Transportmetrica, Vol. 1, No. 2 (2005), 119-149

ECONOMIC FUNDAMENTALS OF ROAD PRICING: A DIAGRAMMATIC ANALYSIS, PART II – RELAXATION OF ASSUMPTIONS[1]

TIMOTHY D. HAU[2]

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

This is Part II of Hau (2005a, *Transportmetica*, **1**, 81-117) on the diagrammatic analysis of the economic fundamentals of road pricing, in which it assumes that the government aims to maximize welfare of the community by simulating the workings of a competitive industry and pricing highway services at marginal cost. There are a few major assumptions that need to be relaxed: 1) constant value of time, 2) static demand, 3) perfect divisibility, 4) constant returns to scale and 5) variability of road thickness. In this paper, we consider the relaxation of each assumption in turn.

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

1. INTRODUCTION

This is the second and concluding parts of Hau (2005a) on the diagrammatic analysis of the economic fundamentals of road pricing, in which it assumes that the government aims to maximize welfare of the community by simulating the workings of a competitive industry and pricing highway services at marginal cost. There are a few major assumptions that need to be relaxed: 1) constant value of time, 2) static demand, 3) perfect divisibility, 4) constant returns to scale and 5) variability of road thickness. Here, we consider the relaxation of each assumption in turn.

2. DIFFERENCES IN TIME VALUATION

The traditional presentation of road pricing and my ensuing critique assume a constant value of time (Walters, 1961).[3] The diagrammatic analysis in Figures 2 and 4 of Part I (Hau, 2005a) implicitly assumes that every driver is identical and maintains the same time valuation. What happens when there are heterogeneous motorists, with different time valuation and tastes? A mathematical proof that generalizes the above result for homogeneous drivers to heterogeneous ones with different values of time is shown by Mohring (1975), Mohring (1976, Chapter 4 Appendix) and Strotz (1964a, 1964b), but the intuition behind it is not difficult. Instead of the optimal toll being based on a representative driver's value of time, the time value is now a weighted average of the different motorists' valuation of time, weighted by the number of trips taken by those

[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. Part of this work also appeared in 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).

[3] In his pioneering work on road pricing, Walters (1961) (and authors thereafter) assumes that traffic is homogeneous, with all vehicles and drivers being the same, with the resultant identical valuation of time for all.

motorists who actually use the facility. If a traveler's time value and the number of trips are close to the average, he will pay the average toll payment. 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. He thus would be willing to, though begrudgingly, pay the difference. The congestion toll, P′ – P′′, in Figure 4 of Part I (Hau, 2005a) then can be labeled the *weighted* congestion toll, and the constant value of time is re-interpreted as the *weighted* average valuation of time. For a trip with a sufficiently higher than average time value, the time saving of a trip, (P^o – P′′), can be even higher than the *weighted* congestion toll, P′ – P′′, thus making the motorist better off. On the other hand, for a (shopping) trip having a lower time valuation, the user still has to make the average payment and therefore would be made 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 the generalized cost of their respective journeys. The use of alternative values of time would relax the point I made earlier that road pricing would make *all* groups except the government worse off. By relaxing the assumption of a constant value of time for everyone, those people with 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, and that a perfectly discriminating monopolistic authority is non-existent.

We conclude that a single transportation facility with differences in values of time would not alter fundamentally our derived result, using the standard assumption of constant returns. Again, with efficient pricing, financial viability and full cost recovery are achievable.

3. DEMAND VARIABILITY AND PEAK-LOAD PRICING

We have in fact considered the case of variable demands, *inter alia*, when we discussed the welfare impact of road pricing on the peak and the off-peak periods. There it was shown that a congestion toll is needed during the peak when there is excess demand but not during the off-peak when there is excess capacity. Notice that because there is free-flowing traffic in the off-peak, no tolling is required because no external cost of congestion is generated. Therefore no quasi-rent is being earned on the invested capital and only the variable costs are paid for by the traveler. On the other hand, during the peak period, a positive quasi-rent is earned. (Suppose further that there is an inter-peak period, some quasi-rents are also generated.) With highway capital stock remaining unchanged, the systematic, diurnal nature of travel demand (as opposed to the static, invariant demand case) means that the *sum* of quasi-rents (rather than just the quasi-rent from the singular peak 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, expansion of the highway is either warranted or not, under constant returns. The same conclusions obtained thus far again hold.[4]

An interesting implication is that the entire capital cost of the highway is 'allocated to' and borne by peak travelers, mainly rush-hour commuters. This surprising result may

[4] The output variable needs to be redefined as vehicles per lane per cycle, with the cycle being the duration of a particular charging period.

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 which require massive infrastructure developments. Without them, the optimal size of the road would be considerably smaller. The result of allocating all capital costs to users of the peak period has long been recognized in the literature on the pricing of public utilities a la Boiteux (1960). Following our earlier example, the extent to which there is another period -- the interpeak or shoulder period -- with even a modest amount of congestion, would allow for differential pricing and thus the allocation of some capital costs to these interpeak travelers. The optimal investment rule is then to expand a road until the *sum* of the quasi-rents over the demand cycle equals the entire capital cost of the facility under constant returns. By implementing peak-load pricing and altering the investment level of the highway facility, depending on whether profits are positive or negative, the highway network is again 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 using peak-load pricing in the presence of independent demands and no indivisibilities.[5] By assuming the demand in each period in fact depends on other periods, i.e., the case of dependent demands, the derived results still go through (Mohring, 1970).[6]

4. 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 bi-directional. In the perfectly divisible case, the long-run average total cost curve which envelopes 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 9b of Part I (Hau, 2005a)). Due to the presence of indivisibilities, however, the formerly neat and continuous pattern of the LRMC curve is broken (Neutze, 1966; Kraus, 1981b). The new long-run average total cost curve is now composed of a series of short-run average total cost curves, where $SRATC_2$, $SRATC_4$ and $SRATC_6$ denote a two-lane, four-lane and six-lane road respectively. The long-run average total cost curve is a series of short-run average total cost curves connected

[5] It is due to the assumption of independent demands that long-run marginal cost pricing (equals 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.

[6] Using the distribution of current demand distribution as given, which is synonymous with assuming independent demands, would result in upward bias in peak periods and downward bias in off-peak periods because of the possibility of substitution (Keeler and Small, 1977).

together in a scalloped-like pattern (see the **solid** LRATC curve labeled ABCDEFG in Figure 1).[7] The long-run marginal cost curve takes on the various short-run marginal cost curves in the respective regions, resulting in a discontinuous shark's tooth-shaped LRMC curve (see the **thick** LRMC curve in Figure 1a). Hence one is always working with the short-run curves themselves since one could only operate with a capital facility which is given. By now, it should be clear 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 region of $Q_2Q_{2\text{-}4}$, $Q_4Q_{4\text{-}6}$ and $Q_6Q_{6\text{-}8}$, and multiples thereof, then the road makes money in the long run under constant returns. *Per contra*, to the left of the outputs, Q_2, Q_4 and Q_6, namely in the region of OQ_2, $Q_{2\text{-}4}Q_4$ and $Q_{4\text{-}6}Q_6$, the road loses money. With a 2-lane road, as traffic increases, the road's large fixed cost is spread over the additional traffic, and as congestion sets in, the road begins to make 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, etc. 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.

4.1 Optimal pricing and investment with indivisibilities: an example

Consider the demand as depicted in Figure 1b, with the actual consideration of indivisibilities. First, we follow the first-best pricing rule of tolling the difference between short-run marginal cost and short-run average variable cost, t*. This yields the optimal traffic level of Q* and a positive profit. Without the indivisibility constraint, the existence of profit would indicate that the road is underbuilt. With indivisibilities, the direct one-to-one correspondence between economic profit and road expansion is lost. One is therefore left with the binary choice of, say, expanding the road from a 2-lane road to a 4-lane road. Using the welfare apparatus that we have developed in the Appendix and Figures 4 and 5 of Part I (Hau, 2005a), however, the net benefit of such a move is shown to be the sum of the welfare gain of going from $Q_{2\text{-}4}$ to Q**, as indicated by the triangular areas $b+c$, and the welfare loss of moving from Q* to $Q_{2\text{-}4}$, as shown by the triangular area a. Such a move would clearly be desirable if $(b+c) - a$ exceeds the capital cost of the move. With a 4-lane road, however, the optimal toll of t** would be insufficient to cover the fixed cost of this new road, resulting in a shortfall. Thus, even supposing that the move from Q* to Q** is a beneficial one after incurring the investment cost, it would mean that the road authority would switch from a profitable regime to a loss regime.

Thus the optimal sequence of decision-making is first to establish the policy of implementing marginal cost pricing and then to plan future adjustments of the road network according to *expected* future demand and established pricing policies. When demand fluctuates, pursuing short-run marginal cost pricing at present would mean setting different prices, or tolls, in response to expected current conditions.

---

[7] A similar set of discontinuous curves is found in, for example, Bennathan and Walters (1979, Figure 2.2). Note that the scalloped-like pattern is asymmetric because the unit cost curves for four and six lane roads are horizontal multiples of those of the two-lane road. In other words, under constant returns to scale, the AVC and AFC curves, and hence ATC curves fan out horizontally -- a mistake that is quite frequently made in the literature (see, for example, Hayutin (1984, Figures 2.8 and 2.16)).

