other vehicles enter the road thereafter, density increases, speed drops and travel time (or delay) lengthens (and accident probability rises). The causality is as follows: traffic density determines speed and not vice versa. Parallelling the theory of fluid dynamics, traffic flow is the product of density, in vehicles per

[10] Note that a relatively uncongested toll road would yield inefficient traffic allocation if its effect is to price traffic off to alternative public roads which are already congested. See Johansen (1989) and Catling and Roth (1987) for a discussion of toll roads and road privatization.

[11] While I am convinced of the advantages of market forces, I have some reservations about their extent. For example, some have argued that many social services should be provided for by the market, and also that transport infrastructure should also be privatized (Roth, 1987, Chapter 6; Catling and Roth, 1987). It is difficult to see, in the absence of road divisibility and competitive forces, how marginal cost pricing would be pursued and maintained by private road owners. From a public choice perspective, the same argument could be reversed and applied with respect to governmental authorities. This explains why I argue that perhaps a mixed system of public/private ownership/regulation would be able to reap the strength of both.

[12] While Walters (1968, pp. 31-34, Chapter 3 Annex) discusses optimal investment, his works emphasize the short run character of road pricing. The approach taken in this paper is to highlight both the short-and-long-run nature of the congestion problem in order to explore the issue of cost recovery.

kilometer, and speed, in kilometers per hour. Note that the rectangular area in Figure 1a -- the speed-concentration relationship -- is equivalent to traffic flow, expressed in vehicles per hour (see Gerlough and Huber (1975, Figure 4.12) and May (1990, Chapters 7, 10), for example). Hence, traffic flow is the product of traffic density and speed, with traffic flow attaining a maximum at F^{max} with speed at S^{m} in Figure 1b -- the speed-flow relationship (see Haight's (1963) "fundamental diagram of road traffic" -- a flow-density curve -- and similar figures in Morrison (1986)). (Touted maximum flow figures of the 'capacity' for a typical expressway are about 1800-2000 vehicles per lane-hour at 50-55 kilometers per hour (30-35 mph) (see Gerlough and Huber (1975, Chapter 4) and any Highway Capacity Manual, for example, Transportation Research Board (1985, Chapters 2-3)).[13]

Given a fixed distance of say a kilometer of road, the traffic engineer's speed-flow curve can be straightforwardly converted to a travel time-flow curve as travel time is the reciprocal of speed, with vehicles-kilometer per lane-kilometer-hour on the horizontal axis (see Figure 1c). Using a constant value of time as a shadow price for the representative driver, travel time is then converted to a money basis which yields time cost, called the average variable cost, AVC (see Figure 2). Low traffic volume corresponds with relatively high speed, so fuel cost would be high. With high traffic flow and low speed, however, fuel cost would also be kept high because of fuel inefficiencies caused by the alternate acceleration and deceleration associated with dense traffic. These two factors roughly cancel one another out, leading to the plausible assumption that the costs of operating an automobile (which include fuel, oil, maintenance and depreciation costs) are approximately independent of the level of traffic flow (Mohring, 1976, Chapter 3; AASHO's (1960) 'Red Book'). A fixed money cost for the vehicle operating cost can therefore be added to the time cost portion to form the generalized cost - an accepted construct of transport economists (Nash, 1976; Button, 1982). Similarly, the road's variable maintenance cost, which is assumed to be proportional to the traffic level, following Walters (1968, pp. 24), can be added up also (see Figure 2).[14] So it is the time cost element that is mainly responsible for the upward-sloping portion of the AVC curve. The AVC curve climbs upwards because significant negative interactions occur before traffic reaches maximum basic capacity, Q^{max}; it is variable in the sense that as traffic flow, Q, is increased, congestion delay actually sets in rapidly at substantially below the traffic level Q^{max} (contrary to the engineering notion of a constant average variable cost curve extending up to the point $Q^{max} \equiv F^{max}$). After the engineering or basic capacity, $Q^{max,}$ is reached, AVC becomes an 'inverse supply' curve. Note that the standard supply curve is nonexistent in the context of roads.[15]

---

[13] Economists' notion of 'capacity' is whenever congestion delay begins, which is considerably less than traffic engineers' concept. Further, vehicular flow is sometimes normalized by the capacity of a certain road to yield a 'volume-capacity' ratio.

[14] Hence the marginal cost curve, MC, when summed up vertically with the vehicle operating cost and variable road maintenance cost, yields a translated marginal cost curve, *MC*. The same notations apply to the average variable cost curve.

[15] The solid portion of the speed-flow curve and travel time-flow curves in Figures 1 and 2 denote the 'normal' part of the curves. The non-solid backward-bending portion of the cost curve (Figure 2) means that time cost increases because traffic flow is reduced after the engineering capacity is reached. The backward-bending curve, far from being fictitious, has been substantiated in the literature (Gerlough and Huber (1975, Chapter 4), Keeler et al. (1975, Figure 1)). Schiff (1991) explores this backward-bending case.

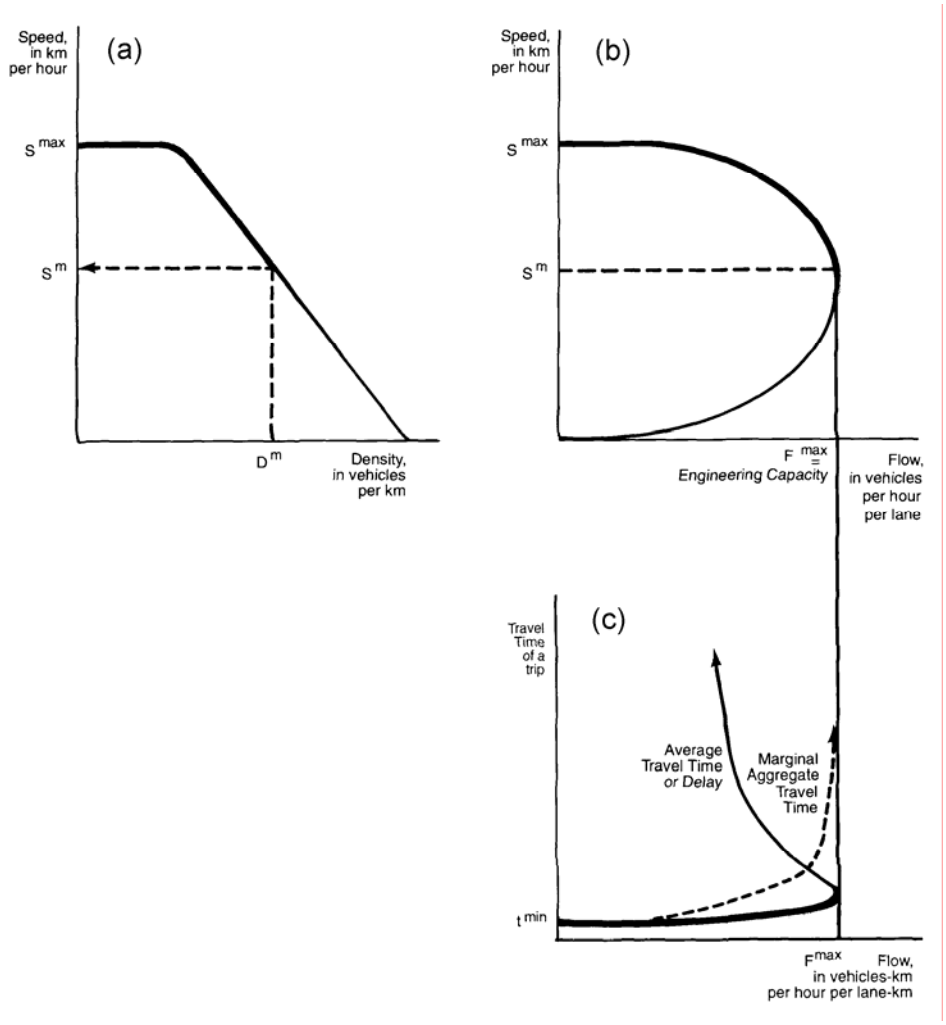

FIGURE 1: Derivation of a travel time – flow curve of an urban highway

5. DEMAND AND SUPPLY

The 'supply' side can be made to be congruent with the demand side when a conventional demand curve is specified to depend on the travel cost (price) facing a traveler for a single trip. When an initial demand function, Q^d, intersects the AVC curve at point U (Figure 2), a (stable) equilibrium is said to exist at Q^o.[16] This is an equilibrium point because travelers' willingness-to-pay curve, i.e., the inverse demand function,

[16] The demand function, Q^d, is an observed, constant-money-income Marshallian demand curve, with the usual regularity conditions. It approximates the exact Hicksian demand curve, which yields the true reflection of willingness-to-pay and marginal benefit. Formally, the functions GC(Q) and $Q^d(P)$ intersect at an equilibrium point, where GC symbolizes the generalized cost. The equilibrium point is expressed as: $Q^d(GC(Q^o)) = Q^o$. Since GC is simply a translation of AVC, the interpretation of one is synonymous with the other.

equals the average variable cost curve -- the function upon which travelers base their travel decisions. After a small excursion of the demand in the neighborhood of the equilibrium point, unfettered use results in a return to the observed traffic level, Q^o, hence the equilibrium is considered stable.

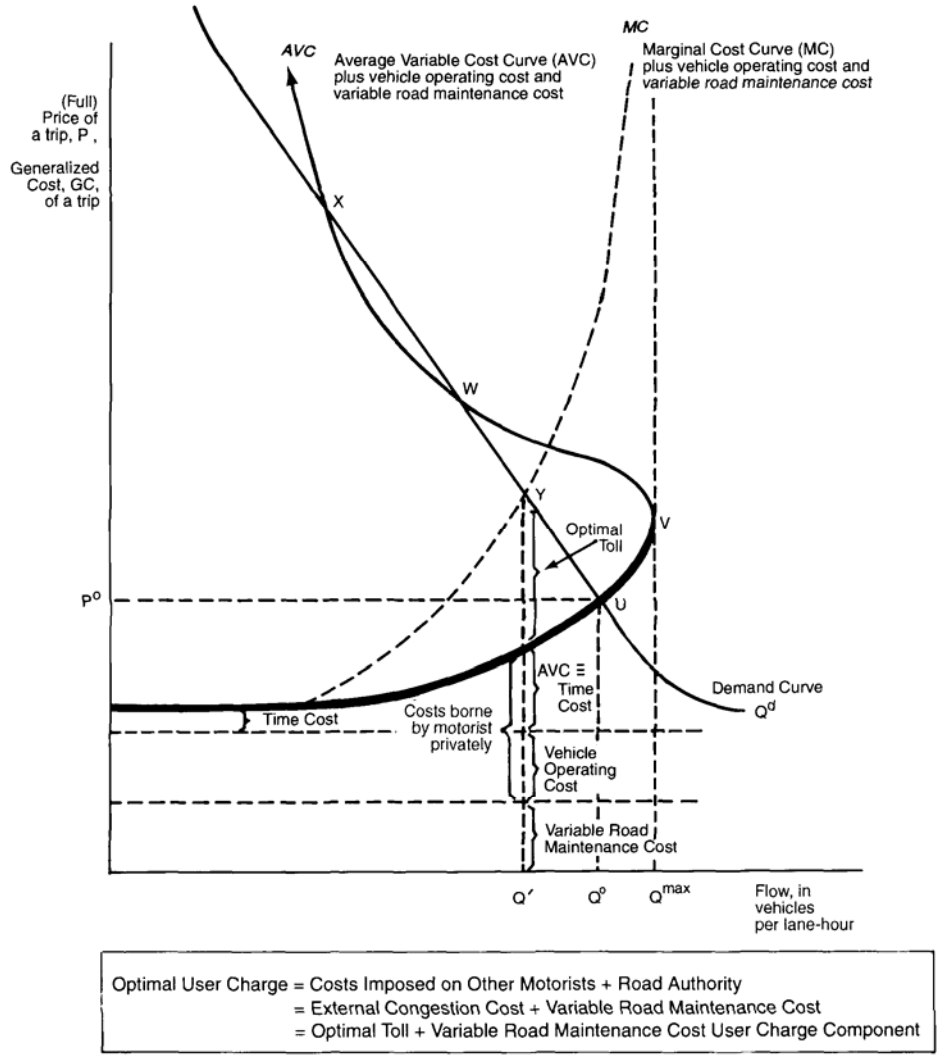

FIGURE 2: Derivation of the marginal cost curve and congestion toll

Basic price theory says that whenever the average variable cost rises, it means the marginal cost curve lies above it.[17] The vertical difference between the two cost curves is

[17] Marginal cost is obtained as follows: $MC \equiv \Delta C(Q)/\Delta Q = AVC(Q) + Q \cdot \Delta AVC(Q)/\Delta Q = AVC(Q) \cdot (1 + \varepsilon)$ where $C(Q)$ is the cost function, ε is the elasticity of the AVC curve, i.e., the rate at which time cost rises with respect to a one percent rise in traffic flow (Walters, 1961a). The first term composes of only time cost and the

the marginal (external) congestion cost -- the additional delay that one driver imposes on the rest -- which is not taken into account by the last driver who joins the traffic stream. In fact, since each driver chooses whether or not to travel according to the AVC curve -- being the decision curve -- he or she totally ignores the resulting external congestion cost imposed on fellow motorists. We thus have the optimal point Y at which the marginal cost curve intersects the (peak) demand curve in Figure 2. In other words, Q′ is the associated optimal output in the sense that the generalized cost which includes external congestion cost and other variable costs (i.e., constant (unit) operating cost of a vehicle and variable maintenance cost of a road), is equated to the price. The congestion cost is the additional time cost that a motorist imposes on others, calculated by taking the increment in average time cost caused by the added trips and multiplied by the number of vehicles in the traffic stream. The Pigouvian tax applied to roads is that optimal toll which closes the wedge between the marginal cost and average variable cost curves by emitting the correct signal and creating appropriate (dis)incentives.

