3. Congestion pricing and road investment

Timothy D. Hau*

I will begin with the proposition that in no other major area are pricing practices so irrational, so out of date, and so conducive to waste as in urban transportation. Two aspects are particularly deficient: the absence of adequate peak-off differentials and the gross underpricing of some modes relative to others. In nearly all other operations characterized by peak load problems, at least some attempt is made to differentiate between the rates charged for peak and for off-peak service. Where competition exists, this pattern is enforced by competition: resort hotels have off-season rates; theaters charge more on weekends and less for matinees. Telephone calls are cheaper at night ... But in transportation, such differentiation as exists is usually perverse.

William Vickrey (1963, p. 452)
Nobel Laureate in Economic Sciences, 1996

3.1 INTRODUCTION

The cost of congestion was regarded by the late Professor William Vickrey to be high. Very roughly, the real economic cost of the transport infrastructure in the US was about three times the total gasoline and vehicular taxes generated by automobile use of city streets (Vickrey, 1963). Yet motorists were always under the misimpression that they pay their way: highway taxes and license revenues were sufficient to cover the highway expenditures made by the US federal and state governments. Even if it were true that gasoline taxes or license fees were increased to meet higher road expenditures, so that motorists were to pay all their way, Professor Vickrey argued back in 1959 – with premonition in hindsight – that the results of not charging motorists for their rush-hour usage can be ‘disastrously expensive’ (Vickrey, 1960, p. 468; excerpt reprinted in Arnott, et al., 1994, p. 44). Why is it so important to charge users for their use of an item at the margin rather than on average? Professor Vickrey gives us a pedagogical illustration: Each member in a group of conferees meeting for dinner inevitably ends up paying for an expensive steak dinner – instead of most members economizing on the goulash – if, in
order to reduce the bookkeeping, the bill is divided evenly amongst the participants. The reason for the excessive consumption is that each person could not lower the group bill significantly by exercising self-restraint unless (s)he is charged according to the true resource costs. Thus everyone in society in fact ends up paying for a costly road system since motorists are not charged at the margin for road use.¹

Over three decades have passed by since Professor Vickrey enunciated his assessment and recommendations on the urban transportation problem to both the American and British governments. Nevertheless, congestion is ubiquitous as ever in major urban areas and incessant during peak periods and often between the hours. The traditional methods of curtailing congestion remain few, and their usefulness limited. On the supply side, the expansion and improvement of roads is restricted by increasingly tight fiscal, physical and environmental constraints. On the demand side, however, the problem can be addressed by pricing or regulation.² This paper argues that the role of peak/off-peak pricing is indispensable in tackling congestion because of its inherent flexibility and power of discrimination. My focus here is on internalizing only the external congestion effects caused by motorists. Congestion is recognized as a significant type of externality from vehicle usage in both developed and developing countries in that it represents a significant share of total estimated road use costs (Newbery, 1988, 1989, 1990; Newbery et al., 1988; Small, Winston and Evans, 1989, Chapter 6).³

This paper interprets the literature on the theory of optimal pricing and investment for roads based principally on the work of Herbert Mohring, Robert Strotz, William Vickrey, Alan Walters, Theodore Keeler and Kenneth Small. It aims to integrate their ideas and principles into a single analytical framework (see Newbery, 1989; Winston, 1991; and Hau, 1992a, for details of combining both congestion and road damage). The rigorous and unified non-mathematical framework derived from first 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 and cost recovery in the provision of road services. The static models relied upon here differ from the bottleneck model pioneered by William Vickrey (1973), which in turn has been extended by Arnott, de Palma and Lindsey's (1988) dynamic models.

Recent technological breakthroughs in automatic road use charging in a multi-lane setting have brought electronic road pricing much closer to reality. While there are a number of electronic toll collection systems in use in parts of Norway, Italy, Spain, France, America, Britain, Hong Kong, Korea and Japan, to name but a few places, only a few congestion pricing systems are currently operating (Hau, 1992b; 1995). The first is the well-known Singaporean Area Licensing Scheme, which has recently been converted into
Congestion pricing and road investment

the smart-card based Electronic Road Pricing Scheme in April 1998. The
electronic charging of vehicles entering Trondheim, Norway, during daylight
hours on weekdays since 1991 (and perhaps the manual charging of vehicles
entering the central business district of Bergen during extended daylight
hours on weekdays since 1986) could be viewed as a crude form of conges-
tion pricing. For traffic returning to Paris every Sunday afternoon since 1992,
the tolled A-1 expressway imposes a surcharge (of 25 per cent) during peak
hours and a discount of equivalent magnitude during the periods before and
after the peak. The French scheme has been successful in spreading conges-
tion during the peak hours onto the shoulder periods. With tremendous strides
made in technological advances, the electronic charging of congestion tolls is
both technically feasible and economically viable (Hau, 1990; Ramjerdi,
1995). This paper therefore attempts to come up with an integrated policy
package derived from economic principles to hasten the arrival of road pric-
ing as the urban transport solution.

3.2 CONCEPTUAL GUIDELINES

Rising real incomes result in increased aspirations for the ownership of private
automobiles (Hau, 1996). Barring major restraint measures, an increasing number
of motor vehicles means that travel demand swells. As municipalities find it
increasingly difficult to finance new road construction and improvements, the
rate of growth of travel demand outstrips the growth of road capacity. The
resulting traffic explosion is an illustration of Downs’ (1962, 1992) law of
peak-hour expressway congestion, in which commuter traffic ascends rapidly
to the level of new capacity in urban areas. 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 im-
proved. The only viable solution to this problem is road pricing.

In this section, I set up the conceptual guidelines which allow authorities to
improve transport efficiency by curtailing traffic congestion in an efficacious
manner while satisfying the World Bank’s general guidelines for public sec-
tor projects and urban transport policy (World Bank’s Operational Manual

In a nutshell, the principles include:

1. implementing short-run marginal cost pricing to generate maximum net
benefits for society: efficiency pricing;
2. undertaking investment in infrastructure whenever the additional benefits
exceed the true resource costs (long-run efficiency) of doing so: eco-
nomic viability;
3. investing in transport services 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.

3.3 FOUNDATIONS OF ROAD CONGESTION: THE CLASSICAL CASE IN THE SHORT RUN

In the first edition of *The Economics of Welfare* (1920, p. 194), Professor Arthur C. Pigou introduced the idea of a congestion toll by coming up with the famous two-road example. He postulates that one road is wide but rough and slow and effectively of unlimited capacity whereas the other is narrow but smooth and fast, and therefore of limited capacity. He argues that (commercial) traffic will distribute itself amongst the two alternative routes until the travel time is the same connecting two points. By imposing a differential tax on the traffic using the narrow road, excessive traffic from the narrow road would be diverted onto the wide road. Total travel time would be lowered and society’s welfare enhanced by the imposition of such a Pigouvian toll-tax.\(^4\) Thus far there is no dispute on Pigou’s contribution. However, Frank Knight (1924) challenged the idea of imposing a Pigouvian tax (on the difference of the marginal and average costs of a trip) to internalize a negative externality by pointing out that the externality arose because Pigou had implicitly assumed a public road. If (private) property rights on roads are delineated and competitive pressure is present, then self-interested road owners would charge users the same differential toll, obviating the need for government intervention in the form of taxation. Even though the two conditions are far from being fulfilled in reality, Pigou nevertheless withdrew the two-road illustration from subsequent editions of *The Economics of Welfare*. Be that as it may, Pigou’s two-road example remains a classic one, not least because almost all roads are publicly owned and subject to scale economies. Public ownership of roads results in ‘market failure’ (or perhaps ‘government failure’, as some would prefer to call it). Where there are private roads, a road without alternatives close by will likely exploit its locational monopoly characteristic, threatening a diminution of society’s welfare. Are we doomed to suffer from the ‘command economy’ characteristics of congestion and chronic shortage of funds in roads?