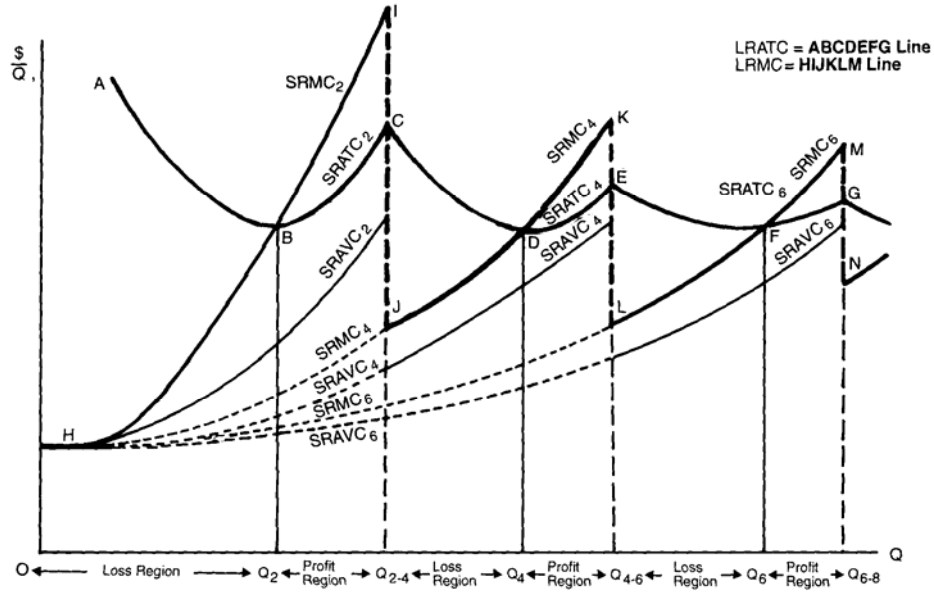

FIGURE 1a: Road indivisibilities under constant returns

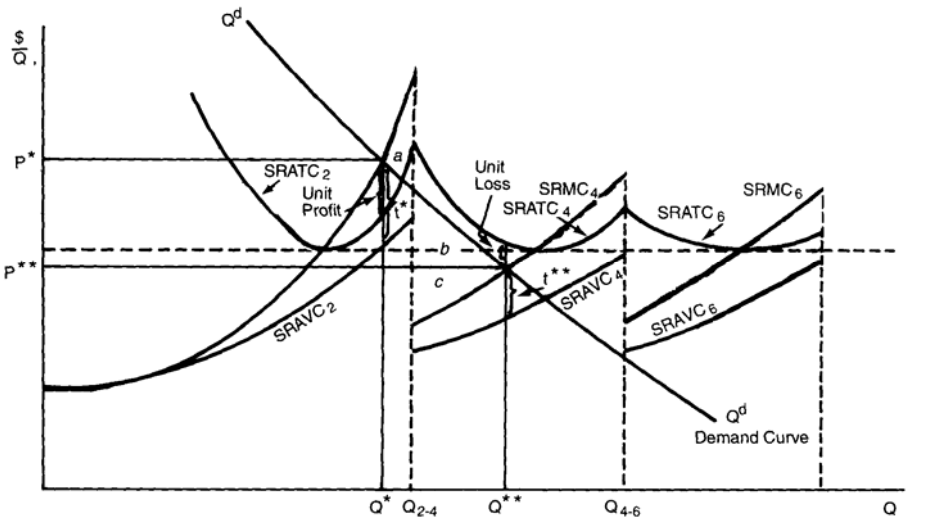

FIGURE 1b: Optimal pricing and investment with indivisibilities: expansion from a 2-lane road to 4-lane road

Suppose the government were faced with a tight budget constraint. It would understandably then be unable to undertake all public projects with positive net benefits. *If* this road authority were mainly concerned about cash flow and financial viability, it could opt not to expand a road, i.e., under-invest, but still charge a congestion toll on the built-up traffic and satisfy economic efficiency in the short run. This option however would not lead to the maximization of society's welfare in the long run.

5. 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 only via careful statistical analysis. The available evidence in road transportation indicates that all three cases exist: decreasing, constant and increasing returns to scale (see Figure 2) --

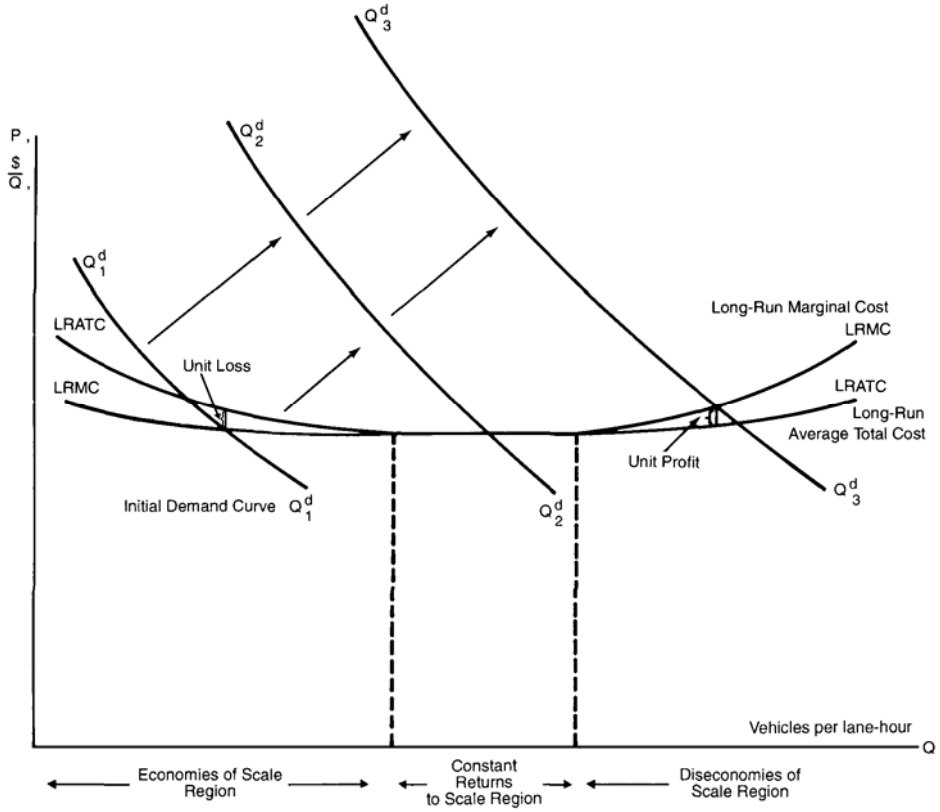

FIGURE 2: Economies and diseconomies of scale in the provision of road capacity with the growth of travel demand

paralleling the case of a competitive private firm and industry -- with profit, zero profit 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 neglected. The implication is that if traffic were to grow to a point where the capacity of a road is reached, congestion delay would set in, and congestion toll revenues could be collected. After all, the short-run marginal cost curve is always non-decreasing, as shown in Figure 1. Profits may occur *despite* the fact that the long-run average and cost curve may be declining and the corresponding long-

run marginal cost curve lies below the average cost curve.[8] Then 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 2). In the case of increasing returns with perfect divisibility (where natural monopoly arguments lie), the long-run marginal cost curve below the long-run average total cost curve pulls it downwards, resulting in losses, beckoning government subsidization. On the other hand, if travel demand is sufficiently high relative to the engineering capacities of roads, the money-making road 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. We turn next to a discussion of the theory underlying the economies vs. diseconomies issue, together with perfect divisibility vs. indivisibilities, and end with a review of the empirical evidence and past work.

5.1 Economies of scale and rural roads

There is a preponderance of evidence -- geometric, engineering and otherwise -- supporting the case that there are significant economies of scale in the construction of *rural* roads (Walters; 1968, pp. 180-182; Mohring, 1976, pp. 140-142)).[9] The illustrations below follow Mohring's scale economy analysis using the geometry of transport right-of-way. 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 non-trivial proportion of the provision of a road's right-of-way involves dead space. These indivisibilities -- required to contribute to the building of a minimum acceptable standard such as a given pavement thickness and road size -- help contribute to economies of scale as the large fixed cost of construction and invariate maintenance costs are shared over greater amounts of traffic. Thus doubling the width of a two-lane road more than doubles its traffic 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 average engineering capacity of 1800 - 2000 vehicles per *lane-hour*, doubling the width of a two-lane road almost quadruples its capacity (see any Highway Capacity Manual, for example, Transportation Research Board (1985, Tables 2-1 and 2-2), yielding economies of scale associated with road use. However, capacity per lane remains constant beyond four-lane roads, resulting in zero economies of road use thereafter. Further, in order to level hilly terrain and/or fill valley for transportation purposes, the earth moving costs rise less than proportionately. In fairly flat or rolling country selected as sites for road building, doubling the width of a two-lane road generally involves less than doubling the earth moving costs. (On steep hillsides, however, the reverse may be true.) Hence, for these three reasons of: 1) the existence of large fixed costs due to indivisibilities, 2) the technology of road capacity, and 3) the possible reduction in earth moving costs, we can claim that there are economies of scale associated with the construction of a two-lane to a four-lane road. Nevertheless, despite the fact that typical

---

[8] This point will be established later in Figures 5 and 6a. Basically, with perfect or almost perfect divisibility and scale economies, only losses will occur, whereas with indivisibilities, either profits or losses will result depending on the level of travel demand.

[9] Meyer et al. (1965, pp. 200-204) also found some evidence of scale economies of urban roads. However, Keeler and Small (1977, pp. 5) query Meyer, Kain and Wohl's findings, stating that their results are in fact based on their initial assumptions.

four-lane highways possess two-thirds dead space and eight-lane highways have only half the dead space, it is not clear from the geometry of highway rights-of-way that economies of scale in *urban* highway construction exist. This is because it is rather difficult to control econometrically 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 authority tends to trade off (and avoid) high right-of-way costs with increased tunneling and flyover 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. (These constraints might explain why capacity per lane is reduced with six- to eight-lane expansion (Mohring and Harwitz, 1962, pp. 97).) This argument is quite 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 firm within a competitive industry as we have seen in Figure 2.

5.2 Diseconomies of scale and urban roads

The discussion thus far has centered on economies of scale to road width for single roads, as opposed to a *system* of roads. Strotz (1964a) conjectures, and Vickrey argues convincingly, that there are considerable diseconomies of scale associated with an urban road networks.[10] The reasoning is based on the geometry of road network. Given a rectangular grid for an urban road network spaced 2 kilometers apart, as in Figure 3a, there are 9 sets of (space-intensive) intersections and traffic lights in a 6 kilometer-wide area. Suppose the number of streets is doubled in order to double road capacity, yielding a grid of a one kilometer-wide spacing (Figure 3b). Quite apart from the possible increase of construction or land acquisitions costs, the number of intersections -- and the required land area and traffic light installations -- is quadrupled to 36. (If no traffic lights are installed, each intersection requires an even costlier overpass or perhaps even a full interchange, a case considered by Mohring (1976, pp. 144-145).) Moreover, despite the fact that trip length remains unchanged between the origin, O, and the destination, D, the number of traffic lights encountered, and hence waiting costs, would then double from 4 to 8 (unless overpasses or interchanges are constructed). Either the installation of traffic lights or the building of overpasses and interchanges would serve to bid up the opportunity cost of land (because less non-highway space would be left for business or other activity) as well as to increase substantially the sum total of the costs of undertaking a trip to the community, resulting in a rising long-run average cost and also a long-run marginal cost curve. The resultant long-run equilibrium for an urban road network in the presence of diseconomies of scale is depicted in Figure 4. The analysis of the relationship between long- and short-run cost curves is similar to the case of constant returns to scale and is not repeated here. However, because the long-run average cost curve is rising, the short-run average total cost curve for a two-lane road, $SRATC_2$, is tangent to the long-run average cost curve to the right of the minimum $SRATC_2$ point.