This Pigouvian toll-tax is equal to the marginal external congestion cost. It is also known as the net-benefit maximizing, economically efficient, (Pareto) optimal and marginal cost toll. (If one prefers, it can be regarded in graphic terms as a sin tax, even though it has not yet included the cost of environmental and other externalities.) Hence the marginal cost pricing of trips in the short run (given that a road is fixed) yields a first-best Pareto optimal allocation of resources. The optimal road user charge is then comprised of a congestion toll and another component which covers the variable road maintenance cost (see Figure 2's legend).

6. VERY CONGESTED ROADS AND BOTTLENECKS

Observe that at the equilibrium point U in Figure 2, the resultant throughput is significantly less than the road's maximal flow capacity of point V. The backward-bending 'supply curve' exists because a one percent increase in density results in more than a one percent decrease in speed when very dense traffic is reached. (That point X is a stable equilibrium point can be seen intuitively by perturbing the price level while point w is unstable in that any disturbance will result in a movement to U or X.) The point here is that quite a few cities, for example, Bangkok, Budapest, Buenos Aires, Hong Kong, Jakarta, Mexico City, Santiago, Sao Paulo, Seoul, Singapore and Taipei are faced with extremely congested situations such as point X, certainly during peak of the peak.

---

second term is the marginal (external) congestion cost, set equal to the congestion toll. Marginal cost pricing of a trip, P, is achieved by setting P = MC. This is known as the first-best optimal pricing rule: our first optimality rule. (Note that AVC depends parametrically on the capacity level K, and can be expressed as AVC (Q, K). Without loss of generality, the inclusion of the average vehicle operating cost and variable road maintenance cost -- both being constant with respect to traffic -- simply alters both the left-hand side and the right-hand side by the same amount). Implementing marginal cost pricing in this case means pricing the *difference* between marginal cost and the average variable cost of a trip, plus a component which covers the cost which the motorist imposes on the community (see Figure 2's legend). We shall return to an intuitive discussion of this subtle but important distinction. Note further that marginal cost rises asymptotically to the engineering capacity level of Q^{max} and is undefined for the AVC curve at points beyond V.

Consider the dynamic phenomenon of traffic growth as shown in Figure 3. Suppose the

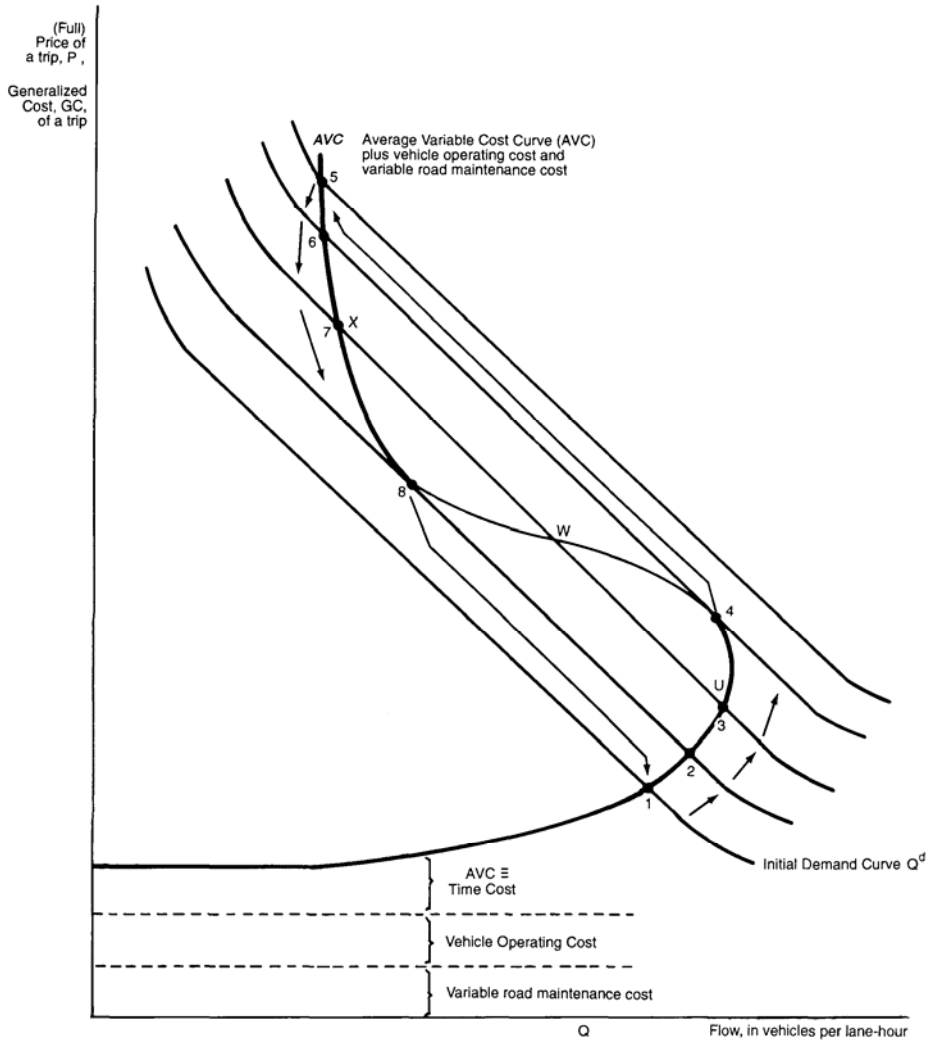

FIGURE 3: 'Dynamic' phenomenon of traffic growth: the relaxation effect

initial demand curve intersects the average variable cost curve at point 1. As traffic grows, the observed number of vehicles per lane-hour increases from point 1 to 2 to 3 (which is identical to point U of Figure 2). A further increase in demand beyond point 4 would result in a discontinuous jump to the backward-bending part of the unit variable cost curve at point 5. Intuitively, traffic congestion worsens rapidly and 'jams up' all of a sudden at times. After a while at this point, travel demand would start to slack off to points 6 and 7 (which is identical to point X of Figure 2). Note that traffic would be moving at a snail's pace as queuing develops. This corresponds diagrammatically to the positive sloping part -- as opposed to the relatively smooth traffic of the 'normal' downward sloping portion -- of the speed-flow curve, Figure 1b. As travel demand continues to diminish towards the end of a rush hour, say, the traffic level would touch

point 2 briefly but would then end up at point 8. Thereafter, further slackening of demand results in traffic returning to the upward sloping portion of the unit cost curve at point 1, say. This is known as a relaxation phenomenon in an engineering and physics context. Therefore, the points between 4 and 8, such as the equilibrium point W, are never reached in reality![18] We conclude that it is precisely the daily recurrent peaking characteristic of travel demand that calls for innovative solutions. That the traffic level at those rush hour periods seems to be near gridlock -- an illustration of 'hypercongestion' -- is too important a case to be ignored (Walters, 1987).

7. THE WELFARE IMPACT OF ROAD PRICING

The purpose of this section is to highlight one of the main points of this paper, the proof of which is relegated to the appendix. Economists *know* that road pricing results in improvement in welfare to society, yet politicians and the public almost unanimously regard it with skepticism. Why? To economists, the increase in welfare comes about because of the imposition of an externality-corrective (toll-) tax. Yet, for those motorists remaining on the road, the congestion toll is similar to a tax increase. Under conditions of 'normal' traffic (or non-hypercongested situations), note that the toll-tax paid by the motorist exceeds the valuation of time savings resulting from road pricing on average, so that the 'tolled' is worse off.[19] Those who are priced off the road to an inferior mode or time of travel in order to avoid paying the toll are also worse off, while those who remain on other indirectly impacted roads are either worse off, if congestion arises there, or just as well off, if there is no resulting congestion. It turns out that the government, in collecting toll revenues, becomes the main party that is better off.

8. THE EFFECTS ON THE TOLLED, THE TOLLED OFF AND THE UNTOLLED

In the Appendix, I establish the proposition that marginal cost pricing of trips maximizes the net benefit to society in the sense that a Pareto-efficient situation is attained, that is, no one can be made better off without making someone else worse off. The quantity approach (the 'American' approach) yields the welfare gain due to road pricing of area l (shown by the triangle UYZ in Figure 4). The equivalent area, $e+f-d$, is the welfare gain using the change in total benefits and costs approach (the 'British' approach). The latter area is the cost saving of area $e+f+g+k$ less the loss of the use value of area $g+k+d$. We note that there are two groups of travelers that are clearly adversely affected by road pricing: the tolled and the tolled off. Briefly, the British approach to the calculation of welfare gain says that those who remain on the road after road pricing is introduced incur a cost by making a toll payment of the rectangular area $b+c+e+f$. On the other hand, the traveler benefits in the form of reduced travel time – 'forcibly' induced -- and the resultant time savings is the smaller rectangular area $e+f$. Hence the consumer-traveler would regard this exchange as not getting good value for money because the traveler is still faced with a net payment of area $b+c$ relative to the

[18] In other contexts such as Walters (1987), point W is described as an unstable equilibrium.

[19] The 'tolled, the tolled off, and the un-tolled' are terms coined by Zettel and Carll (1964). My approach differentiates from both Zettel and Carll and Wohl and Hendrickson's (1984, pp. 114-116) in that I apply the standard neoclassical method of welfare analysis in the appendix to derive my results and policy implications.

no toll situation: this area is his loss of consumer's surplus. Despite such a trade, the traveler would undertake the trip because his willingness-to-pay still exceeds the price.[20]

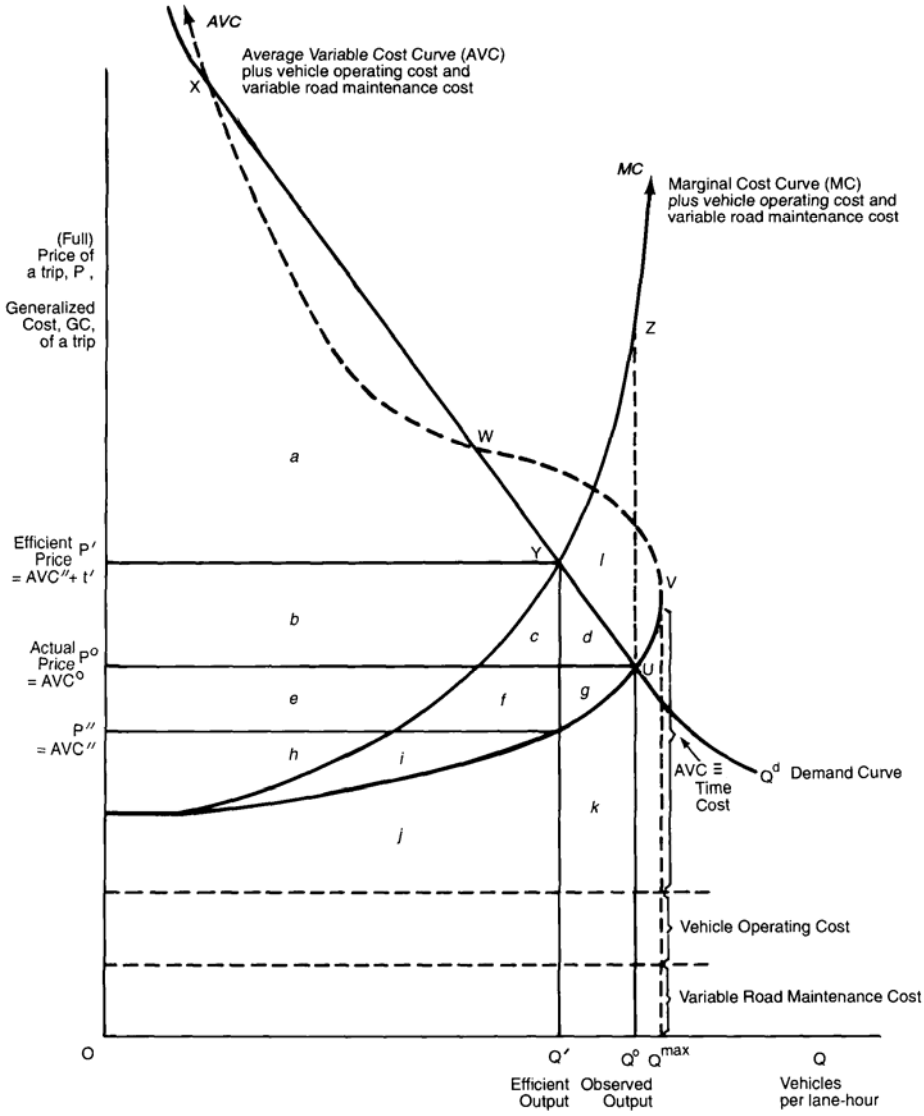

FIGURE 4: Welfare impact due to the introduction of road pricing in the peak period: short-run marginal cost pricing

The other identifiable group includes all those marginal users whose willingness to pay is not high enough and hence are tolled off the road. As a group, their loss in valuation, the vertical trapezoidal area $d+g+k$, exceeds their saving in time cost of the area $g+k$ by

[20] It is as if a blackmailer -- the government -- carves out part of the motorist's consumer's surplus.

the (welfare) loss in consumer's surplus of area *d*. So both groups are necessarily worse off vis-a-vis the original situation. This is shown to be true despite an argument that those who remain benefit from reduced time cost! They *do* benefit from reduced travel time, but they have to exchange money for time. In addition, the "tolled on" is either just as well off, or worse off, depending on whether or not congestion arises. So it appears that, a move from a Pareto inferior position to a Pareto optimal state leaves everyone worse off! How could this happen? It comes about because we have not yet accounted for one agent: the government. The toll revenue collected by the government is considered an actual gain to society in that it is counted once only. Indeed, it may have a greater value than the dollar amount itself if it replaces other (income) revenue sources with an excess burden.