We begin our analysis by following the conventional treatment of the congestion problem in the transport economics literature. Consider a repre-
sentative driver cruising under low traffic conditions along a given stretch of urban road with fixed beginning and end points (Walters, 1961, 1987; Button, 1986). *Ceteris paribus*, as other vehicles enter the road thereafter, density increases, speed drops and travel time (or delay) lengthens. The causality is as follows: low traffic density yields high speed and not vice versa. Paralleling the theory of fluid dynamics, traffic flow is the product of density, in vehicles per kilometer, and speed, in kilometers per hour. Note that the rectangular area in Figure 3.1(a) is equivalent to traffic flow, expressed in vehicles per hour (see May, 1990, for example). Hence, traffic flow is endogenously determined by traffic density and speed, with traffic flow attaining a maximum at \( F^{\text{max}} \) with speed at \( S^m \) in Figure 3.1(b) (see Haight’s 1963 fundamental diagram of road traffic – a flow-density curve – and similar figures in Morrison, 1986). (Touted figures of the ‘capacity’ of a typical expressway are about 1800 vehicles per lane-hour at 55 kilometers per hour, (Gerlough and Huber, 1975, Chapter 4).

Given a fixed distance of a kilometer of road, say, 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 3.1c). Using a

![Figure 3.1 Derivation of a travel time-flow curve of an urban highway](image)

---

"Congestion pricing and road investment"
Figure 3.1 continued
Optimal User Charge = Costs imposed on other motorists + road authority
= External congestion cost + variable road maintenance cost
= optimal toll + variable road maintenance cost user charge component

Figure 3.2 Derivation of the marginal cost curve and congestion toll
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 3.2). 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 independent of the
level of traffic flow (Mohring, 1976, chapter 3). 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. Simi-
larly, the road’s variable maintenance cost, which is assumed to be propor-
tional to the traffic level, following Walters (1968, p. 24), can be added up
also.\(^5\) 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, \(Q^{\text{max}}\); it is variable in the sense that as traffic flow, \(Q\), is increased,
congestion delay actually sets in rapidly at a traffic level substantially below
the level \(Q^{\text{max}}\) – contrary to the engineering notion of a constant average
variable cost curve extending up to the point \(Q^{\text{max}}\). After the engineering or
basic capacity, \(Q^{\text{max}}\), is reached, AVC becomes an ‘inverse supply’ curve.\(^6\)
Note that the standard supply curve is non-existent in the context of roads.

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, or
price rather, facing a traveller for a single trip.\(^7\) When an initial demand
function, \(Q^d\), intersects the AVC curve at point \(U\) (Figure 3.2), a (stable)
equilibrium is said to exist at \(Q^d\).\(^8\) This is an equilibrium point because
travelers’ willingness-to-pay curve, that is, the inverse demand function, equals
the average variable cost curve – the function upon which travelers base their
travel decisions.

Basic price theory says that whenever the average variable cost rises, it
means the marginal cost curve lies above it.\(^9\) The vertical difference of the
two cost curves is the marginal (external) congestion cost – the additional
delay that one driver imposes on the rest. It is not taken into account by the
last driver unless he is unusually altruistic. 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 im-
posed on fellow motorists. We thus have the optimal point \(Y\) at which the
marginal cost curve intersects the (peak) demand curve in Figure 3.2. In other
words, \(Q^*\) is the associated optimal output in the sense that the generalized
cost, which includes time cost, external congestion cost and other variable
costs (that is, constant unit operating cost of a vehicle and variable road maintenance cost), is equated to the price. The (external) 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 trip and multiplied by the number of vehicles in the traffic stream (see note 9). The Pigouvian tax 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 disincentives. (This Pigouvian toll-tax is also known as the net-benefit maximizing, economically efficient, (Pareto) optimal and marginal cost toll.) 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, shown by distance YT in Figure 3.2, and another component which covers the variable road maintenance cost (see Figure 3.2’s legend).

Observe that at the equilibrium point U in Figure 3.2, the resultant throughput is significantly less than the road’s maximal flow capacity of point V. The backward-bending ‘supply curve’ exists when very dense traffic is reached there; a one percent increase in density results in more than a one percent decrease in speed. (That point X (or point W) is a stable (or unstable) equilibrium point can be seen intuitively by perturbing the price level.) We note that many cities are faced with extremely congested situations such as point X, certainly during the peak, if not for most of the day.10

3.4 THE WELFARE IMPACT OF CONGESTION PRICING

Here I highlight what I consider to be an important obstacle to introducing congestion pricing. Economists have long known that road pricing results in an improvement in welfare to society, yet politicians and the public have always regarded it with skepticism. Why? To economists, the increase in welfare comes about because of the imposition of an externality-corrective tax. Yet, for those motorists remaining on the road, the effect of the toll is similar to a tax increase and the payment made (distance YT) exceeds the monetary value of the time saving (distance ST), so that the ‘tolled’ are worse off as a group (by the distance YS) under the assumptions of a uniform value of time across the population and a normal downward sloping demand function (see Figure 3.2). When the toll is multiplied by the total number of vehicle-trips undertaken (Q'), the aggregate toll payment clearly exceeds the valuation of the time savings. When the value of travel time varies across the population, the downward sloping nature of demand ensures that the weighted average of the valuation of travel time savings is still less than the toll payment (see Hau,1992a, Appendix). Those who are tolled off the road to an
in inferior mode or time of travel are definitely worse off, while those who remain on other times of travel or modes — the 'toll on' — are either worse off, if congestion arises there, or just as well off, if there is no resulting increase in congestion.\textsuperscript{11} It turns out that the government, in collecting toll revenues, becomes the main party that is better off (by the area $Q'$ times the distance $YT$). (The other parties that are clearly made better off are the ones with high values of time.) So unless the congestion toll revenues are recycled and the tolled and the tolled off compensated, it is unlikely that road pricing would become a reality on a widespread basis.\textsuperscript{12}

We then ask whether there is a theoretical argument for dedicated funds or earmarking so that society as a whole would benefit from having road pricing implemented. Restated another way, is there a way in which road users can act as beneficiaries and are indirectly 'compensated' for their toll payment by satisfying some commonly accepted notion of fairness, while not violating the first-best pricing rule? I think that the answer is 'yes', although not entirely without qualifications.

3.5 SHORT-RUN EQUILIBRIUM IN TRANSPORT

In a textbook industry such as widgets, the producer uses the revenue which he obtains from selling the good at the market price to pay for all 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. However, transport is unusual in that the traveller is both a producer and a consumer simultaneously. The graphs for the textbook commodity are similar to (but not always the same as) the case of roads considered in Figure 3.2.\textsuperscript{13}

The trip-maker himself supplies some variable inputs, which include vehicle operating cost and time cost, but not the fixed capital infrastructure. Since the competitive level of trips exceeds the efficient quantity in the presence of congestion, and because the quasi-rent of a highway facility would be dissipated due to free competition with public roads, the imposition of an optimal toll would recapture this quasi-rent. Note that because price equals the entire marginal cost of a trip, the optimal toll is equal to the difference of \textit{SRMC} and \textit{SRAVC}. This is a subtle but crucial distinction between transport and widgets. Despite this difference, short-run equilibrium in transport occurs when the government, in the form of a highway agency, behaves in an optimizing manner just as a private competitive firm would were it possible to organize the industry in a competitive fashion.