[10] I am extremely grateful to William Vickrey for pointing out the subtleties of these arguments. See Vickrey (1965, pp. 287-288) and Mohring (1976, pp. 144-145) also for the basic line of reasoning.

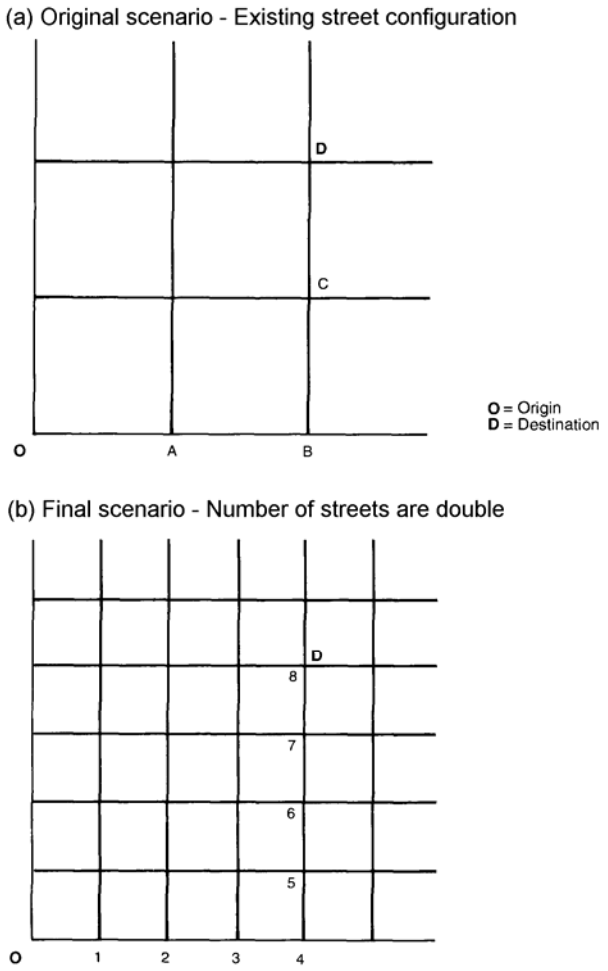

FIGURE 3: Doubling the number of streets – road capacity – results in quadrupling the number of intersections and traffic lights and doubling waiting time

Because the $SRMC_2$ lies above the $SRATC_2$, $SRMC_2$ intersects the LRMC from below. Short-run marginal cost pricing is equivalent to long-run marginal cost pricing at the efficient output level Q_2*. Notice that the optimal toll, t_2*, follows from pricing at short-run marginal cost and tolling the difference between short-run marginal cost and short-run average variable cost. The difference between $SRATC_2$ and $SRAVC_2$ is, by definition, $SRAFC_2$. The optimal toll, t_2*, clearly exceeds the $SRAFC_2$ by the unit profit of π_2*, with the corresponding rectangular area of profit $\pi_2^* \cdot Q_2^*$. As the urban road network expands from an existing single two-lane road to a double two-lane road, say, substantially costlier construction, tunneling and land acquisition costs are encountered.[11] We have also established that time costs rise because of additional

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

intersections and wait time required. Either of the above two factors would serve to push the SRAFC and SRAVC curves up, together with the SRATC curve. Just as the congestion toll filled in the wedge caused by the divergence of short-run marginal and average variable cost curves, the divergence between long-run marginal and average total cost curves serves as an indicator of the unit profit (or loss). In competitive equilibrium, all economic profits are competed away in the long run, so the questions that follow are: in what sense is the case of diseconomies of scale a 'long-run' concept, and what is the interpretation of the profit areas of $\pi_2^* \cdot Q_2^*$ and $\pi_4^* \cdot Q_4^*$ (for demand curves Q_2^d and Q_4^d, respectively)? The existence of economic profits in the long run is attributable to the rents earned by an invaluable fixed factor of production -- land.[12] Even though both the SRAFC and the SRAVC curves reflect the increase in costs mentioned above, the existence of the rising opportunity cost of the fixed factor of the remaining parcels of land is still left unaccounted for. Intuitively, just as the driver, in the short run,

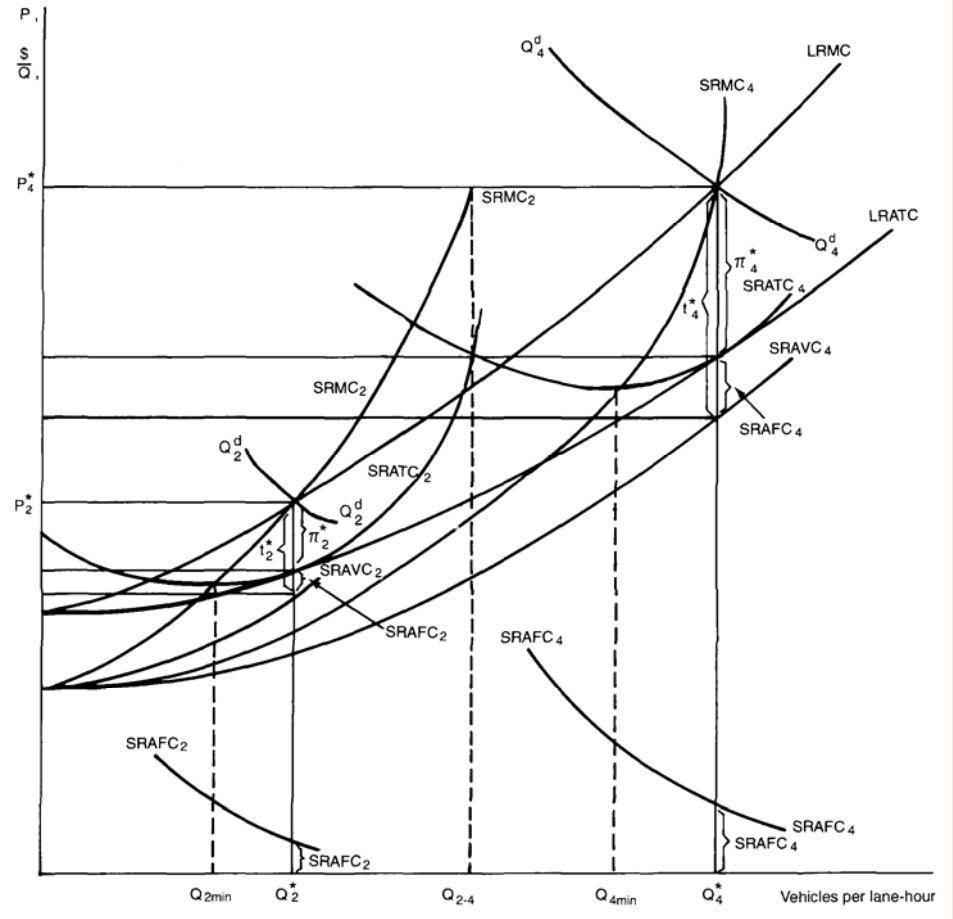

FIGURE 4: Diseconomies of scale: urban road network with perfect divisibility, π = economic profit

[12] A purist might argue that these rents actually accrue to land owners and the usual story of long-run super-profits being zero still holds in the presence of rising long-run average costs.

by imposing external congestion cost on others due to his presence on the road, is charged for it, so also should the urban community, in the long run, charge for the use of scarce urban land in a market economy. Putting it 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 yields (the areas of) economic profits in Figure 4. Notice that we could no longer use of the existence of profits as a surrogate market signal because of decreasing returns to scale. Since the urban road network is *supposed* to generate substantial sums of money because of high land values, relying solely on the profit mechanism and incautiously investing on urban roads until all economic profits are competed away would result in over-investment in road capacity. With diseconomies of scale and divisibility, roads generate positive profits. Performing proper project appraisal of roads cannot therefore be circumvented.

5.3 Diseconomies of scale and indivisibilities

The case of diseconomies of scale combined with indivisibilities is similar in spirit to the analysis of the constant returns case with indivisibilities in Figure 1. For instance, at any moment in time, the regions to the right of Q_{2min} and Q_{4min} in Figure 4 would yield profits. In the case of decreasing returns to scale, with or without indivisibilities, the correct recourse is to invest only if the road project in question passes a stringent cost-benefit test and only if rising costs of roads are compared to the quasi-rents generated in the variable demand periods diurnally. In other words short-run marginal cost pricing, or congestion tolling, as opposed to long-run marginal cost pricing, should be implemented for each period over the demand cycle and the quasi-rents summed up, *regardless* of whether the facility is optimally built. When demand is not known with certainty, probabilistic or expected demand could be used instead and *all* real benefits and costs over the economic life and time periods of a project should be properly measured, discounted and compared with the capital cost of implementing the change. If finely divisible projects are available, we have seen that the procedure is to invest until the marginal benefit of a project -- in the form of time savings -- equals the marginal cost of constructing the capital facility. If indivisibilities abound, then traditional public investment criteria using the net present value criterion or Mishan's (1988, Chapters 35-38) normalized internal rate of return procedure could then be complemented with the diagrammatic welfare analysis presented in this paper. In such a case, unfortunately, the neat linkage between profits and losses as a 'market' mechanism and guide to investment is severed. Thus it is possible for a project with high net benefit to yield a financial deficit. As mentioned before, if a road authority were faced with a severe fiscal constraint, then the road agency could choose to under-invest, rather than over-invest, when faced with an all-or-nothing situation of indivisibilities. Still, it should pursue marginal cost pricing in the short run, while fully taking into account the opportunity cost of congestion toll financing (see Hau (2006)). In this way, short-run marginal cost pricing would yield both economic efficiency and profit, even though an optimal investment strategy would generate an even higher level of net benefit for the community in the long run.