Suppose the toll revenue is collected and then put aside. An 'efficient' amount of traffic and congestion would still exist on the road in question, yet society -- an aggregation of both gainers and losers -- would be definitely worse off. Unless the public in general, or road users in particular, can partake in the tax proceeds either in the provision of public goods and/or reduction in tax revenues, they will definitely be worse off from this imposition of an 'optimal' congestion toll. If funds are not channeled back to road users, the government, or the rest of society, gains, but only at the expense of those faced with road pricing.[21] However, because of the nature of the Pigouvian tax, which possesses an asymmetric price signal, generators of externalities are taxed but those affected by the externality are *not* supposed to be compensated. This is a requirement of optimality in the case of *both* public and private good-type externality-corrective taxes (Baumol and Oates, 1988, Chapter 4).[22]

Two choices to those who are tolled off the road are to use the road during the off-peak period or switch over to public transport during the peak. The analysis is similar in spirit for both classes, so I will illustrate it for the off-peak period, for simplicity, following the same notation used thus far. Substitution of travel demand from peak to off-peak has the effect of shifting the off-peak demand curve to the right, from Q^{d}_{op} to $Q^{d'}_{op}$ (Figure 5). It is easy, as is sometimes done, to regard the area *m* as an additional 'benefit'. This procedure is incorrect because the area bounded by the demand curves in the substituted off-peak period is in fact a pseudo-benefit and is already accounted for by the welfare gain to road pricing in the peak period, i.e., the triangular area *l*, in Figure 3 (Mishan, 1988, Chapter 8). Intuitively, since there are no changes in consumer's or producer's surpluses, there are *no* net changes in benefits or costs: the additional trips made during the off-peak period are entirely self-financing in the short run. If congestion were to set in during an off-peak period with relatively medium levels of congestion, say, during the inter-peak, efficiency suggests that another congestion toll level be set to internalize the congestion externality. In this way, traffic will settle to another equilibrium level with a smaller congestion toll in the inter-peak period. This is the idea behind peak-load or differential pricing. A dynamic process takes place among the peak and inter-peak

[21] The Smeed Report of 1964 and others effectively *assume* away the problem by stating that the congestion toll revenues will be returned to the population in a lump sum nondistortionary manner.

[22] Note that Baumol and Oates (1988, 2nd edition, Chapter 4) corrects the mistake made in the earlier edition (1975, 1st Edition, Chapter 3), which says that only the victims of private good-type externalities ought to be compensated for after the imposition of a Pigouvian tax. This result is contrary to accepted notions of justice. However, Baumol and Oates (1988, pp. 236-240) point out that if the tax revenues were funnelled back *indirectly*, then there would perhaps be only insignificant divergences from Pareto optimality. The idea is that wealth effects on consumption ought to be minimized.

periods until an equilibrium is settled upon. Proper cost-benefit analysis requires that the changes in net benefits be calculated only in the periods which encounter changes in travel time due to congestion or decongestion. With a two market model, the welfare effects of road pricing in the peak period are simply repeated in the inter-peak period whenever congestion is encountered.

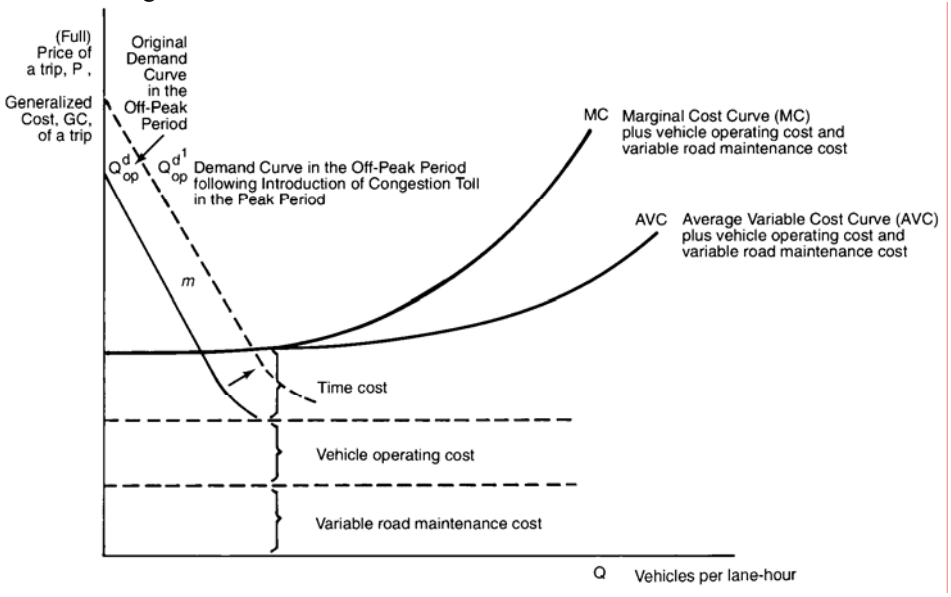

FIGURE 5: Effect of the introduction of road pricing in the peak period on the off-peak period

It turns out that there is a case often overlooked in which *everyone*, including the government, can be shown to be better off. This is the case of 'hypercongestion', where density is beyond the point of maximum flow. Here the traffic density is so high that both traffic flow and speed diminish, with the generalized cost P^o and traffic Q^o occurring at point A (Figure 6). (Even though it violates economic rationality to end up at such a point A, this type of traffic jam does in fact transpire fairly regularly, though limited to peak periods such as peak-of-the-peak.) The implementation of a marginal cost toll would result in travelers reverting to the normal non-hypercongested portion of the speed-flow curve (such as may be observed downstream of a bottleneck), which corresponds to the lower branch of the AVC curve at point C in Figure 6. In addition, the travelers have to pay a unit toll payment of distance BC, resulting in a generalized cost to the motorist of P', which is still *lower* than P^o. Because of the price decrease from P^o to P', traffic therefore increases from Q^o to Q' correspondingly. In this case, literally *everyone* is made better off: the tolled, what I call the "tolled on", and the government.[23] If the speed-flow curve is as depicted in Figure 1b, then hypercongestion appears more often than is commonly realized: it occurs whenever speed drops to half (approximately 60%) of the maximum speed limit.

[23] It can also be shown, using the technique of welfare analysis elaborated in the appendix, that the net benefit to society from such a rational move is equal to area $o+p+q+r+s+t+u+v$ (= area $o+p+q+r+s+t+w+x$).

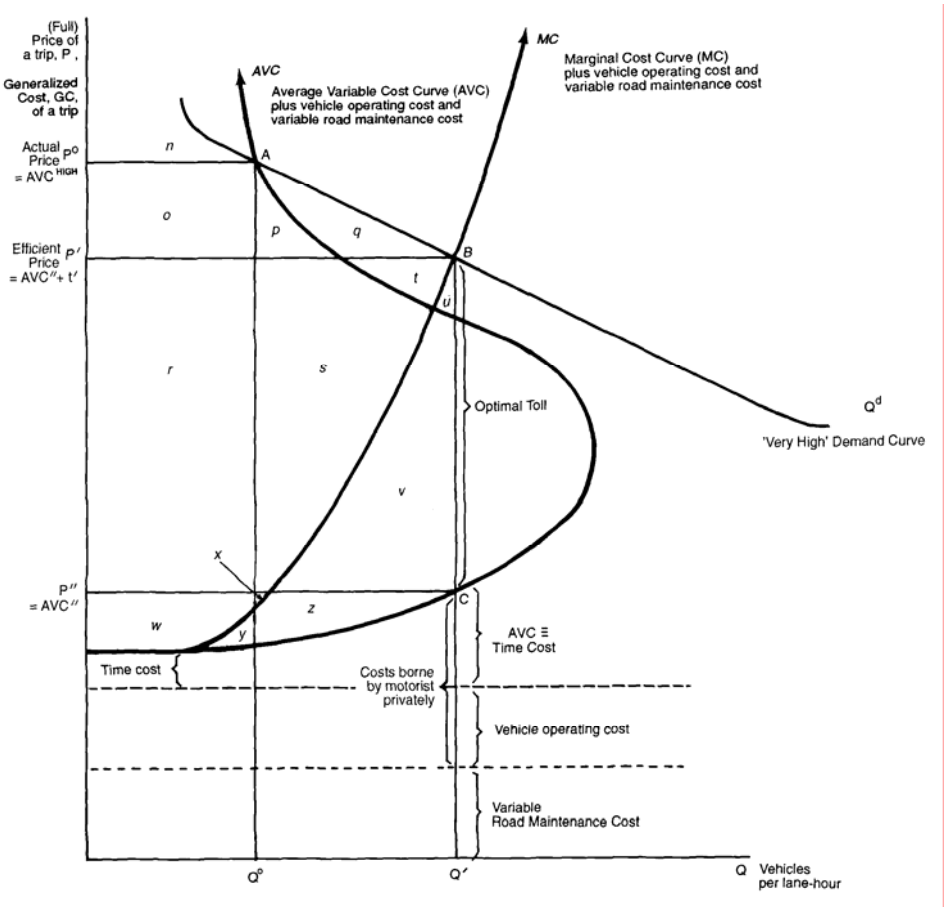

FIGURE 6: Welfare impact due to the introduction of road pricing in the peak period: short-run marginal cost pricing, 'hypercongestion' case

8.1 Policy implications

Ever since the French engineer, Jules Dupuit (1844), introduced the powerful concept of consumer's surplus to analyze issues of the efficient pricing of (toll) roads, economists have embraced that notion, extending and deepening knowledge in that area. Efficiency analysis carefully shown here indicates that society would unequivocally gain from road pricing. The question that arises naturally is why road pricing, with one single exception worldwide, has failed in the sixties, seventies and eighties to get off the ground. I show, using the analytical framework developed above, that road pricing as sold in the past is most likely doomed to political failure. This is because almost all motorists, including both the ones who are tolled and tolled off, find that they are invariably worse off as a result, except in the case of hypercongestion (see footnote for qualification).[24] The sole unmistakable gainer is the government. If road users do not

[24]The policy implications discussed here are based on the assumption of constant value of time -- the assumption used by Walters and others in deriving the average variable cost curve (see Hau (2005b), Section

perceive or are not persuaded that they benefit from the government's newly collected revenues in the form of provision of transport services and worthwhile public expenditures or receive transfers in the form of reduced general tax payments, albeit indirectly, it is highly unlikely that these groups would acquiesce to the pricing of existing roads. Cooperation would be more likely if they are guaranteed a reduction in motor vehicle-related taxes such as import duties, first registration taxes, annual license fees and/or fuel taxes. In particular, the replacement of -- rather than an addition to -- existing vehicle-related taxes by cost-effective congestion tolls would especially be welcome by road users.

A natural question then is as follows: is there a theoretical argument for dedicated funds or earmarking so that society as a whole would benefit from the implementation of road pricing? Restated another way, is there a way in which road users can act as beneficiaries and be indirectly 'compensated for' the payment of tolls by satisfying a commonly accepted notion of fairness, while carefully skirting the first-best pricing rule? I think that the answer is 'yes', although not entirely without qualifications.

9. SHORT RUN EQUILIBRIUM

The proposition that optimal pricing of and investment in a highway system parallels the short and long run equilibrium conditions of a competitive industry of a textbook commodity was first shown by Herbert Mohring (Mohring and Harwitz. 1962; Mohring, 1965, 1976).[25] Economic analysis of transport problems is simplified considerably by explicitly recognizing the traveler as both consumer and producer, and as producer-traveler he purchases factor inputs such as travel time from himself. This short-run/long-run approach is advanced both cogently and lucidly by Mohring and used by established transportation economists such as Keeler, Small, Kraus, Glaister, Morrison, Winston and Oum.