To see how this might be done, we introduce the fixed cost, that is, the cost of construction, together with the invariable maintenance, depreciation and operating costs of a road that faces a road agency in Figure 3.3. We then
Figure 3.3 *Introducing the (short-run average) fixed cost, SRAFC, of a road, short-run optimal toll with economic profit*

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, SRAFC', and the average variable cost curve yields the average total cost curve. With marginal cost pricing, the short-run equilibrium output and price is given by the efficient output, $Q'$, and price, $P'$, respectively.
3.6 LONG-RUN EQUILIBRIUM UNDER CONSTANT RETURNS

The motorist is oblivious to the capital cost of a road, and his behavior is independent of it. However, from the highway agency's planning point of view, even though the size and capital cost of a road is very much taken into account, once a highway is built it is regarded as sunk. The sunk cost of a road 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 unaffected.

In the long run, a highway agency can vary the fixed capital input by road expansion, if the investment is deemed justifiable. Expanding a road until the additional benefit equals the additional cost of building it would yield maximal net benefit to the community. We note that charging the optimal toll of the distance \( t' \) in Figure 3.3 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 \( \pi' \). In general, there is no \textit{a priori} reason why toll revenue collections 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 (following Mohring and Harwitz, 1962, chapter 2). Ultimately, the quasi-rent earned by the existing capital equipment would then be equal to its fixed (opportunity) cost of capital. Upon seeing the existence of economic profits, other firms enter the industry and expand the supply, increasing output and lowering price as a result. The unrestricted mobility of resources and the entry and exit of firms serve as the forces by which profits would be competed away in due course. When capital is freely varying and zero economic profit occurs, long-run equilibrium is reached (see Figure 3.4). This zero profit condition holds under constant returns to scale, where a proportionate increase in all inputs results in the same increase in output. Given fixed factor prices, both average total cost and marginal cost remain constant and flat 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 costs of reproducing the invested capital plus the highway facility's invariable maintenance and depreciation costs, new investment is expected to find its way into that road segment of the highway industry if the appropriate (marginal cost) price signals are given. Equivalently, in the long run, if toll revenues -- which recover quasi-rents over time periods -- exceed the fixed cost of the existing facility, the highway agency 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 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 the collection of congestion toll revenues. In this way, full costs are covered and the problems of cost recovery and cost allocation disappear. There is no need to raise charges over and above marginal cost – unless one wants to internalize other externalities or to impose a pure tax element.14

3.7 OPTIMAL INVESTMENT

Professor Herbert Mohring was the first to show the powerful result that congestion toll revenues would exactly cover the amortized cost of construction, invariate maintenance and depreciation costs of roads in the long run under the technical conditions of constant returns to scale in road construction, maintenance and road use (Mohring and Harwitz, 1962; Mohring, 1965; Arnott and Kraus, 1998). Constant returns to scale in construction and maintenance intuitively means 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.15 By steadily pursuing the policy of marginal cost pricing of a trip via congestion pricing 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 (Figure 3.4). Clearly, collecting a unit congestion toll would cover the entire average fixed cost of the road and yield zero profit only because the existence of economic profit or loss acts as a quasi-market mechanism in the investment decision of whether to expand or contract highway capacity.
3.8 LONG-RUN VERSUS SHORT-RUN MARGINAL COST PRICING

Intuitively, the long-run marginal cost of producing a trip yields the cost of undertaking a trip to the society 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 3.4, long run marginal cost pricing would exactly cover all the variable costs, including time cost, vehicle operating cost and variable road maintenance cost, plus the fixed construction, 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 in the same diagram. After all, in the long run, the users' marginal willingness-to-pay, the short-run and long-run marginal and average costs are all equal. However, if a road is not optimally constructed but underbuilt, say, 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 the travellers' true willingness-to-pay and would also yield positive toll revenues and economic profits as a by-product. Short-run marginal cost pricing is the rule to use whether or not long-run equilibrium is reached. 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. We shall explore this point further in the section on demand variability.

Indeed, Vickrey has emphatically argued that there can be no solution to the urban transportation problem without peak-load pricing. Time-of-day pricing is an application of the concept of short-run marginal cost pricing. Pursuing economically efficient pricing period by period over the long run would not only guarantee the best use of society's given 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 – that is, congestion pricing – rather than long-run marginal cost pricing be used whenever cyclical variations in demand are involved.

3.9 TRADING OFF TRAVELLER'S TIME AGAINST THE GOVERNMENT'S RESOURCES

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 travelling public's desire to save time? By minimizing the sum of these costs, 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 3.3, it can be seen that the least cost for the community involves having a road that is too small, given the demand as depicted. Therefore, there is an incentive for the community to expand the capacity of the road. Long-run equilibrium is reached when the minimum point of the short-run average total cost curve (which 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 from road capacity expansion rises. Intuitively, the highway agency continues to expand the road until the marginal benefit from saving users’ time costs is just offset by the marginal cost of an extra unit of capacity.\(^\text{17}\) It is at the output \(Q^*\) that the valuation of a trip just equals the additional cost to society of taking that trip, plus the vehicle operating cost and the road maintenance cost, given an optimally built road, \(K^*\) (see Figure 3.4).\(^\text{18}\) By pursuing an efficient pricing policy for each stretch of road, the use of an existing, non-optimal 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 common practice of not maintaining roads to preset engineering standards is tantamount to an act of disinvesting in roads.

3.10 FIRST-BEST OPTIMAL PRICING AND INVESTMENT\(^\text{19}\)

Given an estimate of a speed-flow curve and the corresponding travel-time flow curve, we know how the engineering curves can be converted to a short-run average variable cost curve of a trip, using an estimate of the value of time. Hau (1994) shows how speed-flow and travel time-flow curves are derived empirically for a representative Indonesian road type.\(^\text{20}\) Further, a ‘supply’ elasticity (or cost elasticity, rather) estimate yields a one-to-one correspondence between the short-run average variable cost and the short-run marginal cost (see footnote 9 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 agency.

*First Rule – Optimal Pricing Rule.* For each stretch of road, short-run marginal cost pricing is fulfilled by setting a toll equal to the difference of short-run marginal cost and short-run average variable cost. Intuitively, this
congestion toll would serve to internalize the (external) 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.

Second Rule – Optimal Capacity Rule. Under constant returns and optimal pricing, whenever economic profit is found in the operation of a road link, the capacity of that stretch of road should be expanded. The existence of a loss under short-run (as opposed to long-run) marginal cost pricing suggests that the road has been overbuilt. By altering the capacity of each road over 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 agency—by trading its direct resource costs against individuals’ travel time—can follow the rule of setting the marginal value of user cost savings equal to the marginal cost of investment for an additional unit of capacity. Equivalently, the capacity of a road is expanded until the marginal capital cost equals the marginal (external) congestion cost.