5.4 Economies of scale and indivisibilities

Per contra, the above analysis for the case of decreasing returns to scale of an urban road network carries over in reverse to the case of increasing returns to scale for rural roads. There the sum total of the quasi-rents captured via short-run marginal cost pricing is insufficient to cover the entire fixed cost of a particular rural road. Scale economies abound in the construction of rural roads, so that unit losses, L_2* and L_4*, result in the case of an optimized two-lane and four-lane rural road, respectively (see Figure 5). Standard neoclassical arguments then call for subsidization out of the public treasury for the case of scale economies. The presence of both indivisibilities and scale economies could alter the calculation of optimal tolls and subsidies substantially (Kraus, 1981b). It turns out, perhaps surprisingly, that the existence of indivisibilities serves to improve the state of affairs vis-a-vis the government because, as was shown in the constant and decreasing returns to scale cases with indivisibilities, both surpluses and deficits would occur alternately, depending on the level of travel demand. Similarly, in the case of rural roads with both scale economies and indivisibilities, there are regions (such as those to the right of Q_{2min} and Q_{4min} in Figure 5) 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 but possesses a sawtoothed pattern, alternately exceeding its corresponding long-run (and short-run) average total cost curves and rising at an even faster rate. Thus, just as in the case of constant returns with indivisibilities, whenever a SRMC curve exceeds a SRATC curve, profit exists and vice versa. It is therefore quite conceivable to have a congested two-lane road which generates profits even when subject to *increasing* returns to scale for sufficiently large changes in capacity. The existence of losses does not mean that the road agency should cut back on the provision of highway services. On the contrary, it merely means that other sources of funds ought to be sought in order to finance a worthy project. The government authority's decision to cut back services just because of losses in such a situation would yield under-investment and possibly stifle economic development and growth in the longer run. The neat linkage between road expenditures and toll revenues has disappeared. Again, the passage of a tough cost-benefit criterion is then a prerequisite for a project to result in maximizing society's welfare. To repeat, *only if* congestion toll financing is all that is sought by the road agency, and not optimal investment, is it possible to let a two-lane road become congested in the face of rising urbanization and motorization, while differentially pricing it via congestion tolls. In this way, efficient use of a given road is enhanced via short-run marginal cost pricing, but the efficient level of road capacity is not being achieved in the long run.

5.5 The extent of indivisibilities vs. divisibility and their effects on scale (dis)economies

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

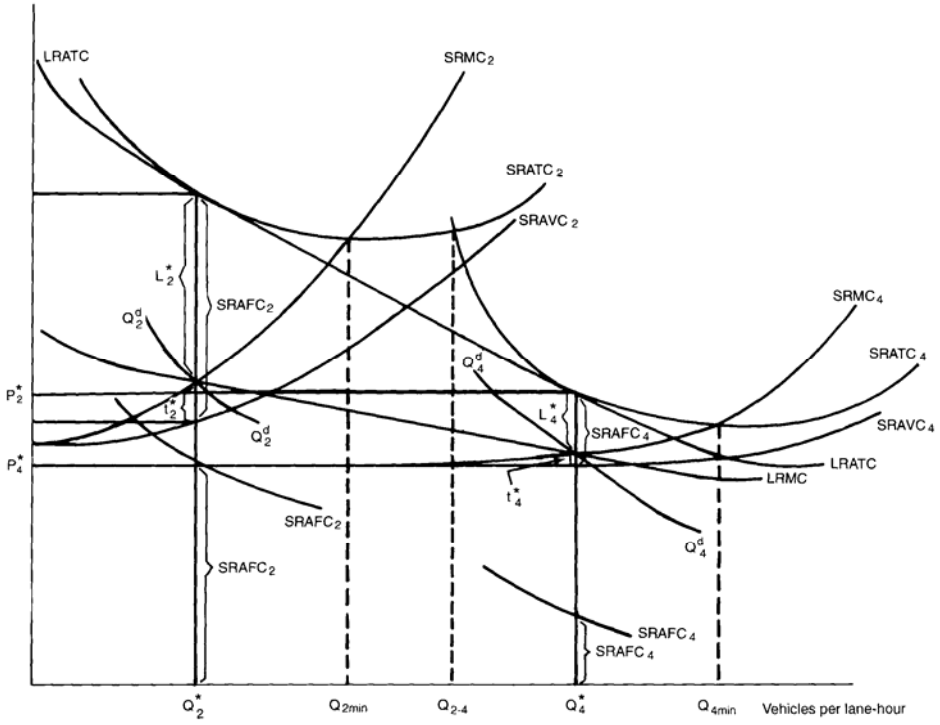

FIGURE 5: Economies of scale: rural roads with perfect divisibility, L = loss

The construction of a road or an additional lane may not be finely divisible in and of itself, 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 (Keeler et al., 1975, Chapter 2). Also, often times, 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 et al., 1989, pp. 103). If the road authority pursues the twin optimizing rules of pricing and investment in capacity, then the road network would be in long-run equilibrium. So with constant returns and a divisible road network, roads would break even.[13] However, some individual roads would make money and some would lose money. On the whole, if the economies and diseconomies of scale are "probably roughly offsetting" as Meyer and Gómez-Ibáñez (1981, pp. 191-192) concluded, then the highway budget would be balanced. Whether or not scale economies and diseconomies are counter-balancing would depend on the degree of urbanization. With increasing urbanization, profits would tend to predominate even after charging land rents as a cost. With indivisibilities, the

[13] Constant or even decreasing returns would be satisfied if two bi-directional roadways are built in proximity to one another.

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 perfect divisibility, we have shown in Figure 4 that profits always occur. The essence of that figure is combined with Figures 6a through 6c. Perfect (Figure 4) and almost perfect divisibility (Figure 6a) and an urban road network would mean that the marginal cost pricing of trips would always be profitable. Note that the continuous regime of profit will not be uniform once a threshold level of indivisibilities is reached, resulting in regions where losses also occur. If the extent of indivisibilities progresses from small but significant (Figure 6b) to severe (Figure 6c), the regions which yield potential losses become even larger. The symmetry carries over to the increasing returns to scale case. Again, the substance of Figure 5 is culled and combined with Figures 7a through 7c. With perfect (Figure 5) and almost perfect divisibility (Figure 7a) in the presence of scale economies, losses would always occur. With scale economies and a 'significant' level of indivisibilities (Figure 7b), smaller regions of profit would become available but would disappear when approaching the neighborhood of the limit (Figure 7a). Nevertheless, even if one were to accept Walters' (1968, Chapter 6) argument that there are significant indivisibilities and scale economies in *rural* roads, we have demonstrated that profits (and losses, of course) would nevertheless arise under congestion tolling. Scale economies in the presence of indivisibilities and financial viability are *not* necessarily incompatible.

5.6 Empirical evidence on the scale economy issue

Fitch et al. (1964, pp. 131) give some numerical support for the case of scale diseconomies in the United States. Walters (1968, pp. 184-185), using Meyer et al.'s (1965, pp. 205) data, shows that there are *diseconomies* of scale in the construction of four-lane, six-lane and eight-lane *urban* road segments. (By employing Walters' straightforward approach, Hau (1989) demonstrates that there are increasing costs associated with a sample of four-lane, five-lane and six-lane roads in Hong Kong.) Without imposing any prior specifications about the extent of returns to scale -- 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 pricing, a result which was also quoted by Newbery (1990).[14]

[14]Keeler and Small's (1977) result is one of the more rigorous econometric analyses bearing on the scale economy issue, in that they are able to separate the confounding effects of size and urbanization. In particular, they regressed construction cost per lane-mile on the number of lanes and various discrete variables which capture the differences between urban, suburban and rural-suburban areas. It is the inclusion of the latter variables which enable them to econometrically control for the effects of expressway capacity on construction cost. The selected sample of 57 roads was based on all state-maintained roads for nine San Francisco Bay Area counties, including arterials, expressways and rural roads. They estimated both a non-linear and log-linear specification, with both yielding statistically insignificant statistics for the estimated degree of homogeneity. Based on the slightly better fit for the log-linear specification, the returns to scale parameter of (–)0.0305 translates itself into the case of "mildly" increasing returns to scale of 1 – 0.0305 = 0.9695. By taking the reciprocal of the estimated returns to scale parameter, the economies of scale degree of 1.0315 is obtained. Since 1.03 is "statistically indistinguishable" from 1, they conclude that there is no firm evidence for scale economies in road construction measured in terms of lanes. The far-reaching policy ramifications of this result could perhaps be buttressed if there were calculations or discussion of the power of their hypothesis tests. This is because not being able to reject a null hypothesis is not necessarily equivalent to accepting the null

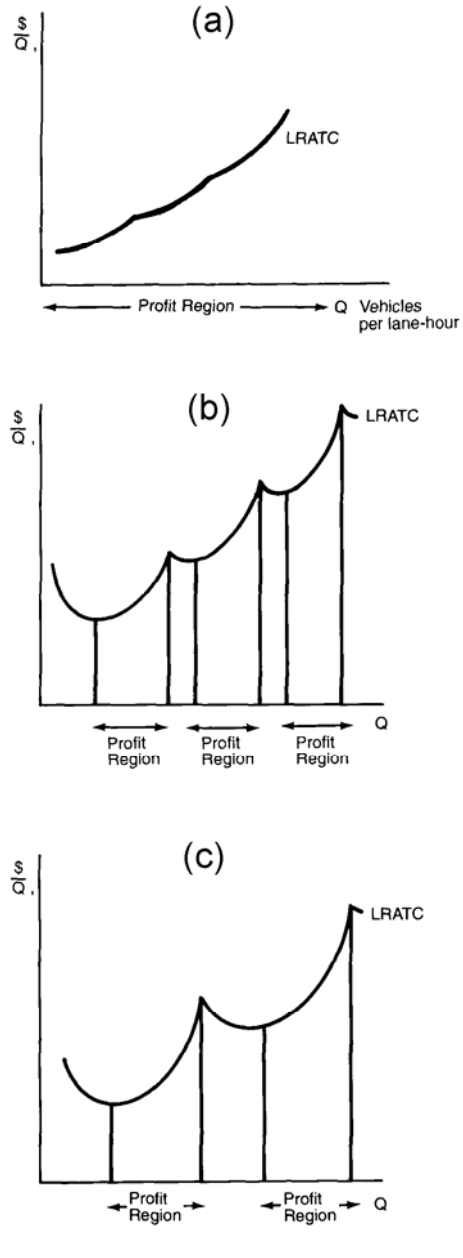

FIGURE 6: Decreasing returns to scale and extent of indivisibilities

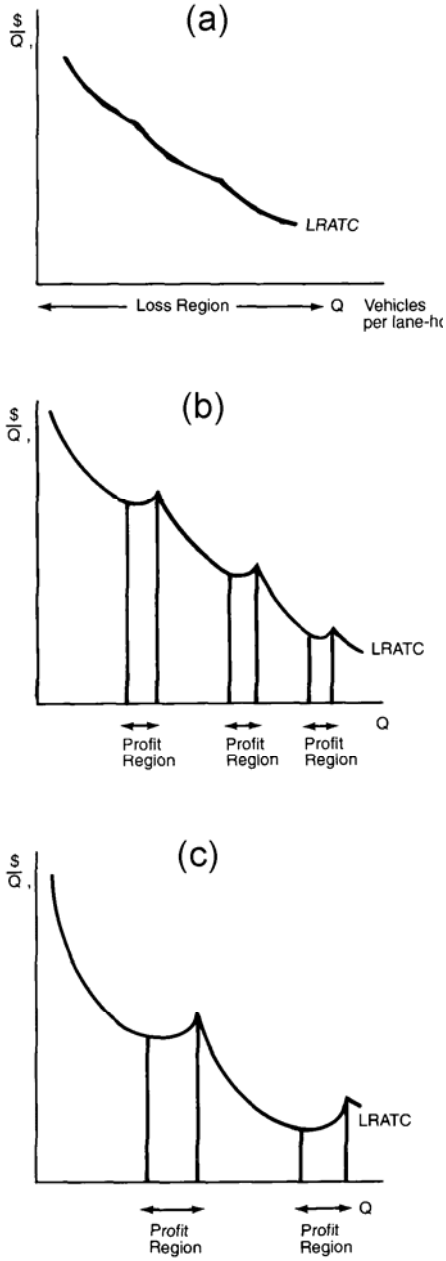

FIGURE 7: Increasing returns to scale and extent of indivisibilities

hypothesis. One hopes that the type II error is relatively small. However, to be fair to their important contribution, the power function for published papers are seldom to be found in the literature.