In the short run, some inputs for producing a textbook commodity are regarded as fixed.[26] Under competition, an economically efficient output level is achieved when the market-determined price equals the short-run marginal cost of producing that good. A competitive producer purchases variable inputs by hiring labor and procuring raw materials, in addition to investing in a fixed input, capital. Thereafter, the firm combines the inputs via the production process and creates commodities to be sold to a consumer. In other words, the producer uses the revenue which he obtains from selling the good at the given price to pay for the variable inputs of labor and raw materials, plus the fixed input in the form of quasi-rent on the capital equipment, normally regarded as the accounting profit. The graphs for the textbook commodity are similar to (but not exactly

---

2). My result that road pricing makes everyone worse off except the government is meant in an average sense. People's time valuations differ in reality, so that this result is modified to say that only those with fairly high valuation of time would be better off, and everyone else still worse off. The condition is that the weighted valuation associated with the time savings rectangle of area $e+f$ must exceed their total money payment of area $b+c+e+f$ based on a weighted congestion toll. Otherwise, even those who remain on the tolled road would continue to pay but would actually be worse off relative to their status quo.

[25] See survey by Winston (1985).

[26] The standard definition of short run is a situation in which some productive inputs are regarded as fixed, and hence certain costs would be fixed. However, the definition of long run refers to a situation in which all inputs vary. In the road context, the duration of the long run depends on the rate at which, say, the size of a road and hence the basic or engineering capacity, can be varied.

the same as) the case of roads considered in Figures 2 and 4.[27] In transport, the short-run marginal cost of a trip, which we derived rigorously from an engineering speed-flow curve, is to be set equal to the 'price' of a trip. Recall that transport is unusual in that the traveler is both a producer and a consumer. Analogous to the parallel case of a textbook commodity, the road user, when undertaking a trip, supplies some of his own variable inputs, which include vehicle operating cost and time cost. We have seen that the competitive level of trips exceeds the efficient quantity in the presence of congestion. Hence, because the quasi-rent of a highway facility would be dissipated due to free competition, an optimal toll should be imposed to capture this quasi-rent. Clearly, even though the dictum of short-run marginal cost pricing prevails in both cases, the optimal toll does *not* equal the short-run marginal cost of producing an output but is equivalent to the *difference* between marginal cost and average variable cost. This is a subtle but important distinction between transport and widgets.

The optimal toll is the efficient charge referred to in Mohring and Harwitz's (1962) mathematical statement of this problem and Newbery's (1989) Proposition 1. The optimal user charge is then the optimal toll plus another component required to cover the variable maintenance cost of a road discussed in Walters' (1968, pp.24) and Newbery's (1989) Proposition 2 (see Figure 2's legend).[28] As discussed before, to focus our efforts here, the optimal user charge should ultimately include air, noise pollution, accident cost and road damage externalities.

Walters (1968, Chapter 2) defines the term 'user charge' as the money charge that governmental authorities levy on travelers for the congestion cost they impose on others and for the variable maintenance cost of the road incurred due to their use of that road. In the absence of congestion, the user charge covers the unit road maintenance cost component only and would be independent of the traffic level. Walters (1968, Chapter 4) then coins the term "economic user charge (EUC)" to be equivalent to the generalized cost concept or user price employed here. This latter usage might lead to a possible misunderstanding about the demand and supply side or even double-counting and is therefore avoided here.

To recapitulate, short run equilibrium in transport occurs when the government, in the form of a highway agency, behaves in an optimizing fashion just as a private competitive firm would were it possible to organize the industry in a competitive fashion. The optimal user charge should not be set equal to the price but to the difference between the marginal cost and the average variable cost of a trip.

10. CONVERGENCE TOWARDS LONG-RUN EQUILIBRIUM UNDER CONSTANT RETURNS

So far we have confined ourselves to short-run equilibrium. The fixed cost component has been deliberately left out of the analysis of the marginal cost of a trip.[29] The motorist is oblivious to the capital cost of a road, and his behavior is independent of it. However,

[27] With standard goods, both the short run marginal and average variable cost curves can decline and swing upwards, whereas I have shown that both the short run marginal and average variable cost curves in transport *never* decline but only rise upwards.

[28] More precisely, Newbery's (1989) Propositions 1 and 2 include the *invariate* maintenance cost due to weather.

[29] The fixed cost of a firm is defined to be the minimum amount of outlay necessary to start production.

from the highway agency's planning point of view, the capital cost of a road is very much taken into account. Once a highway is built, however, it is regarded as sunk. The sunk cost of a road, once incurred, is irrelevant to a planner: only current and future costs, not historical cost, serve as a correct guide to planning future investment. Since the variable road maintenance cost is assumed to be constant, the marginal cost of a trip thus remains the same.

In the long run, however, a highway agency can vary the fixed capital input by expressway expansion, if the investment is deemed justifiable. On the other hand, if a rural road has been built as a result of past planning errors, it can be allowed to deteriorate or be downgraded (or be even auctioned off). Expanding a road until the additional benefit equals the additional cost of building it would yield maximal net benefit to the community. How might this be done without resorting to a full-scale cost-benefit study?

To see how this might be done, we introduce the fixed cost, i.e., the cost of construction, together with the 'invariate' maintenance, depreciation and operating costs of a road that are incurred by a governmental authority in Figure 7.[30] We then convert the entire fixed cost into the cost per time period of a unit of capital for utilizing the flow of highway services. This is done in order to make it commensurate with the average variable cost of a trip discussed thus far. The summation of the short-run average fixed cost and the average variable cost curve yields the average total cost curve. Charging the optimal toll of the distance t′ in Figure 7 seems to be more than sufficient to cover the short-run average fixed cost of the facility. In this case, the optimal toll, t′, exceeds the short-run average fixed cost of the facility, SRAFC′, by the unit profit difference of π'. In general, there is no *a priori* reason why toll revenue collections based on short-run marginal cost pricing cannot cover the non-use-related costs of a given highway facility.

In the case of a textbook commodity, whenever the quasi-rent being earned by a firm's existing capital equipment exceeds its cost, there is an incentive to expand production. Ultimately, the quasi-rent earned by the existing capital equipment would then be equal to its (fixed) cost.[31] Putting it another way, upon seeing the existence of economic profits, other firms enter the industry also, shifting the industry supply outward, increasing output and lowering price as a result. The unrestricted mobility of resources and the entry and exit of firms serve as the instrument by which profits would be competed away in due course. When capital is freely varying, long-run equilibrium is reached when zero economic profit occurs. Equivalently stated, the quasi-rent earned on the firm's capital equipment equals its cost, i.e., the market return of the cost of reproducing the invested capital.[32] This condition holds under constant returns to scale, where a proportionate increase in all inputs is compatible with the same proportionate

---

[30] The term *invariate,* i.e., non-traffic related maintenance cost is found in Walters (1968, pp. 23).

[31] The quasi-rent to a firm is the accounting profits *plus* interest expense on borrowed funds, if any. The accounting profit is the firm's total revenue less its contractual costs, including interest expense on borrowed funds, wages to labor, cost of raw materials, and rental cost of leased buildings. Accounting profit less the market return of the owner-supplied assets is the economic profit. Alternatively, the quasi-rent on invested capital is the total revenue less the variable costs, i.e., including wages, cost of raw materials and rental cost, but excluding interest expense on borrowed funds. Total revenue less total variable and fixed costs yields economic profits. The fixed cost of invested capital includes the entire opportunity cost of capital, regardless of whether funds are borrowed or not. (See the third footnote of Appendix (footnote 45) and Mohring (1976, pp. 8-11)). I am grateful to Herbert Mohring for clarifying these points.

[32] The (opportunity) cost of a resource used *here* is the highest return of it *elsewhere*, hence the cost is the market return on reproduction cost and not historical cost.

increase in outputs. Given fixed factor prices, total cost also doubles, so marginal cost remains constant in the long run. With a slight but crucial modification, this analysis carries over to the case of roads. When the quasi-rent of the existing capital stock exceeds the normal market return on the cost of reproducing the invested capital plus the highway facility's invariate maintenance and depreciation costs, new investment is expected to find its way in that road segment of the highway industry if the appropriate price signals are given. Equivalently, in the long run, if toll revenues -- which recover quasi-rents throughout time periods -- exceed the entire fixed cost of the existing facility, the highway authority would have the appropriate incentive to expand a stretch of that road until all economic profits are eroded away. As we have seen in the case of roads, the variable cost is composed of the user-supplied time and operating costs and are fully self-financing. The non-use related costs are then financed separately by the road agency via toll revenue collections. In this way, full costs are covered and there is no need to raise charges when there are constant returns to scale.

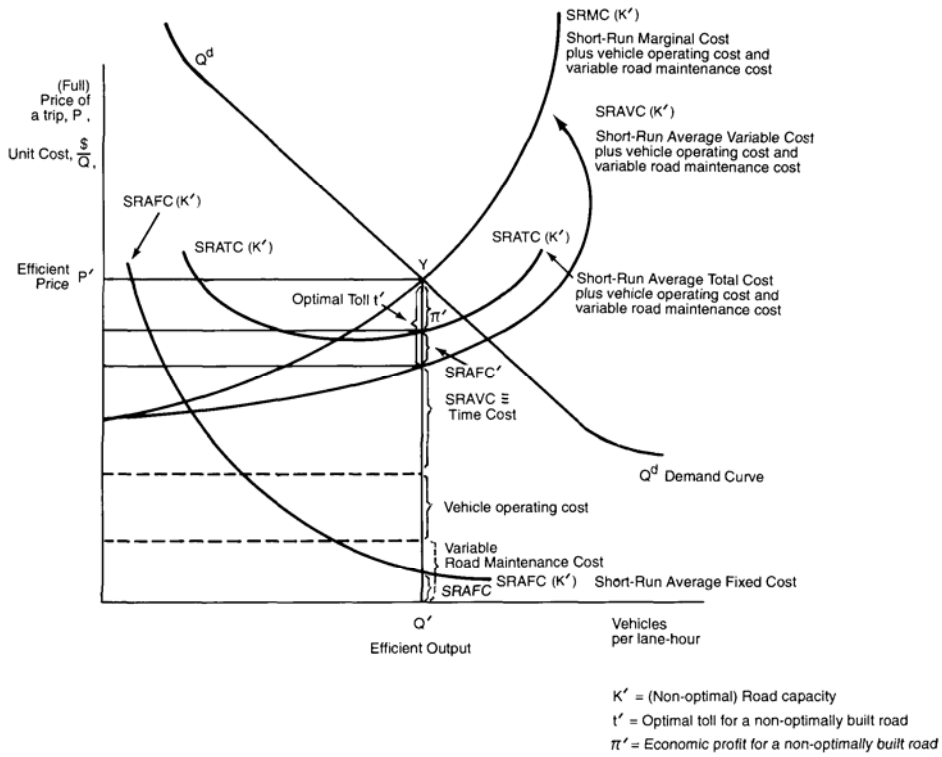

FIGURE 7: Introducing the (short-run average) fixed cost, SRAFC, of a road, short-run optimal toll with economic profit

11. OPTIMAL INVESTMENT

In the long run, toll revenues would then exactly cover the amortized cost of construction, invariate maintenance and depreciation costs of roads -- a powerful result first shown by Mohring and Harwitz (1962, Chapter 2) and Mohring (1965) -- under the technical conditions of constant returns to scale in road construction, maintenance and road use. Constant returns to scale intuitively mean that the cost of building and

maintaining an expressway is proportional to the capacity. Constant returns to road use yields an intuitive interpretation: travel time depends solely on the volume-capacity ratio. If the engineering capacity and the traffic flow were doubled, unit travel times would remain the same.[33] The final long-run equilibrium is shown in Figure 8. By faithfully pursuing the policy of marginal cost pricing of a trip by charging a congestion toll -- the difference between the marginal cost and the unit variable (time) cost -- and by expanding or appropriately reducing the capacity of the road until there is zero economic profit, the output (of vehicle-kilometers per lane-km per hour) is considered optimal. At a moment in time for an existing road, output is optimal in the sense that, given the marginal-cost price, the efficient level of trips is achieved. Undertaking either more or less trips would involve lowering the net benefit to the community. In the long run, output would be 'doubly' optimal if it is the efficient level of trips for that link of road which has been optimally built. Diagrammatically, not only does the implementation of a congestion toll internalize the external congestion cost, it can be seen that the toll covers the short-run average fixed cost of the road in a stationary state. Recall that for homogeneous traffic the average fixed cost of a road is simply defined to be the difference between the average total cost and the average variable cost. Clearly, collecting a unit toll would cover the entire average fixed cost of the road and yield zero profit only because the existence of economic profit or loss serves as a quasi-market mechanism in the investment decision of whether to expand or contract the highway capacity. With zero profit, the minimum point of a short-run average total cost curve is obtained. For any given level of output, the minimum total cost of yielding this output would be obtained only if the optimal investment level in capacity had been chosen. Looking at it the other way, the optimum size of a road is obtained by drawing a locus of all the minima of the various short-run average total cost curves of different sizes (and capital costs) of highways and choosing that particular size of the road associated with the point where the demand curve intersects its marginal cost curve. We do this because the demand reflects motorists' maximum willingness-to-pay and hence the incremental benefit of the last trip. Since the minimum points of the short-run average total costs under constant returns are of the same height, the long-run average total cost is a horizontal line tangent to all the minima. With the long-run average total cost being constant, so also is the long-run marginal cost. Hence, it is only at the minimum SRATC point that *long-run marginal cost pricing* holds.