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 road investment and individual road users’ travel time cost. In our simple framework, congestion delay would never be entirely absent, contrary to what motorists and some environmentalists 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 at all 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 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—fully taking into account the social opportunity cost of capital—has been the conventional dictum.\textsuperscript{21}

By contrast, a road would sooner or later possess the rival consumption attribute of a private good when demand rises. As a result, a congested road can be regarded as a congested variable-use public facility (or a ‘club good’, see Buchanan, 1965). With such mixed goods characteristics, the task of providing road services ought to remain with the public sector unless full exclusion is achievable and imperfect competition is of minor concern. Under the condition of constant returns, we have shown that the optimal toll revenue, which also captures the quasi-rent earned from the invested capital and
reflects the scarcity value of the facility, would cover the 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), and the road is therefore not in long-run equilibrium. The existence of economic profit serves as a surrogate market signal to expand capacity. Updating his earlier work in 1988 on road user charges, Newbery (1990, 1994) estimates that the cost of expanding highway capacity in the United Kingdom is 2.2 pence/km at 1992/93 prices for a private car. This means that transport authorities could simply look at traffic forecasts and decide to expand highway capacity when (external) congestion costs reached 2.2 p/km, and not otherwise.

Similarly, if a road loses money, it suggests that planners may either have invested mistakenly or made overoptimistic forecasts of travel demand, for instance. 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 also 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 invariable maintenance costs that may arise. Efficient pricing, financial viability and cost recovery are therefore entirely consistent with one another under constant returns to scale in long-run equilibrium.

3.11 RELAXATION OF ASSUMPTIONS

Some of the more stringent assumptions employed thus far need to be relaxed. They include: (1) constant value of time, (2) static demand, (3) perfect divisibility, (4) constant returns to scale and (5) variability of road thickness. Here we explore relaxing only the first four assumptions (see Hau, 1992a).

3.11.1 Differences in Time Valuation

The traditional presentation of road pricing assumes a constant value of time (Walters, 1961). The diagrammatic analysis in Figure 3.2 implicitly assumes that every driver is identical and maintains the same time valuation. The question then is what happens when there are heterogeneous motorists, with different time valuations and tastes. A mathematical proof that generalizes the above result of homogeneous drivers to heterogeneous ones with different values of time is shown by Mohring (1976, chapter 4 Appendix) and Strotz (1964), but the intuition behind it is straightforward. Instead of the optimal toll being based on a representative driver’s (marginal) value of time, the optimal toll is now a weighted average of the different motorists’ marginal valuation of time, weighted by the number of trips taken by those motorists
who actually remain on the road. If a traveller’s time value and the number of trips are close to the average, he will incur the toll payment that everyone is faced with. If another motorist’s time value is higher (lower) than average, he would be willing to pay more (less) than the average toll payment for taking a trip. (Each of them would be willing to, though begrudgingly, pay the difference of the toll and his valuation of time saving.) He thus would end up paying the difference. The constant value of time is reinterpreted as the weighted (marginal) valuations of time, whereas the congestion toll, YT, in Figure 3.2 then can be labelled the weighted congestion toll. For a trip with a very high time value, the money equivalent of the time saved, ST, can be even higher than the weighted congestion toll, YT, thus making the motorist better off. On the other hand, for a trip having a lower than average time value, the user still has to make the average payment and therefore would be made considerably worse off. Nevertheless, they both remain on the tolled road, as opposed to being tolled off, because their individual trips’ marginal valuation (or maximum willingness-to-pay) still exceeds (or equals) the generalized cost of their respective journeys. The use of nonconstant values of time would relax the point that congestion pricing would make almost all groups besides the government worse off. By loosening the stringent assumption of a constant value of time across the population, those people with very high values of time would be made better off at the expense of those with low values of time. This intuitive analysis assumes that everyone is faced with the same toll, as in the workings of a competitive economy. Note also that the optimal toll incorporates motorists with both high and low valuations of time but excludes those who are tolled off. With differences in values of time under constant returns, efficient pricing and financial viability are still achievable.

3.11.2 Demand Variability and Peak-load Pricing

Applying short-run marginal cost pricing means that a congestion toll is needed but none when there is excess capacity during the off-peak. With a fixed highway capital stock, the systematic, diurnal nature of travel demand (as opposed to the static, invariant demand character assumed till now) means that the sum of quasi-rents (rather than just the quasi-rent from the single period itself) of the invested capital should be compared with the cost of the highway facility. In other words, when all the quasi-rents over the entire demand cycle are summed up and compared with the capital cost, we can then ask whether expansion of the highway is warranted or not, under constant returns. The conclusions obtained thus far again holds.

An interesting implication is that the entire capital cost of the highway is ‘allocated to’ and borne by peak travellers, mainly rush-hour commuters. This surprising result may seem ‘inequitable’, yet it is perfectly consistent
with efficiency analysis. After all, it is peak users themselves that create congestion and they that demand the use of heavily congested expressways requiring massive infrastructure developments. Without these peak commuters, the optimal size of the road would be much smaller. The result of allocating all capital costs to users of the peak period is long recognized in the literature on the pricing of public utilities as in Boiteux (1960). The optimal investment rule is then to expand a road until the sum of the quasirents over the demand cycle equals the entire capital cost of the facility. By implementing both peak-load pricing and altering the investment level of the highway facility, depending on whether profits are positive or negative, the highway network can again be optimized. Hence, the consideration of demand variability and peak-load pricing would not change the status of our conclusions, in the presence of differences in valuation of time. The fact that the fluctuating demands over the various peak, off-peak and inter-peak periods of a demand cycle are linked by a fixed capital facility and the observation that the consumption of trips must be satisfied by the production of trips during that particular time period combine to yield a simple modification of our result. Pricing, financial viability and cost recovery are again consistent with one another.

Keeler and Small (1977) show rigorously how the Mohring–Harwitz framework developed here is extended to the case of variable demands under peak-load pricing in the presence of independent demands and no indivisibilities. By assuming the demand in each period is in fact dependent on other periods, that is, the case of dependent demands, the derived results still go through (Mohring, 1970).

### 3.11.3 Indivisibilities

While still retaining the assumption of constant returns, but accounting for differences in values of time and demand variability, we proceed to drop the assumption of a road being finely divisible.

Road construction, in fact, involves significant indivisibilities that cannot be ignored. For example, a road must possess the minimum width for accommodating a standard-sized automobile and should also ideally be bidirectional. In the perfectly divisible case, the long-run average total cost curve enveloping a continuum of closely-packed short-run average total cost curves at their minimum points is made horizontal. A flat LRMC curve also coincides with the corresponding LRATC curve (see Figure 3.4). Due to the presence of indivisibilities, however, the formerly neat and continuous pattern of the LRMC curve is therefore broken (Neutze, 1966). The new long-run average total cost curve is now composed of a finite series of short-run average total cost curves. We note that whenever a short-run marginal cost
curve rises above a short-run average total cost curve, profits can be obtained under short-run marginal cost pricing. Thus, if demand happens to intersect the short-run marginal cost curves in the upward-sloping sections of their SRATC curves, then the road makes money in the long run under constant returns. *Per contra*, in the downward-sloping portions of the SRATC curves, the road loses money. When the SRATC is neither rising nor falling (as in Figure 3.4), the road breaks even. With a two-lane road, say, as traffic increases, the road’s large fixed cost is spread out by additional traffic, and as congestion sets in, the road begins to make money. When road expansion is justified by cost–benefit analysis, congestion relief results in the road losing money. In other words, as travel demand continues to grow along the trend, adherence to short-run marginal cost pricing suggests that the road would go through an unavoidable cyclical pattern of deficit, surplus, deficit, surplus, and so on. Whether or not one undertakes a road expansion project from two to four lanes depends on a computation of the net benefits, using welfare gain and loss measures, via cost–benefit analysis (see the example in Hau, 1992a, pp. 35–6).