By contrast, using engineering specifications to estimate each of the cost components of an urban highway network model, 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. He makes the crucial observation that factor prices (such as right-of-way prices) are to be held constant for making relevant comparisons of scale - specific (dis)economies. The reciprocal of his best "pseudo-empirical" estimate of returns to scale in urban highway network capital costs is 0.84, which translates to the economies of scale degree of 1.19. Meyer and Gomez-Ibáñez (1981, pp. 191-192), in assessing the available estimates in the conflicting literature, conclude that economies and diseconomies of scale are "probably roughly offsetting." Hayutin (1984), in an unpublished dissertation, refined the Keeler-Small model and applied it to a sample of U.S. Interstate Highways. By including both intercity (or rural) and urban routes in her data set, she estimates that there are clearly increasing returns with respect to the number of lanes. Her results bear out Mohring's scale economy implications of highway geometry with respect to the "shoulder effect".[15] In reassessing the earlier work of Keeler and Small (1977), Small et al. (1989, Chapter 6) use the result of zero economies (corresponding to the scale economy parameter of 1.00) as well as the mildly increasing returns case of 0.97 (corresponding to the scale economy parameter of 1.03) (see Small (1992, Chapter 3) for a review also). Fully cognizant of the possibility of the existence of increasing returns to scale in urban highway travel a la Kraus, they presented their simulation results on highway finance based on three encompassing parameters for the degree of scale economy: namely 1.00, 1.03 and 1.19.

5.7 Results on cost recovery

As part of a major World Bank research project, Newbery (1988a, b, c, 1989, 1990) and Newbery et al. (1988) extend Mohring's now classic result of highway finance by relaxing the assumption of an infinitely durable highway whose capacity can be continuously adjusted. It is common knowledge that pavements wear out after extended use, but it is less well known that the damage of a vehicle on the road pavement is related to the *weight per axle*, as opposed to the gross vehicle weight *per se*. Common Practice (Highway Research Board (1962)) is to measure this damaging power by the number of equivalent standard axle loads (ESAL), where one ESAL is equivalent to the load of 18,000 lbs. (8.2 tons or 80 kilo newtons) single axle. The American Association of State Highway and Transportation Officials' milestone road test indicates that the damaging power of a vehicle on pavement is proportional to the fourth power of individual axle loads, acquiring the title of the "fourth-power law."[16] For instance, the

[15] Hayutin (1984, Chapter 4) regressed construction investment per mile on the number of lanes and other variables similar to those used by Keeler and Small (1977) (see previous footnote). Further, by including the paved width per lane variable, diseconomies to width measured in feet, rather than in lanes, result. Hayutin (1984, Chapter 5) concludes that the stronger effect of scale economies with respect to the number of lanes dominates the effect of diseconomies associated with building wider footage. In an excellent review of the literature surrounding the scale economy issue, Hayutin (1984, pp. 158-159) observes that Meyer et al.'s (1965, pp. 200-214) conclusion that there are increasing returns to scale in freeway construction stems from their engineering assumptions about costs, as opposed to actual estimation. Her statistical results, however, are supportive of their conclusions. My comment regarding the power of the hypothesis tests applies to her study also.

[16] Formally, the damaging power (in EASLs) of an axle load, l, in tons, is approximately equal to: $(l/8.2)^4$, hence we say that the damaging power is proportional to the fourth power of the axle load. As a hypothetical

rear axle of a 13-ton light truck with two axles can result in over 300 to 2400 times more pavement damage than that of a large car weighing 2 tons.[17] One inescapable conclusion is that almost all damage is caused by heavy vehicles such as trucks and buses: relatively little is due to automobiles. Newbery (1988a) characterized another type of externality called a road damage externality: the *external* damages imposed upon the pavement by heavy vehicles result in increased vehicle operating costs to subsequent motorists for the rest of the periodic life cycle of a road. (Thus, the various costs of undertaking a trip mentioned in the introduction of this paper (footnote 5 of Part I (Hau, 2005a)) should now be more narrowly defined as *external* costs, namely: 1) marginal external congestion cost; 2) road wear cost *and* damage externalities (grouped under the rubric of road damage cost); 3) environmental costs; and 4) costs due to increased risk of accidents not borne by private parties.)[18]

Newbery's (1988a) "Fundamental Theorem of Road User Charges" says that *if*: 1) the maintenance policy pursued by a highway authority is condition-responsive, i.e., road maintenance is carried out whenever a road's pre-determined roughness level is reached, without being optimally set (Paterson, 1987); 2) the age distribution of roads over the life cycle is uniformly distributed; 3) there is no traffic growth, i.e., traffic flow is constant over time; and 4) all road damage is caused solely by vehicles, *then* the external road damage cost is *nil* on average (Newbery, 1988a, Proposition 1). The variable road maintenance user charge component, applied on a per ESAL basis, will fully cover the average cost of repair (Newbery, 1988a, Proposition 2). When viewed diagrammatically, this result is analogous to charging for maintenance cost in another dimension, as opposed to charging solely for the congestive effect of automobiles on a per passenger car equivalent (PCE) basis in Figure 4 of Part I (Hau, 2005a). Intuitively, the damaging force of an additional ESAL has the external effect of raising the vehicle operating costs

---

example, a 24 ton truck whose weight is distributed evenly among 2 axles would cause more than 3 times as much damage as the same truck whose weight is distributed evenly among 3 axles: $(2/3) \cdot (12/8)^4 = 3.375$.

[17] The example assumes that more of the weight -- 8 tons -- is distributed on the rear axle of the light truck whereas a fully loaded automobile has its weight uniformly distributed among the two axles. Small and Zhang (1988) and Small et al. (1989, Chapter 2 Appendix) have disputed the fourth-power law and performed a statistical analysis of AASHO's road test data to show that the equivalence factor for an axle load rises steeply to the third power, thereby proposing a "third-power law" instead. Whether the power term is actually 3, 3.5 or 4 (as respectively claimed by Small, Paterson, or engineers conventionally) is not as crucial as the significant maintenance cost savings to be made if vehicle damage is efficiently charged for. Conceivably, an increase in the power term would enlarge the small deficit that Small et al. obtained by charging for congestion and road damage costs. I argue that the charging of *all* externalities including environmental and accident costs would very likely close any remaining gap.

[18] Based on empirical evidence for the whole road network of Tunisia, Newbery et al. (1988, Table 7, 23 and pp. 58-59) and Newbery (1988c, Table 1) demonstrate that the external urban congestion costs alone account for the "overwhelming fraction" (about nine-tenths or more) of total external costs of road usage (excluding environmental and accident costs) for automobiles and utilities. For heavy vehicles, the reverse appears to apply in both Tunisia and Ghana, with road damage costs dominating instead (see Gronau (1991) also). (Heggie and Fon (1991, Annex 1) argue that some of Newbery's calculations of congestion costs for Tunisia are overestimates.) Congestion costs constitute a large share of total road costs for both the United Kingdom (nine-tenths) and the United States (one-fifth) in the long run, whereas road damage costs are about 3 to 3.5% in the United Kingdom and only 2% in the United States (Newbery (1988a), Small et al. (1989), Chapter 6). Pollution costs appear to account for less than one-tenth of total road costs in the United States, whereas accident costs are on the same order as external congestion costs for the United Kingdom. (Newbery, 1988b, 1990).

of subsequent vehicles over time, just as the external congestive effect of an additional car has on other vehicles behind it within a traffic stream.

In addition, the indirect impact of accumulating more ESALs is to reach the preset, maximum allowable roughness level of the road earlier than expected, thereby precipitating an overlay and a corresponding lowering of vehicle operating costs. It is unclear *a priori* which of these two magnitudes dominate. Hence Newbery's (1988a, 1989) breakthrough was to prove, in a variety of ways, that the additional vehicle operating costs attributed to road damage externalities just cancel out with the vehicle operating cost savings when averaging over roads of different vintages.