---

[33] *If* 1) the capital and invariate maintenance cost of highway capacity, KC, is directly proportional to the engineering capacity, K, i.e., KC (K) = *a*K, where *a* is a constant, *then* there exists constant returns to scale in highway construction (and invariate road maintenance). (In mathematical jargon, KC is homogeneous of degree one in capacity.) The engineering capacity is measured by lane-width and is treated as a continuous variable. Further, *if* 2)(a) traffic can be expressed in terms of a homogeneous unit, Q, in vehicles per lane-hour, and the time cost function AVC(Q,K) depends directly on the traffic flow but is inversely related to the capacity and (b) *if* doubling both highway capacity and traffic flow result in the travel time of a trip remaining the same, *then* there exists constant returns to road use. (Mathematically, the AVC function is homogeneous of degree zero in traffic volume and capacity.) With constant returns to road use, AVC(Q,K) can be formally rewritten as AVC(Q/K), where Q/K is the volume-capacity ratio. Since unit vehicle operating and variable road maintenance costs are both independent of the level of output, and capital cost, KC, is proportional to lane expansion, ATC(Q,K) = ATC(Q/K) holds also. These two technical conditions are crucial to Mohring and Harwitz's (1962, pp. 85-90) so-called theorem and to Keeler and Small's (1977) extension of Mohring's theorem.

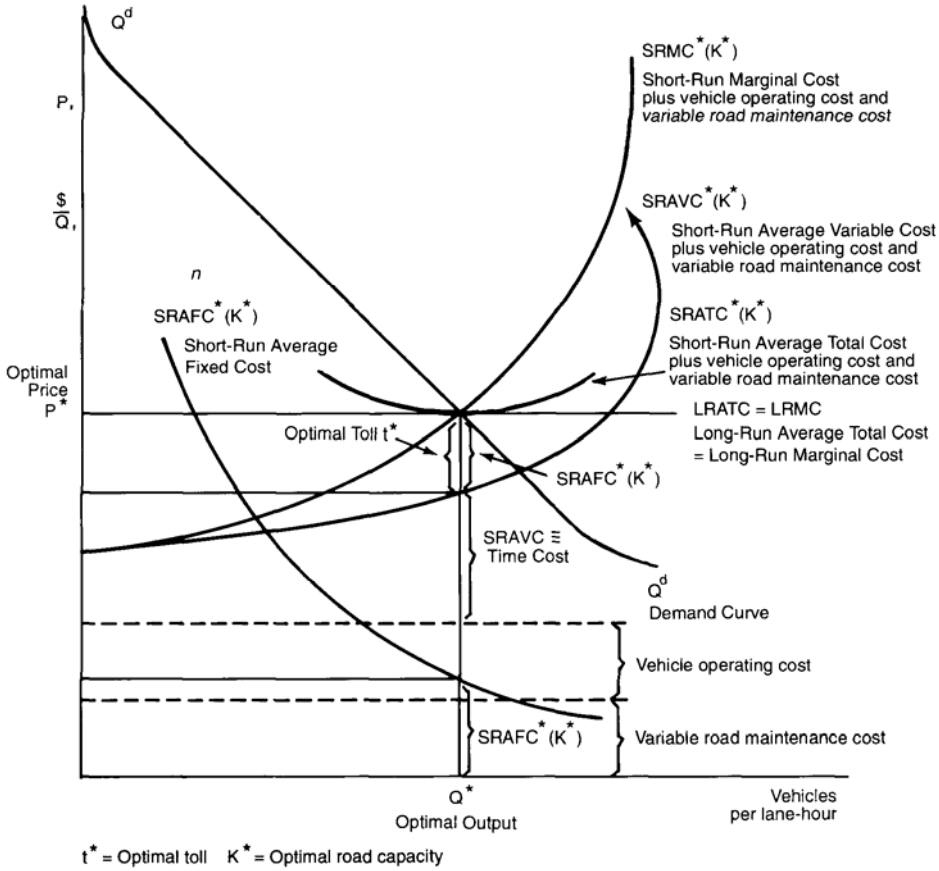

FIGURE 8: Long-run equilibrium of an optimally designed road with both optimal pricing and optimal investment

12. LONG-RUN VS. SHORT-RUN MARGINAL COST PRICING[34]

Intuitively, the long-run marginal cost of producing a trip yields the total cost of undertaking a trip to the community when all fixed and variable inputs can be varied continuously in the long run. Proponents of long-run marginal cost pricing argue that the market return to capital investment would presumably be fully covered. Yet the equivalence of short-run and long-run marginal cost pricing holds only in certain cases, including the static demand and single period case considered here. As shown in Figure 8, long-run marginal cost pricing would cover *all* the variable costs, including time cost, vehicle operating cost and variable road maintenance cost, *plus* the fixed construction,

[34] Prest (1969, pp.8), Walters (1968, pp.33), Bennathan and Walters (1979, pp.33) and Bird (1976, pp.33-39) argue for short-run marginal cost pricing whereas others argue for long-run marginal cost pricing. Since the issue of long-run vs. short-run marginal cost pricing has been with us for some time, a clarification is in order (see the debate between Jordan (1983a, 1983b, 1985) and Vickrey (1985)). Vickrey points out that the concept of long-run marginal cost becomes obfuscated when several demand periods, e.g., peak, interpeak and off-peak, occur diurnally, given the case of a transportation infrastructure that has already been constructed.

invariate maintenance, depreciation and operating costs of the road. In fact, short-run marginal cost pricing covers the entire capital cost of the facility just as much as long-run marginal cost pricing does, as can be seen diagrammatically in Figure 8. After all, both the short-run and long-run marginal and average costs are equal in the long run, with both sets of cost curves intersecting the demand curve at the same point. However, if a road is not optimally constructed but underbuilt, then long-run marginal cost pricing would send out too low a price signal, thereby exacerbating congestion. Short-run marginal cost pricing, on the other hand, would give the correct signal of higher willingness-to-pay and also yield positive toll revenues and economic profits as a by-product. Short-run marginal cost pricing is the rule to use whenever long-run equilibrium is not reached (and of course when it is). Looking at it another way, if short-run marginal cost is below long-run marginal cost at the current output, it means that the road has been overbuilt. But, of course, this does not mean that the size of the expressway should be or indeed can be varied instantaneously whenever demand fluctuates daily. Rather, it means that the price ought to be varied according to demand patterns using short-run marginal cost pricing.

Indeed, Professor William Vickrey has emphatically argued that there can be *no* solution to the urban transportation problem without peak-load pricing. Proper time-of-day pricing can be implemented only using short-run marginal cost. (We shall explore this point further in the section on demand variability.) Pursuing short-run marginal cost pricing period by period by varying road capacity incrementally over time would not only guarantee the best use of society's resources but would also enable road agencies to recover all costs -- as an incidental by-product -- in the long run. It is therefore recommended that short-run marginal cost pricing be used since the concept of long-run marginal cost cannot be unambiguously defined whenever cyclical variations in demand are involved.

13. TRADE-OFF BETWEEN INDIVIDUALS' TIME AND TREASURY ACCOUNTS

Another way of obtaining the optimal investment level for roads is to answer the following question: what is the minimum cost to the community of road building, taking into account *both* the highway agency's desire to minimize the fixed cost of capital facilities and the traveling public's desire to save time?[35] By minimizing the total cost -- the sum of these two costs: the variable (time) cost of trip makers and the fixed cost of the governmental authority -- a trade-off is found between individuals' time and the treasury's accounts. Given a non-optimal capital stock (K′) associated with a particular highway, as in the previous graph, Figure 7, it can be seen that the least cost for the community involves having a road that is too small, for the level of demand depicted. Expanding the capacity of the road may reduce the user's trip cost evaluated at a given traffic level. Long-run equilibrium is reached when the minimum point of the short-run average total cost curve (equals the short-run marginal cost curve) intersects the demand curve. For the governmental authority, road capacity is a choice variable. By increasing its size, the volume-capacity ratio drops in the short run, and so does time cost. However, the cost of road capacity increases. Intuitively, the highway agency continues to expand the road until the marginal benefit from saving users' time costs is just offset

[35] Solving the problem of cost minimization is equivalent to solving the problem of maximization of net benefit to the community.

by the marginal cost of one unit of capacity.[36] It is at the output, Q* with an optimally built road K*, in Figure 8 that the valuation of the last trip taken just equals its marginal cost, that is, the incremental cost of the trip to others, the motorist's own time cost in congested traffic, plus the vehicle operating cost and the road maintenance cost.[37] The highway agency, by setting an optimal road user charge which is equal to the congestion toll and the variable road maintenance cost component, would be able to induce the motorist to travel up to the point where the price of a trip equals its short-run marginal cost. By pursuing this pricing policy for each stretch of road, the use of a non-optimal, existing highway network would be optimized. Further, by expanding highway capacity up to the point where the quasi-rent of each capital facility just covers the cost of reproducing it, with zero (economic) profit remaining, the net benefit to the community would be maximized. *By symmetry*, abandoning or downgrading roads is necessary when economic losses occur. The decision of not maintaining roads is tantamount to the act of disinvesting roads.

Henceforth, to simplify both our discussion and the diagrams, we ignore the individual's vehicle operating cost and the variable road maintenance cost since they are self-financing.[38] Note that our conclusions thus far hold under the assumption of constant returns to scale and perfect divisibility of roads. Consider a three-lane road with capacity K_3 in Figure 9, where output is now measured in vehicles per hour.[39] Since the highway authority has efficiently priced the road by setting the congestion toll t_3* and optimally built the road by investing at K_3*, it can be seen that the toll revenue covers the fixed cost of the road. So far we have simply translated Figure 8 into Figure 9 with the costs borne by motorists conveniently left out but not forgotten. Assume that both traffic volume and road width, i.e., the number of lanes, are doubled and that the inputs to each of the component costs under constant returns are doubled, then i) the fixed costs of construction, maintenance and depreciation, ii) the variable time costs, and iii) the total costs are all doubled.[40] Intuitively, the geometric doubling of the rectangular areas of road construction and maintenance costs is synonymous with the condition of zero economies of scale (given fixed input prices) in road construction and invariate maintenance. Similarly, the horizontal doubling of the rectangular areas of the time costs

---

[36] Formally, given a particular level of output, the cost-minimizing authority would expand the road up to the point where the marginal valuation in time savings due to a unit increase in capacity, $- Q \cdot \Delta AVC(Q,K)/\Delta K$, equals to the marginal cost of a unit of capacity, R. R is the rental cost per time period of capacity, which includes the invariate maintenance and other operating costs of a road, depreciation and imputed interest on invested capital. The negative sign would offset the inverse relationship of AVC and K, yielding a positive magnitude for the entire term. Alternatively, the road is to be expanded up to the point where the marginal external congestion cost just offsets the marginal cost of investment in capacity. This is the second optimality rule: the optimal investment in capacity rule.

[37] The superscript * symbol indicates that that variable is optimized.

[38] I have therefore grouped the non-traffic related maintenance and operating costs as part of the short-run average fixed cost curve. Bear in mind that the variable road maintenance cost (not drawn in Figure 9) is being recovered by a separate road user charge component as shown in Figure 2. With the condition of zero economies of scale in total road maintenance, the total maintenance cost is thus exactly doubled. Also, economic profits from now on will be referred to as 'profits'.

[39] Three lane roads are found in Australia where they are referred to as, 'two-and-a-half lane roads' by Hoban (1987).

[40] The stringent assumption of a road being finely divisible will be relaxed and indivisibilities introduced later on.

supplied by individual users is akin to the technical condition of zero economies of scale in road use.