The optimal sequence of decision-making is to first establish the policy of applying marginal cost pricing and then plan future adjustments of the road network according to expected future demand and established pricing policies (Vickrey, 1987). When demand fluctuates, pursuing short-run marginal cost pricing would mean setting different prices, or congestion tolls rather, in response to expected current conditions.

### 3.11.4 Returns to Scale

The issue of whether constant returns to scale exists or not in road transport is a controversial and important one. Ultimately, it can be answered satisfactorily only via careful econometric analysis of individual cases. The available evidence in road transportation indicates that all three cases exist: decreasing returns to scale, constant returns to scale and increasing returns to scale (see Figure 3.5) – paralleling the case of a competitive private firm and industry – with profit, break even and loss, respectively. (This is but a well-known result of economic theory applied with slight modification to the highway.) It is important to realize at the outset that the case of scale economies, or increasing returns to scale with fixed factor prices under least cost combinations, is merely a case of insufficient demand with respect to the market size in the long run – a point that is sometimes overlooked. This means that if traffic were to grow until congestion delay sets in, congestion toll revenues could be collected. (After all, the short-run marginal cost curve for road use is always non-decreasing, as we have shown in Figures 3.1c and 3.2.) Profits may still occur – due to indivisibilities – despite the fact that the long-run average cost
curve is declining. And if traffic were to continue to grow as real incomes and auto ownership rise, concomitant with expressway expansion, the decreasing returns region would then be encountered (see Figure 3.5). In the case of increasing returns with perfect divisibility – commonly known as the natural monopoly case – efficient pricing will result in losses, beckoning government subsidization. On the other hand, if travel demand is sufficiently high relative to engineering capacities of roads, the money-making enterprises would provide much sought-after funds which could be used to finance efficiently priced but money-losing roads – only if these roads yield positive net benefits to society.
A. Economies of scale
There is evidence of significant economies of scale in the construction of rural roads (Walters, 1968, pp. 180–82; Mohring, 1976, pp. 140–42). In particular, a two-lane road requires a minimum of a twelve-feet width for each lane and a few feet for shoulders and drainage ditches. What this means is that a substantial proportion of the provision of a road’s right-of-way involves dead space. These indivisibilities help contribute to declining unit cost as the large fixed cost of construction and invariate maintenance and depreciation costs are shared over greater amounts of traffic. Thus doubling the width of a two-lane road more than doubles its capacity, the so-called ‘shoulder effect’ (Hayutin, 1984, pp. 106 and 154). Further, we know that the engineering or basic capacity of a two-lane road is about 2000 vehicles per hour. Since the standard four-lane road has an engineering capacity of 1800–2000 vehicles per lane per hour, doubling the width of a two-lane road almost quadruples its capacity. Further, in order to level hilly terrain and/or fill valleys for transportation purposes, the earth moving costs rise less than proportionately. Hence, for the above three reasons of the existence of large fixed costs in the presence of indivisibilities, the technology of road capacity, and the earth moving costs, we can claim that there are economies of scale associated with the expansion from a two-lane to a four-lane road. Nevertheless, despite the fact that four-lane roads possess two-thirds dead space and eight-lane roads have only half the space for usable road capacity, it is not clear that economies from scale in urban highway construction exist. This is because it is rather difficult to control statistically for the effects of urbanization and separate it from the effects of size. For example, four-lane roads tend to be built in rural areas, where interchanges and overpasses are widely dispersed, and right-of-way costs are low. On the other hand, six-lane or eight-lane roads are built mainly near metropolitan conurbations, where expressway interchanges and overpasses are closely spaced together, and land acquisition costs are high. In practice, the road agency tends to trade off (and avoid) high right-of-way costs with increased tunnelling and overpass construction costs. Lane expansion from a six-lane to an eight-lane expressway at the margin, for example, would increasingly encounter alignment constraints associated with the terrain. This argument is independent of whether the expressway is located near urban areas. Hence all three cases of returns to scale occur, resulting in the classic U-shaped long-run average cost curve, paralleling that of a competitive firm of an industry as we have seen in Figure 3.5.

B. Diseconomies of scale
The discussion thus far centered on economies or diseconomies of scale to road width for single roads, as opposed to a system of roads. Strotz (1964)
conjectures, but Vickrey argues convincingly, that there are considerable diseconomies of scale associated with an urban road network. The argument is based on the geometry of road network (see diagrams in Hau, 1992a). As the urban road network expands from a typical single two-lane road to a double two-lane road, say, substantially costlier construction, tunnelling and land acquisition costs are encountered. Either higher construction costs or longer travel time and wait time costs due to the establishment of additional intersections in a road network would contribute to an increase in costs. Just as the congestion toll filled the wedge caused by the difference of short-run marginal and average variable costs, the divergence between long-run marginal and average total cost curves serves as an indicator of the unit profit. In competitive equilibrium, all economic profits are competed away in the long run, so the question that follows is in what sense is the case of diseconomies of scale a ‘long run’ concept? The presence of economic profits in the long run here is attributable to the rents earned by an invaluable fixed factor of production – land. Intuitively, just as the driver, in the short run, is charged for imposing external congestion costs on others due to his presence on the road, so also should the urban community, in the long run, charge for the increasing use of scarce urban land in a market economy. Put another way, if all factors of production – including land – were doubled, so that a scarcity value could be imputed to land, all economic profits would be competed away and vanish in the long run. Clearly, the supply of land cannot be doubled, so it is the existence of land rents which gives rise to long-run economic profits. Notice that we could no longer use the existence of profits as a surrogate market signal because of decreasing returns to scale. Since the urban road network is supposed to recover substantial amounts of revenue from high land values, relying solely on the profit mechanism and injudiciously investing in urban roads until all economic profits are competed away would result in over-investment in road capacity. With diseconomies of scale and divisibility, all roads generate profits. Performing proper economic appraisal of road projects cannot therefore be circumvented (Dodgson, 1997).

C. Indivisibilities and scale (dis)economies

The presence of both indivisibilities and scale economies could alter substantially the calculation of optimal tolls and subsidies (Kraus, 1981b). It turns out, perhaps surprisingly, that the existence of indivisibilities serves to improve the state of affairs vis-à-vis the road agency. For instance, in the case of rural roads with both scale economies and indivisibilities, there are regions where short-run marginal cost pricing yields profits rather than losses. This is because, with indivisibilities, the long-run marginal cost curve (composed of joined segments of the short-run marginal cost curves) is no longer declining all the way (as in Figure 3.5) but possesses a sawtoothed pattern, alternately cutting its corre-
spending long-run average total cost curve and short-run average total cost curves (see figure 12 in Hau, 1992a, which is summarized in Figures 3.7b–3.7c). Thus, whenever short-run marginal cost exceeds short-run average total cost at a given traffic level, profit exists and vice versa. It is, therefore, quite conceivable to have a congested road which generates profits even when subject to increasing returns to scale for a sufficiently large discrete change in capacity. The existence of losses does not mean that the road agency should cut back on the provision of highway services that passes the cost–benefit criterion.