In the general case, the fundamental theorem relaxes Assumption 4) above by allowing for road damages that are independent of use, such as weather-related deterioration. In this case of a condition-responsive maintenance policy, the road damage externality is no longer zero but is quantitatively negligible. While the use-related part of maintenance costs is chargeable because of the road damage cost, the fraction due to weather is not allocable. Hence the presence of weather-related road damage results in only partial recovery of all maintenance costs when only charging by road damage costs. Therefore, in order to close the highway budget deficit, Newbery (1989) needs to price for the marginal external congestion cost within a standard long run framework (see Newbery (1988b) also). That is, under the conditions of constant returns to scale and road use, the efficient congestion toll will yield revenues that cover the capital cost of the highway but *only* the invariant (or non-allocable) portion of road maintenance cost attributable to weather, see Figure 7 of Part I (Hau, 2005a) for definition of short-run fixed cost (Newbery, 1989, Proposition 1). (Following Mohring (1965), we have seen how this is done in our simple investment rule of comparing optimal toll revenues with fixed costs, be they construction or maintenance costs, see Figure 7 of Part I (Hau, 2005a) and the surrounding text.) It then naturally follows that the optimal road user charge, i.e., the optimal congestion toll-cum-road damage charge, will then yield revenues that cover the capital cost of the road and the *total* road maintenance cost in a constant returns world (Newbery, 1989, Proposition 2). In other words, the optimal congestion toll covers both the capital cost of the road and the non-allocable fraction of road maintenance, whereas the allocable fraction of road maintenance is still chargeable to traffic loadings via the average variable road maintenance cost component per ESAL. This explains why the total expenditure on road maintenance is fully recoverable. *If* the same constant returns conditions and assumptions 1) through 4) again apply, and *if* we further accept that 5) heavy vehicles predominantly use the slow lanes and are confined there, *then* the optimal road user charge will recover the capital cost of the highway and *twice* the total maintenance costs of the road (Newbery, 1989, Proposition 3). If all heavy vehicles are confined to the slow lane, then the damage costs are accumulated in a shorter time span than if they were to be spread out evenly over all lanes. This has the effect of raising the maintenance cost of the slow lane and the effective cost of widening the whole road, since all lanes are typically resurfaced together once the performance service index of the road dips below the trigger point. Cost-benefit analysis of lane expansion to reap time savings suggests that the marginal cost of investment in capacity would now have to account for both the increased road maintenance cost for road strengthening *and* the annuitized capital cost on a PCE basis. (This is a slightly modified version of our earlier optimal investment rule.) Moreover, heavy vehicles are charged for the traffic loadings they create. While stronger roads are cheaper to maintain, investment in road strengthening is costly, resulting in costlier upkeep and higher capital cost. The larger fixed costs translate themselves, in the long run, into a requirement for higher congestion

charges, which help contribute to an overall budget surplus for the road authority, despite "massive" increasing returns to road strengthening (Newbery, 1989). Another way of looking at this result is that the congestion toll effectively covers the road's entire capital cost as before, whereas the maintenance cost is recovered twice -- once via the variable road maintenance user charge component applied on a PCE basis and the second time around via heavy vehicles per ESAL.

The most serious drawback in the propositions above is the condition of constant returns to scale.[19] We have reviewed the empirical evidence and find that there are increasing as well as decreasing returns to scale operating on different parts of the road network, meaning that deficits as well as surpluses would most likely co-exist.

6. VARIABILITY OF ROAD THICKNESS

We are now in a position to formally relax the implicitly used assumption of a given road thickness and to incorporate the latest work into our model. In *Road Work: A New Highway Pricing and Investment Policy*, Small et al. provided a comprehensive exposition of their technical extension of Mohring and Harwitz's long-run result on highway economics with optimal durability, within a standard neoclassical welfare maximization framework.[20] Instead of taking current highway design standards as given, another characteristic of a road -- its thickness -- can be varied. A thicker pavement would serve to withstand the damaging power of trucks more and thereby prolong the life of a road. Despite the fact that there are tremendous economies associated with road durability, the additional strengthening of the road would still substantially increase the total investment cost of an overlay. Using the standard optimization technique of consumer's surplus and producer's surplus maximization, the intuition of the three allocation rules derived are again based on the simple notion of setting marginal benefit equals to marginal cost. That is:

1) *The first rule -- the optimal pricing rule* -- says that the traveler should undertake a trip only up to the point where the incremental benefit just offsets the incremental cost to the community. A vehicle's entry into a transport corridor results in two effects: the congestive effect which depends on the number of PCEs, and the damaging effect based on the number of ESALs. Thus our optimal pricing rule derived earlier is adjusted to include an extra component -- a road damage charge -- which balances out any short-fall.
2) *The second rule -- the optimal capacity rule* -- says that, with a condition-responsive maintenance strategy built in, optimal investment for highway capacity is then to

[19] Newbery (1988a) regards the lack of the pursuit of a condition-responsive maintenance strategy as the most serious limitation. Further, Newbery (1989) cites Keeler and Small's (1977) result of constant returns to scale, which is based on a sample of San Francisco Bay Area roads. Newbery (1990) also cites Kraus (1981a) as evidence supporting constant returns when Kraus demonstrates slightly increasing returns (see previous section). Based on an earlier version of Newbery (1989), Heggie and Fon (1991) take issue with many of Newbery's assumptions.

[20] Newbery (1989), for example, derives optimal durability (or strength, rather) and confines the extent of his analysis principally to the constant returns to scale world. By contrast, Small et al. (1989, Chapter 6) explicitly explore the case of increasing returns to scale of road construction, as well as durability, and present simulation results of an urban expressway and arterial -- with sensitivity analysis -- within a long-run equilibrium framework. The first order conditions of their optimization problem yield the optimal pricing, investment and durability rules.

expand the width of the road up to the point where the marginal cost of capacity *plus* maintenance cost is equivalent to the resultant time savings.

3) *The third rule -- the optimal durability rule* -- says that the road is optimally strengthened by investing up to the point where the marginal cost of durability just equals the savings in vehicle operating costs to other motorists *plus* the associated savings in maintenance costs due to a thicker pavement.

The two latter rules could be regarded as an investment rule extended into two different dimensions. Thus the three rules of optimal pricing, optimal capacity and optimal durability yield efficient prices and trips, as well as the optimal number of lanes and inches of pavement thickness.

There is evidence that there exist economies associated with road strengthening (Small et al., 1989, Chapters 2-3; Winston, 1991). By reanalyzing AASHO's road test data within an economic optimization framework, Small and Winston report that the optimal thickness for a rigid pavement is an inch and a half higher than the current ten inch standard, which follows AASHO's guidelines. The remarkable finding is that a mere increase in thickness of 15% would lead to a doubling of pavement life from 13 to 26 years (Small and Winston, 1986, 1988; Small and Zhang, 1988). The logical implication of massive economies to increasing road thickness -- a fact long known to highway engineers -- is losses for the road authority. We therefore ask: is there a theoretical reason that would allow us to still adhere to the impeccable marginal cost pricing principle and yet achieve the goal of cost recovery? To answer this question we turn to the notion of diseconomies of scope.

6.1 Economies of scope vs. diseconomies of scope

The condition of constant, increasing or decreasing returns to scale yields the respective result of break even, loss or profit, respectively (see Section 5 and Figure 2). The empirical findings of returns to scale evaluated earlier suggest that decreasing returns to scale in road construction in urban areas may perhaps offset the increasing returns to scale in rural areas, but that the highway budget may still be in deficit due to economies of road strengthening. The answer can only be established using econometric analysis. Even if we were to accept Keeler and Small's (1977) careful finding of constant returns, would it not seem plausible to argue that the pavement deficit due to significant economies to road durability necessarily yield an overall deficit for the road authority? The answer is no.

Up till now, we have confined ourselves to the single product world of traffic volume. But road transport involves two products: namely traffic volume and loadings, which carries us into the literature of multiproduct industries and returns (Bailey and Friedlaender, 1982). The issue is whether a multiproduct firm can jointly manufacture the various products cheaper than if each firm were to produce an output separately. If it is cheaper to combine operations and share in the joint costs, there are economies of scope. The notion of diseconomies of scope can best be grasped by illustrating the design of a railway track (Kim, 1987). A railroad track that is built to withstand the axle loadings of freight would need greater strength and thickness, which conflict with the requirement of securing a smooth ride for train passengers. Ensuring both thickness and smoothness of tracks is more costly, resulting in diseconomies of scope. The analogy carries over to roads, where it would cost more to produce a highway that is thick enough to handle heavy vehicles and yet wide enough to accommodate the considerably larger number of automobiles. Thus there are no economies of scope even though one

product -- the number of loadings -- is clearly subject to increasing returns to durability, whereas the other product -- the traffic volume -- is potentially subject to increasing returns to scale due to road construction. In other words, it is the other characteristic of production -- the scope -- that would tip the budget balance away from a deficit (Small et al., 1989, Chapter 6).

Automobiles cause virtually no road damages compared with heavy vehicles but trucks are fewer in number and therefore cause lesser amount of congestion. Hence one could separate, roughly speaking, traffic flow from loadings by identifying automobiles with the former and trucks with the latter. Traffic flow, in turn, requires road capacity whereas loadings require road strengthening. The empirical finding of "modest but significant" diseconomies of scope by Small et al. (1989, Chapter 6) of about 6 to 10% tips the product-specific economies of scale closer to one -- the constant returns case. (The multi-product economies of scale range from 1.00 to 1.06.) Their bottom line simulation results of an urban expressway and arterial and sensitivity analysis demonstrates that a budget balance and hence cost recovery are achievable. The shortfall of a few percent of total road costs can still be recovered by maintaining some level of first registration fees, annual licenses and/or fuel taxes. Finally, since *all* externalities ought to be internalized in principle, air, noise pollution and accident costs should also be appropriately charged for (Carbajo, 1990; Cameron, 1991). In this way, the highway budget would most likely make a profit (see footnote 18).

The basic intuition behind this remarkable result is as follows: because most roads are currently built to accommodate both automobiles and heavy vehicles, a neat dichotomy of allocating pavement wear costs between automobiles and trucks cannot be achieved. Thus 1) the marginal cost pricing of traffic flow requires a congestion charge -- which effectively covers the capital cost of investment in the long run -- and 2) the marginal cost pricing of pavement wear -- associated with the investment cost of strengthening the road -- results in a relatively steep road damage charge. The marginal cost-based road user charges combine to yield the double-charging of roads, depicting diseconomies of scope.[21] One logical implication of the diseconomies of scope argument is that savings could be reaped from building a thinner autos-only road system (the savings of which are estimated to be on the order of 23% by Keeler and Small (1977)). With a universal road system, roads should then be built or resurfaced durably on the slow lane only and heavy vehicles should be confined to that lane. (This experiment is being carried out in California and Florida (Small et al., 1989, pp. 15).

7. SUMMARY AND CONCLUSIONS

One of the earliest contributions to the economic analysis of road pricing was from a French engineer, Jules Dupuit (1844). He was the one (and not Alfred Marshall) who introduced the concept of consumer's surplus -- the cornerstone of the welfare analysis of any public project -- and brought it to bear on the subject of toll roads. It is a similar tack that this paper has taken, in the belief that a picture would perhaps speak a thousand words. I have synthesized the dominant thinking to date on the topic of road pricing by the main protagonists and integrated them into a consolidated, analytical framework.

[21] The theoretical finding of the double-charging of roads is also derived by Newbery (1989) for the case of constant returns to scale and road use. However, he neither cites nor employs specifically the concept of multiproduct returns in his work.