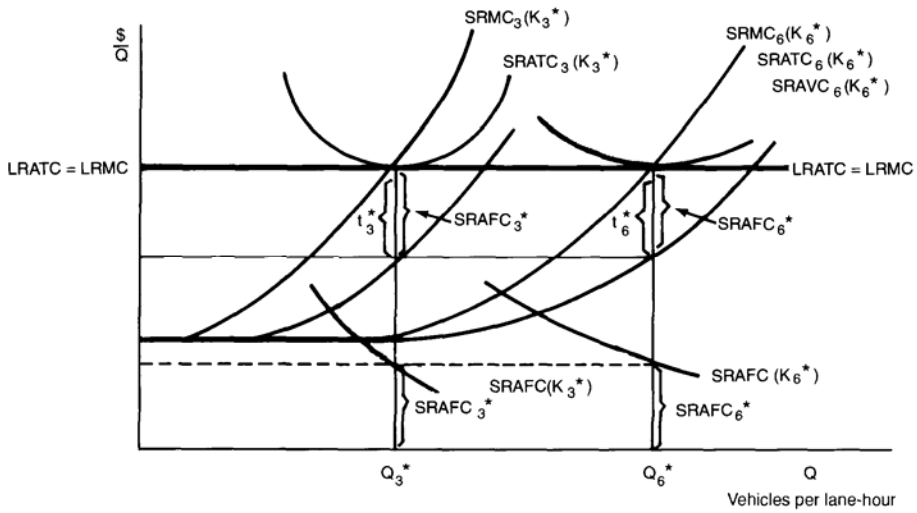

FIGURE 9a: Constant returns to scale with road divisibility: doubling optimal road capacity (K) and Traffic (Q), result in doubling fixed cost, variable cost and total cost (FC, VC, TC) and toll revenues (t_L*· Q_L*)

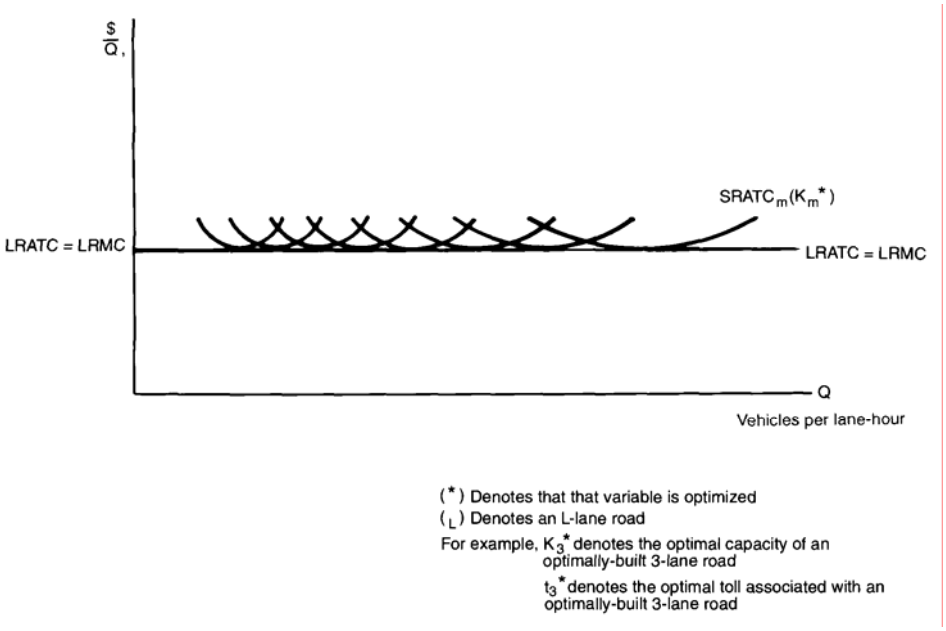

FIGURE 9b: The relationship between short-run average total cost and long-run average total cost and marginal cost with perfect road divisibility and constant returns to scale

14. FIRST-BEST OPTIMAL PRICING AND INVESTMENT RULES

14.1 Empirical considerations

By estimating behavioral travel demand functions using state-of-the-art logit mode choice models, one could obtain empirical estimates of marginal valuation of time (as a function of income levels) (Hau, 1986). When combined with a fine-grained, parametric transportation corridor supply model of the San Francisco Bay Area (Talvitie et al., 1978), multi-market demand and supply could be equilibrated and cost-benefit analysis of alternative policies performed (Hau, 1987). Heuristically, the adjustment process towards the final equilibrium parallels that of the cobweb equilibrium model of adjustment. Optimal tolls and welfare gains and losses could thus be simulated with appropriate specification of the marginal travel time function. The results obtained are for a short-run equilibrium model of demand and supply.

Even with poor data, one could make some progress in empirical work. Given an estimate of a speed-flow curve and the corresponding travel-time flow curve, we know how these engineering curves can be converted to a short-run average variable cost curve of a trip, using an estimate of the value of time. A 'supply' elasticity estimate together with a value for unit variable cost would yield a one-to-one correspondence between the short-run average variable cost and the marginal cost (see footnote 17 for formula). A rough estimate of the demand elasticity and the traffic level of a particular road would yield a first order approximation of the proper congestion toll. Now, in order to maximize aggregate net benefit, two operating rules should be followed by the road authority.

14.2 The first rule -- the optimal pricing rule

For each stretch of road, short-run marginal cost pricing is fulfilled by setting a toll at the excess of short-run marginal cost over short-run average variable cost. The intuition is that this congestion toll would serve to internalize the congestion cost that a driver imposes on others. In addition, the motorist is charged another component which covers the variable maintenance and operating costs of a road which he imposes on the road authority. Thus the public authority's imposition of an optimal road user charge would cover both the external cost of congestion as well as the variable road maintenance and operating costs.

14.3 The second rule -- the optimal capacity rule

Under constant returns to scale and optimal pricing, whenever economic profit is found in the operation of one road link, the procedure would be to expand the capacity of that stretch of road. The existence of a loss under short-run marginal cost pricing suggests that the road has been overbuilt. By altering the capacity of each road in the long run according to the quasi-market signal of profits and losses, the entire highway network's investment level in capacity would be optimized, with the fixed cost of each road covered. Alternatively, the road authority, by trading its direct resource costs against individuals' travel time, follows the rule of setting the marginal travel time savings equal to the marginal cost of investment for an additional unit of capacity. The capacity of a road is expanded until the marginal capital cost equals the marginal (external) congestion cost.

14.4 Perspective on the result

What we have described is the long-run equilibrium of an optimally designed capacity of a road network under constant returns. If the road authority were: 1) to pursue the efficiency-enhancing policy of pricing according to the marginal cost of a trip, and 2) to minimize the sum of the direct resource costs of providing a road and the value of user-supplied travel time inputs, then the road would be both efficiently utilized and optimally expanded. Notice that the optimally designed road has a positive amount of external congestion cost. This results from the road agency's desire to minimize both the sum of the direct cost of the investment in capacity and individual drivers' travel time cost. In our simple framework, congestion delay would never be entirely absent, contrary to what environmentalists and road users would prefer, because achieving zero congestion is very costly to the community. In other words, an optimal amount of congestion externality is a valid concept, just as an optimal amount of pollution has long been recognized in the environmental economics literature. What if there is no congestion on a particular road? Zero congestion means that that stretch of road has been overbuilt (or priced non-optimally) and should perhaps be downgraded or even abandoned. If excess capacity occurs all the time, the road possesses the non-rival consumption characteristic of a pure public good. Then we are faced squarely with the standard task of provision of public goods. If resources are plentiful, financing the shortfall via general revenue taxation has been the conventional dictum. Similarly, if public resources are scarce, the opportunity cost of public funds must be accounted for.[41] *If* lump sum taxation is infeasible and resources are severely inadequate because of political constraints, then it is possible for one to consider the feasibility of financing via Ramsey pricing.[42] If it is uncertain whether the (marginal) deadweight losses from general revenue financing exceed those obtained from Ramsey taxation, financial and other considerations such as equity may then have to be appealed to in order to justify the potential use of Ramsey pricing in the road sector.

By contrast, a road would possess the rival consumption characteristic of a private good when excess demand occurs. Hence a congested road is also regarded as a congested variable-use public facility. Because of this mixed good nature, and based on the theory derived here from first principles, the provision of road services ought to reside with the public sector. Under the condition of constant returns, the optimal toll revenue, which captures the quasi-rent earned from the invested capital, would cover the

[41] In the companion paper (Hau, 2006), I present calculations of the opportunity cost of financing alternative charging mechanisms.

[42] Frank Ramsey's (1927) inverse elasticity formula under the case of independent demands was discovered in response to Pigou's question of how to set tax rates in order to minimize the welfare losses associated with meeting a tax revenue requirement. The problem of the choice of optimal tax rates which are subject to a revenue constraint is formally equivalent to that of the setting of optimal prices which are budget-constrained. Ramsey pricing in the presence of externalities requires that it be computed on the basis of marginal cost and a *fraction* of the marginal external cost (Oum and Trethaway, 1988). Clearly, in the presence of leakages, non-excludability and partial rivalry in consumption, the requirements for the implementation of Ramsey pricing are considerably more stringent than those of pursuing marginal cost pricing. On conceptual, empirical and implementation grounds, marginal cost pricing is superior to even the simplest form of Ramsey pricing: the inverse elasticity rule. The computation of Ramsey prices, for instance, requires the estimation of marginal cost as a prerequisite and the determination of revenue targets. Despite all the problems with which Ramsey pricing is fraught, research into this interesting issue is potentially useful. After all, increasingly tight fiscal constraints, countries will demand alternative funding mechanisms.

entire fixed cost of the road in the long run. No residual or overhead cost need be allocated. If profit exists, then it is because there is insufficient road capacity (or pricing at a level above marginal cost). The road is therefore not in long-run equilibrium. The existence of profit serves as a surrogate market signal to expand capacity. The motto is as follows: "If a road makes money (i.e., economic profit), expand it, else not." Similarly, if a road loses money, it suggests that planners made the wrong decision or were given over-optimistic forecasts of travel demand. In that case, marginal cost pricing is still to be adhered to, with the congestion toll set close to nil. A user charge component is still needed to cover the variable road maintenance cost. Thus it may even be worthwhile to abandon a money-losing road and save on any annual invariate maintenance costs that might arise. Efficient pricing, financial viability and cost recovery are therefore entirely consistent with one another under constant returns to scale in long-run equilibrium.

ACKNOWLEDGMENTS

A rudimentary version of this research paper was presented at the 2nd International Conference on Privatization and Deregulation, Tampere, Finland, June 16-20, 1991. Part of the findings of this research was presented at the 6th World Conference on Transport Research, Lyon, France, June 29 - July 3, 1992; Seminario de Tarificación Vial 1992 (Road Pricing Seminar), sponsored by the Ministry of Transport, Santiago, Chile, July 20-21, 1992; the Symposium on Road Pricing, sponsored by the Swedish National Road Administration and the Department of Regional Planning, Royal Institute of Technology, Sigtuna, Sweden, November 9-10, 1992; and the Southwest Congestion Pricing Conference, sponsored by the University of Houston's Center for Public Policy and Citizens Advocating Responsible Transportation, Houston, Texas, January 4-5, 1993. Comments by participants at these places are gratefully acknowledged. I have benefited from the discussions and comments of many people on the controversial subject of road pricing and I take this opportunity to thank William Vickrey, Michael Beesley, Kiran Bhatt, Richard Bird, Andrew Evans, Stephen Glaister, Phil Goodwin, José Gómez-Ibáñez, Chris Hendrickson, Peter J. Hills, Jake Jacoby, Jan Jansson, Peter Jones, Theodore Keeler, Damien Kulash, Odd Larsen, Henning Lauridsen, Douglass B. Lee, Peter Mackie, Anthony May, Robert McGillivray, Herbert Mohring, Richard Musgrave, Yew Kwang Ng, Esko Niskanen, Tae Oum, Farideh Ramjerdi, Gabriel Roth, Kenneth Small, Antti Talvitie, Tom Tietenberg, Burkhard von Rabenau and Sir Alan Walters. I am also grateful to World Bank staff, including Esra Bennathan, Philip Blackshaw, José Carbajo, Anthony Churchill, Shanta Devarajan, Asif Faiz, Jeffrey Gutman, Clell Harral, Ian Heggie, Chris Hoban, Gregory Ingram, Frida Johansen, Kyu Sik Lee, William McCleary, William Paterson, Richard Scurfield, Zmarak Shalizi and Larry Summers.

Support services from Margarat Callan, Pam Cook, Barbara Gregory, Beatrice Sito and Gabriella Steinmetz are gratefully acknowledged.

The work that is described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (HKU7126/04E).

These acknowledgements do not mean that they share in the views expressed in this paper. I alone retain full responsibility for its contents.

REFERENCES

American Association of State Highway Officials (AASHO) (1960) *Road User Benefits for Highway Improvements*. Committee on Planning and Design Policies, American Association of State Highway Officials, Washington, D.C., pp. 1-152.

Baumol, W.J. and Lee, K.S. (1991) Contestable markets, trade, and development. *The World Bank Research Observer*, **6**, 1-17.

Baumol, W.J. and Oates, W.E. (1975, 1988) *The Theory of Environmental Policy*. First Edition, 1975, and Second Edition, 1988, Cambridge University Press, New York.

Baumol, W.J., Panzar, J.C. and Willig, R.D. (1982, 1988) *Contestable Markets and the Theory of Industry Structure*. First Edition, 1982, and Revised Edition, 1988, Harcourt Brace Jovanovich and Academic Press, San Diego, California.

Beesley, M.E. and Walters, A.A. (1970) 'Some problems in the evaluation of urban road investments' and 'A Reply'. *Applied Economics*, **1**, 241-259, and **4**, 317-320.

Behbehani, R., Pendakur, V.S. and Armstrong-Wright, A.T. (1984) *Singapore Area Licensing Scheme: A Review of the Impact*. Water Supply and Urban Development Department, The World Bank, Washington, D.C., pp. 1-55.