How often do we encounter surpluses in the presence of scale economies and deficits in the presence of economies? The answer depends on the extent of the presence or absence of indivisibilities. There are two views on this issue. The first perspective argues that the aggregate road network could be regarded as divisible (see the works by Keeler and Small, 1977; Starkie, 1982). The other view, presented by Walters (1968, chapter 3) and Kraus (1981b), contends that roads are indivisible because the number of lanes – the main measure of highway capacity – is discrete.

The construction of a road or an additional lane may not be finely divisible, but taking the road network as a whole, a single newly constructed facility can be regarded as an incremental addition to the network, resulting in the applicability of the foregoing marginal analysis. Also, often varying some dimensions of road features other than the number of lanes increases the capacity of the road network. For example, the lane width, the provision of auxiliary lanes, horizontal and vertical alignments and the surfacing of road shoulders can all be varied incrementally (Starkie, 1982). One could characterize this view by treating the lane capacity as a continuous variable rather than a discrete one (Small, Winston and Evans, 1989, p. 103). If the road agency follows the twin optimizing rules of pricing and investment, then the road network would be in long-run equilibrium. So with constant returns and a divisible road network, roads would break even. However, some individual roads would make money and some would lose money. On the whole, if the economies and diseconomies of scale roughly cancel one another, the highway budget would be balanced. With indivisibilities, the profit (or loss) regime occurs about half the time but it is unclear what the relative weights would be when travel demand is reasonably assumed to grow over time.

Under decreasing returns and (almost) perfect divisibility, profits always occur as shown in Figure 3.6a. Perfect (and near-perfect) divisibility and an urban road network would mean that marginal cost pricing would always be profitable. This regime of ubiquitous profits disappears once indivisibilities set in sufficiently to admit downward sloping portions of the LRATC curve to occur as in Figure 3.6b. If the extent of indivisibilities progresses from small but significant (Figure 3.6b) to severe (Figure 3.6c), the regions which yield potential losses become larger initially. The symmetry carries
over somewhat to the increasing returns to scale case (see Figures 3.7a 3.7c). With perfect (and almost perfect) divisibility in the presence of scale economies (Figure 3.7a), losses would always occur. With scale economies and an intermediate level of indivisibilities (Figure 3.7b), say, smaller regions of profit would become available but would disappear when approaching the neighborhood of the limit (of divisibility). Nevertheless, if one were to accept Walters' (1968) argument that there are significant indivisibilities and scale economies in rural roads (as depicted in Figure 3.7c), we have demonstrated that profits (and losses, of course) would still arise under congestion tolling. Scale economies and financial viability are not necessarily incompatible.

Insights by Newbery (1988, 1989) and Small, Winston and Evans (1989) about the economic implications of the extensive damage that heavy vehicles cause to roads enrich the basic Mohring model. Charging for both the external and variable costs of road damage, by assigning a fee based on vehicle weight per axle (as opposed to weight alone), can go a long way towards covering any deficit arising from congestion tolling. Even if a road network is broadly characterized by increasing returns to scale in building and strengthening roads, the deficit could be closed by scope diseconomies. Diseconomies of scope means that a road network that accommodates both cars and trucks costs more than the sum of an autos-only and a (smaller) trucks-only road system (Winston,

![Diagram](image)

*Figure 3.6  Decreasing returns to scale and extent of indivisibilities*
Figure 3.6 continued
Figure 3.7  Increasing returns to scale and extent of indivisibilities
1991). So the surplus associated with diseconomies of scope offsets the potential loss from congestion pricing in the presence of scale-specific economies.

Recently, Arnott and Kraus (1998) extends the basic static Mohring-Harwitz result to the intertemporal setting. He shows that the self-financing result of congestible facilities does extend to this new environment in terms of present value. Arnott and Kraus demonstrates that discounted cost recovery in a growing economy depends not only on static returns to scale but also on the technological attributes of road capacity expansion and maintenance.

3.12 CONCLUSIONS

It is hardly surprising that congestion pricing as advanced in the past encountered its share of difficulties. This is because imposing a congestion toll has the effect of a tax increase on trip-makers, despite the fact that it is an externality-corrective tax. In the transport context – unlike in the textbook commodity case – the fact that the traveler is both a consumer and producer has interesting policy implications. With road use, the consumer-producer is a ‘perpetrator’ and a (willing) ‘victim’ (of negative externalities), as well as a ‘beneficiary’ of road use. As a generator and ‘perpetrator’ of negative externalities such as congestion and pollution, the traveler should be justifiably toll-taxed (as with the polluter pays principle). Yet the traveler also suffers
from the congestion and pollution externalities he helps engender. Hence as ‘victims’ of congestion externalities, perhaps travelers ought to be compensated. However, it has been argued that Pigouvian toll-tax revenues should accrue to the public treasury and should not be used to compensate ‘victims’ of externalities (Baumol and Oates, 1988, pp. 23–9). The intuition is that motorists would be induced to drive more because the level of compensatory payments would depend on their car usage, so economic efficiency would be violated. In this context, a road fund would be consistent with first-best pricing only if the funds were used in an indirect manner. As a consumer, the traveler pays for the benefits of taking a trip. So travelers are also ‘beneficiaries’ of road transport by virtue of their use of congested roads and their contributions to the toll revenue component of user charges. In the absence of lump sum transfers, earmarking of toll revenues could serve as a useful device in principle to approximating benefit taxation as a way of satisfying a commonly accepted notion of ‘fairness’. Combining these intuitive arguments and our stated principles of first-best optimal pricing and investment suggests that some form of dedicated funds is perhaps necessary – either in the form of a road fund or a transport fund – if congestion pricing is to gain political acceptance. When the proceeds of tolls are channeled back to the users in this way, congestion pricing and road taxes become a (road) user ‘fee’ (or mobility ‘fee’) and not regarded as a ‘tax’ per se.

Even without dedicated funds, it is essential to pursue steadfastly efficient pricing in the short run and efficient investment over the long run. Thereafter, the results can be presented for public scrutiny, thereby improving managerial efficiency and public accountability. By exploiting private profit incentives where competitive elements are strong, the competitive tendering and private provision of some transport services could also serve to enhance managerial efficiency in the public sector. However, the welfare gain from managerial efficiency due to private initiatives of road provision via increasingly popular build–operate–transfer (BOT) projects, for instance, should be measured against the welfare loss from monopoly abuse when parallel roads are next to nonexistent. Because many roads possess natural monopoly characteristics and since it is difficult to price various component parts of an integrated road network, the ownership of roads should best reside with the public sector (Newbery, 1994). The market failure resulting from the common property resource problem where no one really owns the roads would still call for the diligent application of optimal pricing and investment rules by an independent public road authority. Thus commercialization is in order rather than privatization. Vickrey (1996) also insists on ‘marketization’ – that is, the setting of quasi-market prices which enhances efficiency and acts as signals for (dis)investment – in transport and argues strongly against privatization in transportation.
Subject to further research, the idea of setting up a transportation (or road) fund to pursue marginal cost pricing in all its dimensions would enable us to satisfy the quintipartite principles of the World Bank's general guidelines for improving transport efficiency, as stated at the outset of this paper, which is namely to: (1) implement efficiency pricing; (2) meet economic viability; (3) meet (to a considerable extent) financial viability; (4) achieve (some degree of) 'fairness' among beneficiaries; and (5) attain (somewhat) managerial efficiency of the public authority. The conception of such a fund passes many of the tests for a 'good' earmarking arrangement as presented in McCleary (1991). The implementation of marginal cost pricing in the traffic (and loading) dimension could be achieved by employing the recent technological breakthroughs in automatic road use charging. These electronic toll collection devices using smart card technology and automatic classification systems all face remarkable scale economics. Alternatively, less powerful instruments and reversible setups such as area licensing and simple cordon pricing schemes can be used as stepping stones (Glaister, 1991).