According to my analysis of the traditional road pricing arguments, it is hardly surprising that road pricing as advanced in the past encountered its share of difficulties. This is because the congestion toll has the effect of a tax (increase) on trip-makers, despite the fact that it is an externality-corrective tax. The gist of the argument is as follows: people are against congestion pricing because: 1) those who are tolled would face a higher price relative to a no tax situation on average;[22] 2) those who are priced off the road in order to circumvent paying the toll are clearly worse off as a result of the 'forced' switch onto a different mode or time of day; and 3) the other road users who are not tolled -- the tolled on -- are no better off and, indeed, may even be worse off if congestion is encountered. My applied welfare analysis actually takes into account the increase in revenue to the government and the issue of transfer payment. If each of the parties is separated out as: 1) the tolled, 2) the tolled off, 3) the tolled on, and 4) the government, the group that stands to gain the most is the government (and the untolled -- the rest of society), unless the toll revenues are earmarked. The other party that is primarily better off are those with very high values of time. Only in the hypercongestion case could *all* groups be made better off on average.

It has been argued that the disposition of the revenues of externality corrective (toll-) taxes should accrue to the public treasury (Baumol and Oates, 1988, pp. 29).[23] Conventional cost-benefit analysis treats a dollar as a dollar to whomsoever it accrues, and also implicitly assumes that only consumers derive satisfaction from revenues. Hence, *unless* toll revenues are channeled back through reduced transportation related taxes, user charges or improved public services, neither the tolled nor the tolled off would endorse road pricing.

In this paper, I have also shown step-by-step how to implement short-run marginal cost pricing in transport following Walters and others. In particular, I have established that implementing the *optimal pricing rule* -- the first rule -- is equivalent to setting an optimal road user charge, where: 1) a congestion toll on the *difference* between the marginal cost and the average variable cost of a trip is imposed, and 2) the maintenance cost of road use is also charged. Further, I have shown that the process of determining an optimally priced and invested road system is similar, albeit with a couple of important differences, to the process of achieving long-run equilibrium of a typical product within a competitive environment, inspired by the basic Mohring-Harwitz model. The issue of short-run vis-a-vis long-run marginal cost pricing is clarified *inter alia*. For instance, implementing short-run marginal cost pricing is equivalent to pursuing long-run marginal cost pricing in a steady-state world in the long run. However, only short-run marginal cost pricing would be able to capture the peak/off-peak nature of travel demand. In the absence of scale economies, the *optimal capacity rule* -- the second rule -- says that the existence of economic profit, i.e., toll revenue collection less the fixed and non-use-related costs of a road, would serve as a surrogate market mechanism indicating that the road ought to be expanded. Putting it another way, the investment rule says that a road ought to be expanded to the point where the additional cost of investment in capacity equals the additional savings in travel time. In the long run, toll revenues would cover the interest on the capital investment, invariate maintenance, depreciation and

[22] Consider a more realistic situation where the existence of a uniform fuel tax effectively translates itself into a trip price that is higher (but not as high as the peak-period price), and exactly the same analysis follows.

[23] This is because a Pigouvian tax is one which imposes a positive price on a producer of an externality and a zero price on a consumer of an externality.

operating costs of the road. Maximizing society's welfare dictates that one should implement short-run marginal cost pricing *over the long run* by varying the size of the highway capital stock. In this way, the pursuit of efficient pricing and a self-financing road system would be compatible with one another and no residual cost need be covered. In one stroke, the same congestion tolling mechanism solves the pricing-cum-investment problem, satisfying the conceptual guidelines of efficiency pricing, economic and financial viability as set out in the introductory section of the first paper (Hau, 2005a).

Relaxing and flushing out major assumptions indicate that the result derived here is robust and applicable to: 1) a multiplicity of roads, 2) a road which is subject to diurnal variation of demand and the peak-load problem, and 3) differences in values of time. If constant returns to scale can be shown to hold on average for a particular city with severe congestion, it could potentially aid in greatly simplifying the planning of highway investment. It could also be used as a yardstick, against which scale economies or diseconomies could be measured.

Economic efficiency would be enhanced if marginal cost pricing of a trip were done in the short run and optimal investment in capacity were pursued over the long run. I have established that, if governmental authorities were to charge correctly for congestion, it is possible for them to make money on a road while satisfying economic efficiency. Profitable roads arise in heavily utilized or urban areas because land rents of real estate are high and congestion tolls reflect the rising opportunity costs. Yet, it is possible that congestion pricing in the presence of both indivisibilities and diseconomies of scale in urban roads may curtail the extent of profitable undertakings. Similarly, pursuing marginal cost pricing under the restrictive conditions of both indivisibilities and scale economies of rural roads could also result in profits in the short run. Thus far the points I have made are based on first principles.

In the long run, the simple pricing and investment model implies that the marginal cost pricing of trips covers all the fixed costs of the road and the congestion toll (which captures the quasi-rent) behaves *as if* it is a capital charge. This analysis also means that an optimally invested road system is in fact one where a road should not always be uncongested during the peak period. An optimally congested road is akin to the commonly accepted notion of an optimal pollution level in the field of environmental externalities. An uncongested road for every time period of the day would suggest that that road is over-invested, either because of indivisibilities or nonmarginal cost pricing. If a road is indeed overbuilt, abandoning or downsizing it in the long run may be unavoidable on narrow cost-benefit criteria. The act of downgrading lightly used roads in order to save on the costs of maintaining higher standards of road pavement is a form of disinvested.[24] Alas, given that almost all existing road systems are non-optimally designed and that costs are considered sunk in the short run, the efficient usage of such a network would still call for marginal cost tolls. Any increase above the road user charge should then be regarded as a 'pure tax element' or surcharge, whose contribution to general revenues should perhaps be made on either fiscal or non-economic grounds.

One may ask: starting off with an overbuilt road system, say, is there a way in which road pricing based on the marginal cost concept can be implemented within an institutional context where severe fiscal constraints on public expenditures prevail? To answer this question requires that we go beyond first principles.

---

[24]The common practice of downgrading roads is consistent with the findings of road deterioration in developing countries (Harral and Faiz, 1988, pp.32).

Extensions a la Newbery-Small-Winston have enriched the basic model developed diagrammatically here by incorporating the fact that heavy vehicles are the cause of road damage. Charging for both the external and variable cost of road damage on a vehicle weight *per axle* basis would help close the deficit that may arise from congestion tolling. Further, a road needs to be strengthened to the point where the additional cost of investing in durability just balances out the incremental savings from maintenance and vehicle operating costs. Thus the third rule -- the *optimal durability rule* -- is born. As a natural extension of pricing for externalities, air, noise pollution and accident costs ought to be charged for. Surely in this way the highway budget would more likely involve profits than losses *if* the issue of cost recovery of the sector cannot be ignored.

Given point estimates (or preferably, functional specifications) of speed-flows, demands and the value of time, one can estimate and simulate some of the analytical results developed here. When combined with the associated optimal pricing and investment rules, the efficient level of prices, user charges, speed, volume-capacity ratios, and trips, as well as the optimal number of lanes and inches of pavement thicknesses can be obtained.

The fact that in the transport context, the consumer-producer is both a willing 'victim' as well as a 'beneficiary' has policy implications. As 'victims' of congestion externalities, perhaps travelers ought to be compensated. Note, however, that Pigouvian toll-tax revenues are not supposed to be used to compensate 'victims' of externalities (Baumol and Oates, 1988, pp. 23). Also, intuitively, 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. Travelers are also 'beneficiaries' of road transport by virtue of their being present on congested roads, and their contributions to the toll revenue component of 'user charges' reveal their willingness to pay. 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'. Similarly, heavy vehicles ought to incur their 'fair' share of hefty pavement wear fees. Combining these plausible arguments and our earlier results of optimal pricing and investment principles, suggests that some form of dedicated funds is perhaps necessary -- either in the form of a road fund or a transport fund -- if road pricing is to gain political acceptance.[25]

---

[25] Developments in electronic toll collection and electronic road pricing in Norway, Sweden and Cambridge (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. (With optimal tolling, however, high purchase taxes and registration/license fees of vehicles ought to be reduced to a level sufficient to cover the administrative and enforcement costs of collection. If the road maintenance cost is constant with respect to the traffic level, an appropriate fuel tax could perhaps be used to approximate usage.)

Indeed, a national survey conducted in England indicates that when people were asked whether they are for or against road pricing, about 57% 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 public transport, 57% of the *same* surveyed population were in favor of road pricing rather than against it (Jones, 1991; Goodwin, 1989). (May and Gardner's (1990) simulation results also confirm the case for an integrated approach to road pricing, and buttress the analytical results presented here.) Using an equilibrium travel demand model of modal choice applied to the San Francisco Bay Area, Small (1983) concludes that congestion pricing combined with the redistribution of toll revenues would result in benefits to *all* income groups.

As an illustration, the Oslo Toll Ring, currently in operation, charges private cars (light vehicles) going into the city center a toll of 10 kronos (approximately US$1.50). Roughly 60% of motorists opt for the subscriber lanes

7.1 The role of a road fund

If a road fund were to be set up, compensation would need to be made sufficiently indirect to satisfy Pareto efficiency. Thus the funds for road construction should be used both to maintain existing roads and to finance new roads, and the profits generated from urban roads could be used to finance the fixed capital cost of *worthwhile* rural roads in a nondistortionary manner.[26] 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 and should be examined on a country by country and case by case basis using econometric analysis.

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, however, deficits from lightly used rural roads (with increasing returns to scale) would demand subsidization from the public 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 priced at a *premium*.[27] These welfare losses and premiums would presumably offset one another if viewed within the same (transport) sector -- with a nominal value of a dollar being treated at its face value -- so that we are back to the case of pure efficiency concerns.

Even if a certain city in a developing country, say, 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. Meeting the requirement that a road network be both large enough, in terms of capacity, and strong enough in terms of pavement thickness, can be quite costly. Scope diseconomies in highways mean that a road network that accommodates both loading and traffic volume found universally is more costly than the sum of an autos-only *and* a tailored trucks-only road system. Hence, the surplus associated with diseconomies of scope offsets the potential deficits 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 recovered 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 feasible.

---

in 1991, which are operated via electronic toll collection, rather than manually-operated toll lanes. About 80% of the earmarked funds are used to finance road construction and 20% for busways, buses and trams. The optional nature of the successful Oslo Toll Ring raises the intriguing idea that congestion tolling should perhaps be implemented on a voluntary basis: with the choice of a combination registration fee/fuel tax based on 'average' usage, or a subscription to the use of electronic devices.