Bennathan, E. and Walters, A.A. (1979) *Port Pricing and Investment Policy for Developing Countries*. Published for the World Bank, Oxford University Press, Washington, D.C.

Bird, R.M. (1976) Charging for public services: a new look at an old idea. Canadian Tax Papers No. 59, December, Canadian Tax Foundation, Toronto. Excerpt reprinted as "Principles of Public Pricing" In Pricing Policy for Development Management, Gerald M. Meier (1983), pp. 171-176, ed., EDI Series in Economic Development, published for The World Bank, Baltimore, Johns Hopkins University Press, Maryland.

Boadway, R.W. and Wildasin, D.E. (1984) *Public Sector Economics*. Second Edition, Little, Brown and Company, Boston, Massachusetts.

Buchanan, J.M. (1965) An economic theory of clubs. *Economica*, **32**, 1-14.

Button, K.J. (1982) *Transport Economics*. Heinemann Educational Books Ltd, London.

Catling, I. and Roth, G. (1987) Electronic road pricing in Hong Kong: an opportunity for road privatization? Paper presented to the Transportation Research Board conference on 'Roles of Private Enterprise and Market Processes in the Financing and Provision of Roads', and *Transportation Research Record*, **1107**, 51-55.

Coase, R.H. (1960) The problem of social cost. *Journal of Law and Economics*, **3**, 1-44.

Coase, R.H. (1988) *The Firm, the Market and the Law*. University of Chicago Press, Chicago.

Cooter, R.D. (1982) The cost of coase. *Journal of Legal Studies*, **11**, 1-34.

Downs, A. (1962) The law of peak hour expressway congestion. *Traffic Quarterly*, **16**, 393-409.

Dupuit, J. (1844) On the measurement of the utility of public works. Annales des Ponts et Chaussées, 2nd series, Vol. 8, translated from the French essay "De la mesure de l'utilité des travaux publics," by R. H. Barback for International Economic Papers, No. 2, 1952, pp. 83-110. Reprinted in Transport, ed., Denys Munby, Penguin Modern Economics, London, 1968, pp. 19-57.

Dupuit, J. (1849) On tolls and transport charges. Annales des Ponts et Chaussees, 2nd series, 4th part, 1849, pp. 207-248, translated from the French essay "De l'influence des pearges sur l'utilite des voies de communication," by Elizabeth Henderson for International Economic Papers, No. 11, 1962, pp. 7-31.

Glaister, S. (1981) *Fundamentals of Transport Economics*. Basil Blackwell, Oxford.

Gerlough, D.L. and Huber, M.J. (1975) *Traffic Flow Theory, A Monograph*. Special Report 165, Transportation Research Board, National Research Council, Washington, D.C.

Gwilliam, K.M. and Mackie, P.J. (1975) *Economics and Transport Policy*. George Allen and Unwin, London.

Gwilliam, K.M. and Nash, C.A. (1972) Evaluation of urban road investments – 'A Comment' and 'A Rejoinder'. *Applied Economics*, **4**, 307-315 and 321.

Haight, F.A. (1963) *Mathematical Theories of Traffic Flow*. New York, Academic Press.

Hau, T.D. (1986) Distributional cost-benefit analysis in discrete choice. *Journal of Transport Economics and Policy*, **20**, 313-338.

Hau, T.D. (1987) Using a Hicksian approach to cost-benefit analysis in discrete choice: an empirical analysis of a transportation corridor simulation model. *Transportation Research*, **21B**, 339-357.

Hau, T.D. (1988) Urban transport. In *The Economic System of Hong Kong*, H.C.Y. Ho and L.L.C. Chau (eds.), Asian Research Service, pp. 190-226.

Hau, T.D. (1990) Electronic road pricing: developments in Hong Kong 1983-89. *Journal of Transport Economics and Policy*, **24**, 203-214.

Hau, T.D. (1992a) Economic Fundamentals of Road Pricing: A Diagrammatic Analysis. World Bank Policy Research Working Paper Series, WPS 1070, World Bank, Washington D.C., pp. 1-96.

Hau, T.D. (1992b) Congestion Charging Mechanisms for Roads: An Evaluation of Current Practice. World Bank Policy Research Working Paper Series, WPS 1071, World Bank, Washington D.C., pp. 1-99.

Hau, T.D. (1998) Congestion pricing and road investment. In Kenneth J. Button and Erik T. Verhoef (eds.) Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility, Chapter 3, Edward Elgar, Cheltenham, U.K. and Northampton, M.A., U.S.A., pp. 39-78.

Hau, T.D. (2005b) Economic fundamentals of road pricing: a diagrammatic analysis, Part II – Relaxation of assumptions. *Transportmetrica*, **1**, 119-149.

Hau, T.D. (2006) Congestion charging mechanisms for roads: an evaluation of current practice. *Transportmetrica* (forthcoming).

Hoban, C.J. (1987) Evaluating Traffic Capacity and Improvements to Road Geometry. World Bank Technical Paper No. 74, October, The World Bank, Washington, D.C., October, pp. 1-145.

Johansen, F. (1989) Earmarking, Road Funds and Toll Roads. A World Bank Symposium, Infrastructure and Urban Development Department, Report INU45, The World Bank, Washington, D.C., pp. 1-196.

Jordan, W.J. (1983a) Heterogeneous users and the peak load pricing model. *The Quarterly Journal of Economics*, **93**, 127-138.

Jordan, W.J. (1983b) The theory of optimal highway pricing and investment. *Southern Economic Journal*, **50**, 560-564.

Jordan, W.J. (1985) Capacity costs, heterogeneous users, and peak-load pricing. *The Quarterly Journal of Economics*, **95**, 1335-1337.

Julius, D. and Alicbusan A.P. (1989) Public Sector Pricing Policies: A Review of Bank Policy and Practice. Policy, Planning, and Research, Working Paper Series 49, The World Bank, Washington, D.C.

Keeler, T.E. and Small, K.A. (1977) Optimal peak load pricing, investment, and service levels on urban expressways. *Journal of Political Economy*, **85**, 1-25.

Keeler, T.E., Small, K.A. and Associates (1975) with contributions by George S. Cluff, Jeffrey K. Finke, Leonard A. Merewitz, Randall J. Pozdena and Philip A. Viton, The Full Costs of Urban Transport: Part III, Automobile Costs and Final Intermodal Cost Comparisons, Monograph No. 21, Institute of Urban and Regional Development, University of California, Berkeley, pp. 1-154.

Knight, F.H. (1924) Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics*, **38**, 582-606. Reprinted in Readings in Welfare Economics, eds., Kenneth J. Arrow and Tibor Scitovsky, American Economic Association Series, Richard D. Irwin, Inc., 1969, pp. 213-227.

Kraus, M.C., Mohring, H.D. and Pinfold, T. (1976) The welfare costs of non-optimum pricing and investment policies for freeway transportation. *American Economic Review*, **66**, 532-547.

Lee, D.B. (1982) Net benefits from efficient highway user charges. *Transportation Research Record*, **858**, 14-20.

May, A.D. (1990) *Traffic Flow Fundamentals*. Prentice Hall, Englewood Cliffs, New Jersey.

Mills, D.E. (1981) Ownership arrangements and congestion prone facilities. *American Economic Review*, **71**, 493-502.

Ministry of Transport (1964) *Road Pricing: The Economic and Technical Possibilities*. Her Majesty's Stationery Office, London, pp. 1-61.

Mishan, E.J. (1988) *Cost Benefit Analysis*. Fourth Edition, Unwin and Hyman, London.

Mohring, H.D. (1965) Urban highway investments. In Measuring Benefits of Government Investments, Robert Dorfman (ed.), papers presented at a Conference of Experts held November 7-9, 1963, The Brookings Institution, Washington, D.C., pp. 231-291.

Mohring, H.D. (1976) *Transportation Economics*. Ballinger Press, Cambridge, Massachusetts.

Mohring, H.D. and Harwitz, M. (1962) *Highway Benefits: An Analytical Framework*. Northwestern University Press, Evanston, Illinois.

Morrison, S.A. (1986) A survey of road pricing. *Transportation Research*, **20A**, 87-98.

Musgrave, R.A. and Musgrave, P.B. (1989) *Public Finance in Theory and Practice*. Fifth Edition, McGraw Hill, New York.

Nash, C.A. (1976) *Public versus Private Transport*. Macmillan Studies in Economics, The Macmillan Press Limited, London.

Newbery, D.M.G. (1988a) Road user charges in Britain. *The Economic Journal Supplement* (Conference 1987), **98**, 161-176.

Newbery, D.M.G. (1988b) Charging for roads. *The World Bank Research Observer*, **3**, 119-138.

Newbery, D.M.G. (1989) Cost recovery from optimally designed roads. *Economica*, **56**, 165-185.

Newbery, D.M.G. (1990) Pricing and congestion: economic principles relevant to pricing roads. *Oxford Review of Economic Policy*, **6**, 22-38.

Newbery, D.M.G., Hughes, G.A., Paterson, W.D.O. and Bennathan, E. (1988) Road Transport Taxation in Developing Countries: The Design of User Charges and Taxes for Tunisia. World Bank Discussion Paper No. 26, World Bank, Washington, D.C., pp. 1-94.

Oum, T.H. and Trethaway M.W. (1988) Ramsey pricing in the presence of externality costs. *Journal of Transport Economics and Policy*, **22**, 307-318.

Pigou, A.C. (1920) *The Economics of Welfare*. First edition, Macmillan Company, London.

Pozdena, R.J., Schmidt, R. and Martin, D. (1990) Market-Based Solutions to the Transportation Crisis: The Concept. A Two-Part Report, *Bay Area Economic Forum*, May, pp. 1-17.

Prest, A.R. (1969) *Transport Economics in Developing Countries: Pricing and Financing Aspects*. London, The Trinity Press, Weidenfeld and Nicholson. Excerpt reprinted as "Public Transport Pricing and Cost Recovery," in Pricing Policy for Development Management, ed., Gerald M. Meier (1983), pp. 216-222, EDI Series in Economic Development, published for The World Bank, Johns Hopkins University Press, Baltimore, Maryland.

Ramsey, F.P. (1927) A contribution to the theory of taxation. *Economic Journal*, **37**, 47-61.

Roth, G.J. (1987) *The Private Provision of Public Services in Developing Countries*. EDI Series in Economic Development, published for The World Bank, Oxford University Press, Washington, D.C.

Schiff, M. (1991) New findings in the theory of optimal congestion taxes, with an application to road transportation. *Revista de Análisis Económico*, **6**, 81-92.

Small, K.A., Winston, C.M. and Evans, C.A. (1989) *Road Work: A New Highway Pricing and Investment Policy*. The Brookings Institution, Washington, D.C.

Talvitie, A.P., and Associates (1978) Policy Analysis of a Transportation Corridor. ITS Research Report UCB-ITS-RR-78-5, Urban Travel Demand Forecasting Project, Final Report Series, Vol. 10, Institute of Transportation Studies, University of California, Berkeley, California, pp. 1-306.

Tanner, J.C. (1963) Pricing the Use of Roads -- a Mathematical and Numerical Study. Road Research Laboratory Note No. L/N 319, Department of Scientific and Industrial Research, Hammondsworth, England, pp. 317-345.

Thomson, J.M. (1970) Some aspects of evaluating road improvements in congested areas. *Econometrica*, **38**, 298-310.

Thomson, J.M. (1974) *Modern Transport Economics*. Penguin Modern Economics, Bungay, Suffolk, Richard Clay (The Chaucer Press) Limited.

Transportation Research Board (1985), Highway Capacity Manual, Special Report 209, Transportation Research Board, National Research Council, Washington, D.C., pp. 1-1 to 14-4.

Transpotech (1983) Electronic Road Pricing in Hong Kong, Pilot Stage: Economic Background and An Example. Technical Paper 1, October, background paper prepared for the Hong Kong Government, pp. 1-14.

Transpotech (1985) Electronic Road Pricing Pilot Scheme: Main Report. Main Report prepared for the Hong Kong Government, pp. 1.1-4.40.

Vickrey, W.S. (1968) Congestion charges and welfare: some answers to Sharp's doubts. *Journal of Transport Economics and Policy*, **2**, 107-118.

Vickrey, W.S. (1985) The fallacy of using long run cost for peak load pricing. *The Quarterly Journal of Economics*, **95**, 1331-1334.

Walters, A.A. (1954) Track costs and motor taxation. *Journal of Industrial Economics*, **2**, 135-146.

Walters, A.A. (1961a) The theory and measurement of private and social cost of highway congestion. *Econometrica*, **29**, 676-699. Reprinted in Readings in Urban Economics, eds., Matthew Edel and Jerome Rothenberg, (1973), Macmillan, New York, pp. 417-437.

Walters, A.A. (1961b) Empirical evidence on optimum motor taxes for the United Kingdom. *Applied Statistics*, **10**, 157-169.

Walters, A.A. (1968) The Economics of Road User Charges. World Bank Occasional Paper Number 5, International Bank for Reconstruction and Development, Johns Hopkins University Press, Baltimore, Maryland.