NOTES

* Associate Professor of Economics, The University of Hong Kong. This paper draws extensively from research on my 'Economic Fundamentals of Road Pricing: A Diagrammatic Analysis', World Bank Policy Research Working Paper Series WPS 1070, The World Bank, Washington DC, December 1992, and research materials published since then. The conclusions in this paper do not reflect the views or policies of the World Bank. I take this opportunity to especially thank my mentor and friend the late Professor William Vickrey -- the grandfather of road pricing -- without implicating him for my errors. I should like to dedicate this paper to the memory of William S. Vickrey. Further, I thank others with whom I have sparred on road pricing. They include: Esra Ben Nathan, David Bernstein, Richard Bird, Ken Button, José Carbajo, Shanta Devarajan, Frank Englmann, Phil Goodwin, Clell Harrall, Tatsuo Hatta, Jake Jacoby, Jans Jansson, Frida Johansen, Odd Larsen, Kyu Sik Lee, Lars-Göran Mattsson, Herbert Mohring. Max Neutze, Gabriel Roth, Ken Small, Farideh Ramjerdi, Larry Summers and Sir Alan Walters. While making these acknowledgements, I retain full responsibility for the contents of this paper.

1. Professor Vickrey once told me that he thought that roads in the US were about a quarter to a third overbuilt. Vickrey's views on the severity of the congestion conundrum remained the same up until his unexpected demise in 1996 (Vickrey, 1996).

2. The regulatory approach 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.

3. Road use costs (both 'private' and 'social') include: (1) congestion (which is borne by road users); (2) pavement wear (which is typically covered by the road agency); (3) air and noise pollution; and (4) costs of accidents (both of which are borne by society at large). This paper deals mainly with congestion pricing (and only tangentially with pavement wear charges) as opposed to marginal social cost pricing, which is defined to include both the private costs and the external costs of congestion, air pollution, noise pollution, accidents, road damages and externalities (see Hau, 1992a).

4. See Beckmann, McGuire and Winsten (1956, figure 4.1); Mogridge (1990, figure 6.3); Jansson (1993, figure 9.4); and Johansson and Mattsson (1995, figure 1.5). In the next
section, we first assume that there are only two alternatives (i.e. periods): a peak and an off-peak.

5. 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 in Figure 3.2.

6. The relation AVC = AVC(Q) means that time cost depends on traffic level, and not vice versa. The backward-bending portion of the cost curve means that time cost continues to rise when traffic flow is reduced after engineering capacity is reached. The backward-bending portion of the cost curve has been substantiated in the literature (Gerlaough and Huber, 1975, chapter 4).

7. Recently, Alan Evans (1992) and Peter Hills (1993) have argued that output should be specified as the number of vehicles over a stretch of roads (that is, density), rather than the number of vehicles per unit of time as used conventionally in Nash (1982), Button and Pearman (1983), De Meza and Gould (1987) and Andrew Evans (1992), for instance. I maintain that the output variable used in the standard analysis is more appropriate because: (i) the (engineering) capacity of a road, in terms of vehicles per hour, is clearly built into the analysis and (ii) any commodity or service consumed clearly takes place in a specific time period (that is, in quantity per unit of time) and is implicit in the demand analysis of the commodity in question. Else (1981) argues that the demand variable is in fact a demand for the number of completed trips, but I argue that it is still necessary to express demand in quantity per unit of time. Hence the number of (vehicle) trips completed per unit time – which is simply the traffic flow – is the correct output specification. Demand and supply as I have interpreted them here are fully consistent with one another.

8. Formally, the functions GC(Q) and Q^e(P) intersect at an equilibrium point, where GC symbolizes the generalized cost. The equilibrium point is expressed as: Q^e(GC(Q^e)) = Q^e. Since GC is simply a translation of AVC, the interpretation of one is synonymous with the other.

9. Marginal cost is obtained as follows: MC = \Delta C(Q)/\Delta Q = AVC(Q)+ Q \cdot \Delta AVC(Q)/\Delta Q = AVC(Q) \cdot (1 + \epsilon) where C(Q) is the cost function, \epsilon is the elasticity of the AVC curve, that is, the rate at which time cost rises with respect to a one percent rise in traffic flow (Walters, 1961). The first term composed of only time cost – where trip time is converted into time cost via the (marginal) valuation of travel time – and the 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 (or efficient) 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 vehicle operating cost and variable road maintenance cost – both being constant with respect to traffic – simply shifts the right-hand side of the equation upwards.) 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 point V.

10. These major cities include: Bangkok, Bombay, Budapest, Buenos Aires, Jakarta, Mexico City, Pusan, Santiago, Sao Paulo, Seoul, Shanghai and Taipei.

11. The terms 'the tolled' and 'the tolled off' are coined by Zettel and Carll (1964) but their approach differs from our approach and results derived from first principles.

12. Cooperation is greatly enhanced if road users were guaranteed a reduction in motor vehicle-related taxes such as import duties, first registration taxes, annual license fees and/or fuel taxes (Small, 1992).

13. With standard commodities, 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 cannot decline but can only rise upwards.

14. With congestion tolling, note that high purchase taxes and registration/license fees of vehicles (if applicable) ought to be reduced to a level just sufficient to meet the administrative and enforcement costs of collection. If the variable road maintenance cost is constant with respect to the traffic level as we have assumed, an appropriate fuel tax could perhaps be used to approximate usage (as well as to tackle other externalities). As with the
services of public utilities, a pure tax element (such as a value-added tax) could also be imposed on top of the marginal cost of road use to meet general tax revenue requirements.

15. If (1) the capital and invariable maintenance cost of highway capacity, KC, is directly proportional to the engineering capacity, K, that is, KC(K) = aK, where a is a constant, then there exists constant returns to scale in highway construction (and invariable 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 input and the output variable of 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 average 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 vital to Mohring and Harwit's (1962, pp. 85–90) so-called theorem.

16. Prest (1969, p. 8) and Walters (1968, p. 33) argue for short-run marginal cost pricing whereas others like Meyer et al. (1959, chapter 4) argue for variants of 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 (1983, 1985) and Vickrey (1985)).

17. Formally, given a particular level of output, the cost-minimizing authority would expand the road up to the point where the marginal valuation in user cost savings due to a unit increase in capacity, – Q · ΔAVC(Q,K)/ΔK, equals the marginal cost of an extra unit of capacity, R(K). K(K), which depends on the level of highway capacity K, is the marginal rental cost per time period of capacity. It includes the invariable 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.

18. The superscript * symbol indicates that that variable is optimized.

19. First-best rules would yield economic efficiency only by assuming that the rest of the economy is marginal cost-priced. When that assumption is not satisfied, the theory of second best, with all its limitations, applies. Verhoef (1996) has recently handled these complex second-best issues in congestion pricing.

20. Estimates for the marginal congestion costs, congestion tolls and revenues for urban road use in Indonesia are reported in Hau (1994).

21. Alternatively, the costs of uncongested rural access roads could be covered by access charges such as annual license fees or local rates (Newbery, 1994). Vickrey (1996) argues strongly for a tax on land values.