[26] The common sense notion of using the profits from heavily used roads to finance lightly used roads has its intellectual roots in Alfred Marshall's 'tax and bounty' system of pricing:

> "One simple plan would be the levying of a tax by the community on their own incomes, or on the production of goods which obey the law of diminishing return, and devoting the tax to a bounty on the production of those goods with regard to which the law of increasing return acts sharply."
>
> Alfred Marshall, Principles of Economics, 1920, pp. 392

Marshall also observes that it is necessary to consider the cost of administering such a tax-subsidy system.

[27] I am indebted to Sir Alan Walters for this insight.

7.1 The role of a transport fund

Alternatively, taking the transport sector as a whole, a transportation fund ought to be set up.[28] If dedicated funds are set up in this way, indirect 'compensatory' payments can be achieved and would not depart far from 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 encompasses a substantial if not the lion's share of trips undertaken in both newly industrializing and developing countries (Deaton, 1987).[29] Typically, the production of bus services is subject to consumer-side bus route and frequency economies of scale. Hence additional funds in the form of *user-side subsidies* are required to meet the financial shortfall arising from the capital equipment, if bus usage is priced at marginal cost. Road pricing would result in more crowded and inferior public transport services unless bus companies were to offer more bus services (and hence lower generalized prices) as a supply response. Then 'untolled' public transport users or captive riders would be made better off.[30] Here the double charging of automobiles via traffic volume and heavy vehicles via loadings would help to close the deficit gap. Increasing by popular rapid mass transit and light rail systems -- both of which are subject to significant scale economies -- 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 pricing practiced in *all* modes (in contrast to tackling individual, non-optimally priced modes), the urban transportation problem will continue to be pervasive.

Even without dedicated funds, it is essential to pursue efficient pricing and stringent benefit-cost analysis link by link and mode by mode on both a volume and loading dimension. Thereafter, the results can be presented for public scrutiny, thereby improving managerial efficiency and public accountability. The competitive tendering and private provision of certain transport services could also serve to enhance managerial efficiency in the public sector. Issues warranting further investigation include the corporatization of certain transport agencies.

Subject to further research, the idea of setting up a transportation or road fund and the pursuit of marginal cost pricing in all its dimensions would enable us to satisfy the quinpartite principles of the World Bank's general guidelines, as stated at the outset of this paper, 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 a fund passes many of the criteria for a 'good' earmarking arrangement as presented in McCleary (1991). The implementation of marginal cost pricing in both the traffic and loading dimension could be done with the advent of technological breakthroughs in automatic road user charging utilizing automatic vehicle

[28] Using an entirely different model than the one used here, Vickrey (1977) establishes the result that cities should use land rent tax revenues arising from agglomeration economies to finance mass transit and public transportation which are subject to increasing returns. Indeed, he argues forcefully and proves the case that *not* subsidizing these fixed costs would be inefficient.

[29] The provision of public transport mentioned in Section 8 of Part I (Hau, 2005a), assumed to operate under constant returns was used merely as an illustrative convenience but this assumption does not result in loss of generality.

[30] Notably, captive bus passengers would benefit from road pricing if equilibrium transit travel times are lowered.

identification and classification, all of which are subject to remarkable scale economies (Hau, 2006). Alternatively, less powerful road pricing instruments such as area licensing, simple cordon pricing schemes and the monitoring of vehicle and axle loading via weigh-in motion scales can be used.

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

Bailey, E.E. and Friedlaender, A.F. (1982) Market structure and multiproduct industries. *Journal of Economic Literature*, **20**, 1024-1048.

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.

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.

Boiteux, M. (1960) Peak load pricing. *Journal of Business*, **33**, 157-179. Translated from the French article "La tarification des demandes en pointe: application de la theorie de la vente au coût marginal," by H. W. Izzard, as revised by the author, in the Revue Générale de l'Electricité, August 1949, Vol. 58, No. 8, pp. 321-340. Reprinted in Marginal Cost Pricing in Practice, ed. James R. Nelson, Prentice Hall, Inc. Englewood Cliffs, New Jersey, 1964, pp.59 89.

Cameron, M. (1991) Transportation Efficiency: Tackling Southern California's Air Pollution and Congestion. Environmental Defense Fund, Regional Institute of Southern California, March, pp. 1-103.

Carbajo, J. (1990) Accident and Pollution Externalities in a System of Road User Charges. Unpublished Working Paper, December, The World Bank, Washington, D.C., pp. 1-22.

Deaton, A.S. (1987) The demand for personal travel in developing countries. Assisted by Duncan Thomas, Janet Neelin and Nikhilesh Bhattacharya, Discussion Paper No. INU-1, Infrastructure and Urban Development Department, The World Bank, Washington, D.C., pp. 1.1-9.26.

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.

Fitch, L.C. and Associates (1964) *Urban Transportation and Public Policy*. Chandler Publishing Company, San Francisco.

Goodwin, P.B. (1989) The rule of three: a possible solution to the political problem of competing objectives for road pricing. *Traffic Engineering and Control*, **29**, 495-497.

Gronau, R. (1991) Are Ghana's Roads Paying Their Way? Assessing Road Use Cost and User Charges in Ghana. Policy Research and External Affairs Working Paper series, WPS 773. The World Bank, September 1991, pp. 1-44.

Harral, C.G. and Faiz A. (1988) Road Deterioration in Developing Countries: Causes and Remedies. A World Bank Study Policy, with contributions by Esra Bennathan, Graham Smith, and Anil Bhandari, June, The World Bank, Washington, D.C., pp. 1-61.

Hau, T.D. (1989) Road pricing in Hong Kong: a viable proposal. *Built Environment*, **15**, 195-214.

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. (2005a) Economic fundamentals of road pricing: a diagrammatic analysis, Part I – Fundamentals. *Transportmetrica*, **1**, 81-117.

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

Hayutin, A.M. (1984) Scale Economies in Highway Capacity: Empirical Evidence and Policy Implications. Unpublished Ph.D. dissertation, Department of Economics, University of California, Berkeley, California, pp. 1-182. Also available from University Microfilms International, Dissertation Information Service, Ann Arbor, Michigan.

Heggie, I.G. and Fon, V. (1991) Optimal User Charges and Cost Recovery for Roads in Developing Countries. Policy Research Working Paper Series, WPS 780. World Bank, Washington, D.C., pp. 1-50.

Highway Research Board (1962) The AASHTO Road Test: Report 5, Pavement Research. HRB Special Report 61E, National Academy of Sciences, National Research Council, Washington, D.C., pp. 1-352.

Jones, P.M. (1991) Gaining public support for road pricing through a package approach. *Traffic Engineering and Control*, **32**, 194-196.

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.

Kim, H.Y. (1987) Economies of scale and scope in multiproduct firms: evidence from U.S. railroads. *Applied Economics*, **19**, 733-741.

Kraus, M.C. (1981a) Scale economies analysis for urban highway networks. *Journal of Urban Economics*, **9**, 1-22.

Kraus, M.C. (1981b) Indivisibilities, economies of scale, and optimal subsidy policy for freeways. *Land Economics*, **57**, 115-121.

Marshall, A. (1920) *Principles of Economics*. 8th Edition, McMillan Press Limited, London.

May, A.D. and Gardner, K.E. (1990) Transport policy for London in 2001: the case for an integrated approach. *Transportation*, **16**, 257-277.

McCleary, W.A. (1991) The earmarking of government revenue: a review of some World Bank experience. *The World Bank Research Observer*, **6**, 81-104.

Meyer, J.R. and Gómez-Ibáñez J.A. (1981) *Autos, Transit, and Cities*. Harvard University Press, Cambridge, Massachusetts.

Meyer, J.R., Kain, J.F. and Wohl, M. (1965) *The Urban Transportation Problem*. Harvard University Press, Cambridge, Massachusetts.

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. (1970) The peak load problem with increasing returns and pricing constraints. *American Economic Review*, **60**, 693-705.

Mohring, H.D. (1975) Pricing and transportation capacity. In Better Use of Existing Transportation Facilities, Special Report 153, Transportation Research Board, National Research Council, Washington, D.C., pp. 183-195.

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.

Neutze, G.M. (1966) Investment criteria and road pricing. *Manchester School of Economics and Social Studies*, **34**, 63-73.

Newbery, D.M.G. (1988a) Road damage externalities and road user charges. *Econometrica*, **56**, 295-316.

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

Newbery, D.M.G. (1988c) 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.

Paterson, W.D.O. (1987) Road Deterioration and Maintenance Effects: Models for Planning and Management, Vol. III, Highway Design and Maintenance (HDM) Standards, Johns Hopkins University Press for the World Bank, Baltimore, Maryland.

Small, K.A. (1983) The incidence of congestion tolls on urban highways. *Journal of Urban Economics*, **13**, 90-111.

Small, K.A. (1992) *Urban Transportation Economics*. Fundamentals of Pure and Applied Economics 51, Harwood Academic Publishers, Chur, Switzerland.

Small, K.A. and Winston, C.M. (1986) Efficient pricing and investment solutions to highway infrastructure needs. *American Economic Review*, **76**, 165-169.

Small, K.A. and Winston, C.M. (1988) Optimal highway durability. *American Economic Review*, **78**, 560-569.

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.

Small, K.A. and Zhang, F. (1988) A reanalysis of the AASHO road test data: rigid pavements. Unpublished paper, December, pp. 1-33.

Starkie, D.N.M. (1982) Road indivisibilities: some observations. *Journal of Transport Economics and Policy*, **16**, 259-266.

Strotz, R.H. (1964a) Principles of urban transportation pricing. 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. 113-121.

Strotz, R.H. (1964b) Urban transportation parables. In *The Public Economy of Urban Communities*, Julius Margolis (ed.), papers presented at the Second Conference on Urban Public Expenditures held February 21-22, 1964, Resources for the Future, Johns Hopkins University Press, Baltimore, Maryland, 1965, pp. 127-169.

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

Vickrey, W.S. (1965) Pricing as a tool in coordination of local transportation. In Transportation Economics, Proceedings of a Universities-National Bureau Committee

Conference on Transportation Economics, National Bureau of Economic Research, New York, pp. 275-296.

Vickrey, W.S. (1977) The city as a firm. In The Economics of Public Services, Martin S. Feldstein and Robert P. Inman (eds.), Proceedings of a Conference held by the International Economic Association at Turin, Italy, published by the Macmillan Press, Ltd., London, pp. 334-343.

Walters, A.A. (1961) 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. (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.

Winston, C.M. (1991) Efficient transportation infrastructure policy. *Journal of Economic Perspectives*, **5**, 113-127.