Walters, A.A. (1987) Congestion. In The New Palgrave: A Dictionary of Economics, Vol. 1, John Eatwell, Murray Milgate and Peter Newman (eds.), The Macmillan Press Limited, London, pp. 570-573.

Watson, P.L. and Holland, E.P. (1976) Congestion pricing: the example of Singapore. *Finance and Development*, **13**, 20-23.

Watson, P.L. and Holland, E.P. (1978) Relieving Traffic Congestion: The Singapore Area License Scheme. World Bank Staff Working Paper No. 281, June, The World Bank, Washington, D.C., pp. 1-286.

Winston, C.M. (1985) Conceptual developments in the economics of transportation: an interpretive survey. *Journal of Economic Literature*, **23**, 54-87.

Wohl, M. and Hendrickson C. (1984) Transportation Investment and Pricing Principles: An Introduction for Engineers, Planners and Economists, Wiley Construction and Engineering Series, Wiley-Interscience, New York.

World Bank Operational Manual Statement (1977) Cost Recovery Policies for Public Sector Projects: General Aspects. No. 2.25, March, The World Bank, Washington, D.C., pp. 1-7.

World Bank (1986) Urban Transport: A World Bank Policy Study. Water Supply and Urban Development Department, The International Bank for Reconstruction and Development, Washington, D.C., pp. 1-61.

World Bank (1991) Urban Policy and Economic Development: An Agenda for the 1990's. Urban Division, Infrastructure and Urban Development Department, The International Bank for Reconstruction and Development, Washington, D.C., April, pp. 1-87.

Zettel, R.M. and Carll, R.R. (1964) The Basic Theory of Efficiency Tolls: The Tolled, The Tolled-Off and The Un-Tolled. Highway Research Record, No. 47, paper presented at the 43rd Annual Meeting of the Highway Research Board, January 13-17, 1964, National Research Council, Washington, D.C., pp. 46-65.

APPENDIX. MEASUREMENT OF THE WELFARE IMPACT OF ROAD PRICING

A brief analysis of the measurement of the welfare impact of road pricing would help explain why road pricing is unpopular. We will consider several acceptable approaches to measuring the net benefits offered by the introduction of road pricing on a non-perturbed equilibrium. Each casts different light and insights on the controversy surrounding road pricing.

A.1 Quantity approach

The first approach, which is more popular in the U.S. literature, is to measure the so-called welfare gain or loss areas (see Kraus et al. (1976), for example). This standard method is labeled the quantity approach or the 'American' approach (see Figure 4). The loss in valuation to the consumer-traveler from a reduction in trips from Q^o to Q' as a result of increasing the generalized travel cost to him from P^o to P' is the vertical, trapezoidal area $d+g+k$. The saving in resource cost to travelers from the reduction in

traffic, together with the saving of congestion in the form of externality reduced, is the vertical area *l+d+g+k*. The net benefit to society of the introduction of road pricing is given by the triangular area *l*.

Net benefit approach. A variant of this approach is the net benefit approach (see Figure 4). The net benefit in the case of the optimal traffic level of Q´ is typically a large triangle (the pie area *a+b+e+h* between the demand function and the marginal cost curve), with the pie triangle emanating from the point of optimum. Similarly, the net benefit in the case of the non-optimal level of Q^o is given by the difference of the pie area *a+b+e+h* and the small triangular area *l*. The latter area is of course the welfare cost saved when the traffic level is induced to be lowered from Q^o to Q´. This variant is intuitively appealing as it graphically illustrates that net benefit is maximized with marginal cost pricing. Any departure from the point Q´, either in a positive or negative direction, would slice into this maximal net benefit pie. To the left [or right] of Q´, travelers' marginal valuation would exceed [or be less than] the marginal cost.

A.2 Change in total benefits and total costs approach

The above procedure, and its variant, is an impeccable one. However, there is an alternative intuitive method to calculating the net benefit of introducing road pricing. This approach is widely used in the British literature (Ministry of Transport, 1964; Tanner, 1963, pp. 318; Gwilliam and Mackie, 1975, pp. 105-106; Thomson, 1970; Thomson, 1974, pp. 142-145). The findings of the Ministry of Transport, known as the Smeed Report, present a different calculation of the areas of gains and losses indicated above, yielding different insights into the problem (see their Appendix 3). The 'British' approach uses the change in total benefits and change in total costs. The change in total benefits accruing to those who are tolled off the road are negative because they suffer a loss in valuation equivalent to the vertical area *d+g+k*. The change in total costs -- expressed as the reduction in the total expenditure on travel in the form of savings in time cost -- accrues to all motorists and is given by AVC^oQ^o - AVC″Q′ (or P^oQ^o - P″Q′), that is, the area *e+f+g+k*. The net gain to society is the area *e+f-d.* Heuristically, the remaining users find that they derive satisfaction from the savings in time cost of the area *e+f*. The losers -- those tolled off the road -- would clearly experience a welfare loss of the area *d*.

The discussion thus far gives the conclusion and mistaken impression that those who remain behind are in fact better off by the *entire* savings in time cost of area *e+f*. In fact, drivers who remain on the road have to make toll payments of the area *b+c+e+f*, which in turn become a gain to the government in the form of toll revenues. (This is the notion of a transfer payment excluded in cost-benefit calculations using the British approach, see Gwilliam and Mackie (1975, pp. 105-106).) Yet, paradoxically, it is precisely the imposition of this tax -- resulting in a transfer payment -- which enables those who remain on the road to benefit the time savings of an additional area *e+f*.[43] Without the

[43] The rectangular area *b+c+e+f* should be counted as accruing *either* to the government in terms of toll revenues *or* returned to consumers (via a hypothetical lump sum transfer mechanism). A lesson to be learned regarding the issue of transfer payment is to avoid double-counting. *If* a dollar is treated as a dollar to whomsoever it accrues *and* the transfer mechanism is implemented, *then* the move to road pricing results in positive net benefit to society of area *e+f–d*. Putting it another way, the standard notion of a transfer payment of the area *b+c+e+f* says that money goes from the consumer's pocket into the government's. It is important to view the time savings of area *e+f* as an additional layer on top of the transfer payment itself. The bottom layer

tax, motorists are not properly induced to save valuable time resources: the time is completely lost. The ones who remain on the road, however, actually suffer a loss of consumer's surplus of the rectangular area *b+c*. It is as if a discriminating monopolist -- in the guise of the government's tax department -- carves away part of the users' consumer surplus. Also, to 'benefit' from time savings of the area *e+f*, drivers are in fact trading a dollar of money for a dollar's worth of time, implying that both a standard and constant value of time and efficiency analysis are assumed. (The rest of the transfer payment of area *e+f* also accrues to the government in the form of tax revenues.)

Prima facie, whether or not the net benefits of introducing road pricing using this latter approach (i.e., area *e+f-d*) and the former approach (i.e., area *l*) are equal is not at all obvious. The latter procedure gives less indication of the notion of optimality when compared to the first approach, especially with regard to its variant. In the quantity approach, one could move to the left or right of Q´ and observe that the net benefit pie to society would clearly be eroded, suggesting that Q´ yields maximal net benefit. Using the latter approach, however, as Q increases past Q´, a welfare loss area would increase. This would have to be offset with a new rectangular area of saving in resource cost. The point is that it is unclear whether Q´ can be shown to be optimal, at least diagrammatically, because the new rectangular area *may* not offset the new (trapezoidal) welfare loss area. Formally, the proof is as follows: the move from Q´ to Q^o yields a change in cost to society of area *l+d+g+k* because the vertical area below the marginal cost curve is a proper measure of cost. Equivalently, the change in variable cost of going from Q´ to Q^o is the inverted L-shaped area *e+f+g+k*. By definition, these two areas must be equal, implying that *e+f=l+d* or *l=e+f–d*. Diagrammatically, it may *appear* as if the change in total benefits and total costs approach yields larger net benefit area-use. However, it needs to be clearly shown here.[44]

A.3 Consumer's surplus and producer's surplus approach

The third approach using the summation of changes in consumer's surplus and producer's surplus involves the term quasi-rent.[45] The traveler is both a consumer (in the sense that he derives benefits from purchasing a transport service) and a producer (in the sense that he himself has to purchase the inputs with both his own time and operating costs). In the absence of road pricing, because drivers travel up to the point where average variable cost intersects the demand (at output level Q^o), the entire receipt (from the consumer traveler's expenditure) goes to cover the 'payment' of user-supplied factor inputs, so zero quasi-rent is thereby generated. However, it could be equivalently stated that the nil area can be expressed as the difference of two triangles, i.e., area (*e+h*) –

---

goes from the motorist's pocket to the road agency's; the top layer is obtained because the motorists who are tolled are forced to trade money with time. Remarkably, it is this coerced payment of the tax revenue area *e+f* which brings about a real saving of time of an area of equal size.

[44] Lee (1982) claims that the two areas based on the different methodologies are the same but does not prove it. The spirit of the analysis I show here underlies Gwilliam and Nash's (1972) comment on Beesley and Walters' (1970) evaluation of urban road investments.

[45] The notion of rent is a slippery one and warrants clarification. Rent is the analog of producer's surplus in the input market. Rent is a *permanent* payment to a factor over and above that which is required to draw forth its resources. Quasi-rent is a *temporary* payment and would continue only until the capital asset is depreciated or possibly transferred to another use (see footnote 31 and Mohring (1976, Chapter 2) also). Note that a high price and willingness-to-pay yields high quasi-rent, and not the reverse.

($c+d+l$), by simply exploiting the meaning and geometric relationship of average and marginal cost curves. In the advent of road pricing, the quasi-rent -- the return to a fixed factor of production -- is essentially the amount which the traveler-as-producer 'receives' over and above his total variable costs. This quasi-rent, instead of accruing to the drivers as such, is captured by the government in the form of toll revenue or a user charge, and hence should be accounted for properly in benefit-cost calculus. Note that the quasi-rent of area $b+c+e+f$ can be re-expressed as the area $b+e+h$. Clearly, the change in the quasi-rent would be equal to the area $b+c+d+l$. Coupled with the loss in consumer's surplus of area $b+c+d$, the net benefit area l emerges. Hence, we have shown in different ways that the three approaches are identical.[46] Note, also, however, the second approach is used and extended because it graphically illustrates the broad distributional implications of road pricing.

Perhaps one reason why there is confusion regarding the two approaches above is because of Walters' (1961a, b) treatment of the MC and AVC curves as marginal social cost and *marginal* private cost curves respectively. (My interpretation here is at variance with the common use of the latter term since I regard it as somewhat of a misnomer.) Walters' use of the AVC as MC curve immediately brings to mind standard diagrams of an externality such as the classic economics text example of a polluting factory, with the consequent changes in consumer's surplus, producer's surplus and externality valuation. As carefully shown above, this example is not valid in our analysis because the marginal private cost curve is only marginal with respect to the driver himself. The individual perceives and bears the average variable cost only: it is merely a decision curve and no more. Since the area below the AVC curve is not the total private cost, only an incorrect interpretation can be drawn by mathematically integrating the area under the marginal private cost curve, which turns out to be an *average* variable cost curve. (Further, producer's surplus should really be interpreted as quasi-rent to avoid possible confusion, especially in understanding the relationship between pricing and investment.) In the absence of the optimal pricing of trips, average (variable) cost pricing prevails with the associated inefficiencies. This illustrates strongly the need to reserve the term 'short-run marginal cost pricing' to be consistent with the World Bank's policy guideline (World Bank Operational Manual Statement, 1977). I employ the term marginal *social* cost pricing when accounting for all the other externalities such as environmental pollution and accident costs.[47]

A technical paper by consultants hired by the Hong Kong Government indicates that the net benefit due to introducing road pricing corresponds to the area $e+f+d+g+k$ (Transpotech, 1983, Figure 4). The consultants' explanation of the extra area $g+k$ is either based on a possible misunderstanding of the first two approaches, or simply a matter of double-counting areas $g+k$. The authors state that "this money [referring to the areas $g+k$] is available to be spent in other ways, perhaps on other modes of traveling". Having already included the resource saving as a reduction in the expenditure on travel, $P^oQ^o - P''Q'$, the resource saving from the tolled off drivers of the vertical area $g+k$ should *not* be counted twice. The point here is that unless care is taken to ensure rigorous

[46] In fact, if Q^o is very close to Q', the marginal cost is equivalent to the change in total variable cost. By evaluating the difference of the change in variable cost with the marginal valuation, the two methods discussed above (the quantity approach and the *change* in total benefits and total costs approach) are seen to be equal.

[47] For example, Glaister's (1981, Chapter 5) use of the term 'marginal social cost pricing' is synonymous with the marginal cost pricing concept employed here.

cost-benefit analysis, the benefit (or cost) figures would be biased, as has been the case with the evaluation of the electronic road pricing experiment in Hong Kong.[48]

[48] Based on the numbers presented for an illustrative case, the bias is 40% upwards. It should be stressed, however, that it is unclear from a reading of the Hong Kong Government's Main Report on the Electronic Roading Pricing Pilot Scheme whether the final report followed the methodology outlined in Technical Paper 1 (Transpotech, 1983, 1985).