22. It is due to the assumption of independent demands that long-run marginal cost pricing (equal to short-run marginal cost pricing) still holds at each time period. The concept of long-run marginal cost pricing is blurred in the case of jointness of demand.

23. Increasing financial costs of construction via tunnelling and/or flyovers, together with high land resumption costs, are consistent with the findings of Hau (1989) for Hong Kong.

24. Perhaps surprisingly, the symmetry of the LRMC curves in Figures 3.6a and 3.7a does not carry over exactly to the other cases in Figures 3.6b, c and 3.7b, c (see Hau, 1992a).

25. Space precludes us from elaborating on the empirical evidence of the scale economy issue (see Hau, 1992a). Walters (1968, pp. 184–5), using Meyer, Kain and Wohl's (1965, p. 205) data, shows that there are diseconomies of scale in the construction of four-lane, six-lane and eight-lane urban road segments. Keeler and Small (1977) find evidence of constant returns to scale for a sample of San Francisco Bay Area roads. Their often cited econometric study is important because of the balance budget implication for congestion pric-
ing. By contrast, using engineering specifications, Kraus (1981a) finds that there are increasing returns to scale in road construction in terms of length of freeway and interchanges but not for overpasses and length of arterials. Meyer and Gomez-Ibáñez (1981, pp. 191–2), in assessing the available estimates in the conflicting literature, conclude that economies and diseconomies of scale are ‘probably roughly offsetting’. Newbery (1989, 1994, p. 239) observes that ‘[i]f there are constant returns to expanding road capacity (as seems empirically plausible for those roads carrying the larger fraction of total traffic), then these [efficiency] prices will equal the maintenance costs and the interest on the infrastructural capital involved in an optimally adjusted road network’.

26. If a road fund were to be set up, compensation would need to be sufficiently indirect to satisfy Pareto efficiency. Thus the funds generated from money-making roads should not be tied to those roads but be made available for road construction and maintenance of the road network in general. The funds from profitable urban roads could be used to finance the fixed capital cost of worthwhile rural roads in a non-distortionary manner, for instance. To what extent can the profits collected from heavily used roads offset the losses arising from the construction of lightly used roads? The answer depends on the extent of the interaction of both scale economies and indivisibilities. A road fund is attractive because of the high marginal cost of raising a tax dollar. Moreover, a road fund run by an autonomous authority would increase the ‘linkage’ between revenues and expenditures, currently lacking in a politically-based budgeting process, thereby improving managerial efficiency. Without the setting up of such a fund, deficits from lightly used roads (with increasing returns to scale) would demand subsidization by the treasury, and would thus compete for tax money valued at a high opportunity cost. By symmetry, surpluses that accrue in heavily utilized urban areas (with decreasing returns to scale) should then be valued at a premium as (toll-)tax revenues. If these welfare losses and premiums offset one another when viewed within the same (transport) sector, then the nominal value of a dollar could be treated at its face value. This allows us to retreat to the standard case of pure efficiency concerns where a dollar is treated as a dollar to whomsoever it accrues. Even if a certain place is found to be faced with mainly increasing returns to scale, the deficit could be closed, in principle, by appealing to the notion of diseconomies of scope (see Hau, 1992a). The surplus associated with diseconomies of scope balances the potential deficit associated with scale-specific economies of road construction or use. The viability of the fund is enhanced by the fact that the maintenance cost of the road pavement is charged twice: once when traffic flow creates congestion, and the second time when traffic loadings cause road damage. Thus, the idea of a trust fund administered by an independent agency according to strict cost–benefit principles is likely to be viable.

27. Similarly, heavy vehicles ought to incur their ‘fair’ share of hefty pavement wear fees.

28. Alternatively, taking the surface transport sector as a whole, a transportation fund ought to be set up. If dedicated funds are set up in this way, indirect ‘compensatory’ payments can be achieved and would satisfy optimality. I recommend this both because the problem of highway congestion is tied intrinsically to the provision of poor transit alternatives and because public transport plays an important role in most places. The fact is that the production of bus services is subject to increasing returns due to scheduling frequency when passengers’ travel time is taken into account (based on the system economy of scale effect or the so-called ‘Mohring (1976) effect’). Hence additional funds in the form of a subsidy – preferably a user-side subsidy – is required to meet the financial shortfall arising from (first-best) optimal bus service provision. Road pricing would result in more crowded and inferior public transport services unless bus companies were to run more buses as a supply response. When this results in lower bus fares for passengers, the ‘untolled’ public transport users and captive riders would then be made better off. (Here the double charging of automobiles via traffic volume and heavy vehicles via loadings would help to close any deficit gap.) Increasingly popular rapid mass transit and light rail systems – both of which are subject to significant scale economies due to its large infrastructure costs – also require capital funds, the construction of which should be based on economic viability. Unless a global view is taken of the congestion problem and more rational time-of-day
Congestion pricing and road investment

pricing practiced on all modes (in contrast to tackling individual, non optimally priced modes), the urban transportation problem will continue to be pervasive.

29. Newbery (1994) calls for a road authority to be established like a public utility and subject to regulation. Because of roads' natural monopoly attributes, clear safeguards must be put in place via regulation to prevent, for instance, the curtailment of road supply to rake in congestion toll revenues (possibly in concert with environmental lobby groups). Freed from the fiscal shackles that governments face, Newbery argues that the Road Authority, when vested with the capital value of the existing road infrastructure, could finance efficient (but not necessarily profitable) road expansions. One advantage of establishing such an authority is that the politically sensitive issue of the nonhypothecation of tax revenues by the Treasury would be obviated. Further, if the Authority were to be run like an old electricity area board, gas council or water board in the United Kingdom, subcontracting or franchising could take place, all within an open and transparent manner to the (motorist)customer. Newbery (1994, p. 39) puts it aptly by observing that: '[t]he test of commercialisation is whether there is some intermediate allocation of the powers of pricing and investment to a Road Authority and away from the Treasury, subject to regulation on this narrower range of powers, and which nevertheless provides good incentives for efficient management and investment'. The feasibility of Newbery's proposal is buttressed by his estimates showing that the introduction of optimal road user charges in the United Kingdom to replace current road taxes there would adequately cover the cost of the road infrastructure. He also notes that with proper road use charges in place, competing rail services could raise their fares and thus pose less of a drain on government coffers. (Thus introducing road pricing would eliminate the second-best theoretical rationale for continuing to (nonoptimally) subsidize public transport.) Newbery's proposal is fleshed out in Roth (1996), who suggests that the application of commercial principles to roads would reduce congestion and pollution and simultaneously raise funding for roads.


31. Recent developments in electronic tolling in Norway, Sweden and England point to the fact that travellers do not object to road pricing when the toll revenues are earmarked for both road construction and improvement and/or the provision of better public transport. Indeed, a national survey conducted in England indicates that when people were asked whether they are for or against road pricing, about 57 percent are against it. However, when the question was posed in a different way: would they be supportive of a package approach to road pricing, with the revenues from road pricing used only to finance road construction and/or public transport, 57 per cent of the same surveyed population were in favor of road pricing (Jones, 1991; Goodwin, 1989; Grieco and Jones, 1994). A 1995 House of Commons Transport Committee report in the United Kingdom reveals the results of a survey conducted in Cambridge in 1994: the most effective combination of measures in combatting congestion was public transport improvements coupled with road pricing (Ison, 1996).

REFERENCES


Newbery, David M.G. (1990), ‘Pricing and congestion: Economic principles relevant


