TI 2007-068/3
Tinbergen Institute Discussion Paper

Self-Financing Roads

Erik T. Verhoef[1]
Herbert Mohring[2]

[1] *VU University Amsterdam, and Tinbergen Institute;*
[2] *University of Minnesota.*

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

SELF-FINANCING ROADS*

Erik T Verhoef**
Department of Spatial Economics
VU University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
+31-20-5986090
everhoef@feweb.vu.nl

Herbert Mohring
1425 E River Pkwy
Minneapolis, MN 55414-3625
Professor of Economics Emeritus
University of Minnesota
+1-612-332-1462
mohring@umn.edu

Key words: Traffic congestion, Road pricing, Road capacity choice, Road financing
JEL codes: R41, R48, D62

Abstract

Mohring and Harwitz (1962) showed that, under certain conditions, an optimally designed and priced road would generate user toll revenues just sufficient to cover its capital costs. Several scholars subsequently explored the robustness of that finding. This paper briefly summarizes further research on the relationship between congestion-toll revenues and road costs. Despite its transparency, the self-financing theorem can lead to erroneous interpretations. The paper's second part discusses three such possible fallacies. It uses a simple numerical model to investigate them. The model shows that the naïve interpretation of the Mohring-Harwitz rule may lead to substantial welfare losses. These losses are particularly prominent when the difference between capital and investment cost is confused and when balanced-budget constraints are imposed under second-best network conditions. In contrast, losses from imposing a balanced-budget constraint when economies or diseconomies of scale exist are surprisingly small.

* The authors thank Robin Lindsey and two anonymous reviewers for helpful comments on an earlier version of this paper. Any remaining deficiencies are ours.

** Corresponding author. Affiliated to the Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam.

1. Introduction

That maximizing the benefits an economy provides to its members requires setting prices equal to marginal costs is a long accepted economic principle. One cost of a road's use is the external congestion cost that each user imposes on all other users by adding to its level of congestion. Economists interested in transportation have long regarded incorporating congestion costs into road prices as essential to an efficient use of roads (*e.g.*, Pigou, 1920; Walters, 1961).

In 1962, one of us participated in pointing out that, under certain technical conditions (to be spelled out below), an optimally designed and priced road would generate user tolls just sufficient to cover its capital costs in the long run (Mohring and Harwitz, 1962). A number of scholars have explored the robustness of that finding. They asked, "Would optimal toll revenues cover optimal capital costs under a variety of more realistic circumstances?" and, "If optimal toll revenues would not cover optimal capital costs but if roads must be self-supporting, what adjustments in tolls and road design would be required to maximize road benefits given a break-even constraint?"

The first part of this paper briefly summarizes the results of research on the relationship between congestion-toll revenue and road costs. We present the self-financing result in its most basic form, and review some of the extensions that have been discussed in the literature. The self-financing theorem, despite its transparency, easily lends itself to erroneous interpretations. The second part of this paper discusses three such possible fallacies, and develops a simple numerical model to investigate the potential relative welfare losses that may result from them. The model shows that the naïve interpretation of the Mohring-Harwitz rule may lead to substantial welfare losses. These losses are particularly prominent when the difference between capital and investment cost is confused and when balanced-budget constraints are imposed under second-best network conditions (the example presented considers the rather common situation where an unpriced substitute exists). In contrast, losses from imposing a balanced-budget constraint when economies or diseconomies of scale exist are surprisingly small.

2. Congestion tolls and road costs: the simplest case

Household travel choice

Consider a set of identical households that enjoy consuming trips on a given road. However, each household dislikes spending the time (t per trip) required to make the trip. Assume that we can characterize traffic conditions in our period of analysis by a simple travel time function $t(F/K)$, where t is travel time, F (for flow) denotes the number of trips per hour being taken on the road and K gives the road's hourly capacity. Assume that besides travel time, there is only one other price component of trip making for a household, namely a toll τ (if levied). If we denote the value of time by α, the perceived *generalized price* or *full price* of the trip, p, can be written as the sum of the generalized cost c and the toll τ:

$$p = c(F/K) + \tau = \alpha \cdot t(F/K) + \tau \,. \qquad (1)$$

The equilibrium flow will then be such that the marginal benefit – the benefit attached to the final trip added – is equalized to the generalized price: if marginal benefit is higher, more trips will be taken; if it is lower, some trips will be suppressed. The marginal benefit function $MB(F)$ therefore determines the equilibrium demand (measured in flow) as a function of generalized price p. $MB(F)$ is therefore also referred to as the *inverse demand*, $D(F)$: "inverse", because quantity as a function of price, $F(p)$, is expressed as price as a function of quantity, $D(F)$. Aggregate household behaviour can thus be represented by the equilibrium condition:

$$D(F) = c(F/K) + \tau\,. \tag{2}$$

Toll and capacity optimization

A public highway-authority might wonder what the 'best' toll level is. The answer of course depends on the objective chosen. An (economic) efficiency-enhancing objective would be to maximize social surplus (or: net benefits): the difference between aggregate benefits of trip making and the social cost of making these trips possible. With $D(F)$ representing marginal benefits, its integral between 0 and F gives total benefits (per unit of time). The social cost consists of two components. One is the total user cost, $F{\cdot}c()$; being the product of flow and average cost. The other is total capacity cost, which we assume will depend on capacity K only and that will be written as $C_K(K)$. It is to be interpreted as a per-unit-of-time cost, so it should include capital and depreciation; it is not the investment cost. The highway-authority's optimization problem thus reads:

$$\begin{aligned} &\underset{F,K}{\text{Max}}\, S = \int_0^F D(x)\mathrm{d}x - F\cdot c(F/K) - C_K(K) \\ &\text{s.t.: } D(F) - c(F/K) - \tau = 0. \end{aligned} \tag{3}$$

The first-order condition with respect to flow F shows that it is optimal to equate marginal benefit $D(F)$ to the marginal social cost of a trip, which is the sum of the private cost $c(\cdot)$ incurred by the individual, and the marginal external cost $F{\cdot}\partial c(\cdot)/\partial F$ that a road user imposes on fellow road users due to congestion:

$$\frac{\partial S}{\partial F} = D(F) - c(\cdot) - F\cdot\frac{\partial c(\cdot)}{\partial F} = 0 \quad\Rightarrow\quad \tau = F\cdot\frac{\partial c(\cdot)}{\partial F}\,. \tag{4}$$

The second expression in (4) follows from substitution of the equilibrium constraint in (3), and shows that optimal road use requires imposition of the so-called *Pigouvian* toll, which is equal to the marginal external cost as just defined. The total toll revenues R are then:

$$R = F^2\cdot\frac{\partial c(\cdot)}{\partial F}\,. \tag{5}$$

The first-order condition of (3) with respect to K tells us to expand capacity up to the point where the marginal benefits of doing so (*i.e.*, the value of aggregate travel time saving) is equal to the marginal cost of capacity:

$$\frac{\partial S}{\partial K} = -F \cdot \frac{\partial c(\cdot)}{\partial K} - \frac{\mathrm{d}C_K}{\mathrm{d}K} = 0 . \tag{6}$$

Note that, because capacity only enters the two cost components of equation (3), cost minimization (by setting K for a given F) is directly implied by the maximization of social surplus. With the elasticity of capital cost with respect to capacity κ defined as follows:

$$\kappa = \frac{\mathrm{d}C_K}{\mathrm{d}K} \cdot \frac{K}{C_K} \tag{7}$$

we can rewrite (6) as follows:

$$-F \cdot \frac{\partial c(\cdot)}{\partial K} = \kappa \cdot \frac{C_K}{K} . \tag{8}$$

As a brief side-step, we note that the quotient rule of differentiation tells us that for any function $c(F/K)$, the following holds true (this is, in fact, an application of Euler's Theorem):

$$F \cdot \frac{\partial c(\cdot)}{\partial F} = -K \cdot \frac{\partial c(\cdot)}{\partial K} . \tag{9}$$

Bringing K to the left-hand side in (8) and substitution of (9) and then (5) finally produces the following equation:

$$R = \kappa \cdot C_K \quad \Rightarrow \quad \phi \equiv \frac{R}{C_K} = \kappa . \tag{10}$$

The first expression in equation (10) tells us that provided an optimal toll is charged (equation (4) applies) and capacity is optimized (equation (6) applies) the per-unit-of-time revenues from optimal pricing R are equal to the per-unit-of-time capacity cost C_K, multiplied by the elasticity κ. Phrased differently, the second expression in (10) states that the *degree of self-financing*, which we define as $\phi \equiv R/C_K$, is equal to the elasticity of capital cost with respect to capacity κ. This is the celebrated self-financing theorem of Mohring and Harwitz (1962, Chapter 2); the special case with neutral scale economies ($\kappa=1$) can be referred to as the "exact" self-financing theorem.

Thus, with neutral scale economies, an optimally priced road designed to minimize the sum of user and provider costs would generate toll revenues that would just cover its provider's costs. Optimally designed and priced roads would thus exactly support themselves. When there are economies of scale in capacity provision ($\kappa<1$), there will be a deficit; with diseconomies ($\kappa>1$) a surplus results. These results are entirely consistent with basic micro economic insights that tell us that a firm that is forced to apply marginal cost pricing will face a deficit under economies of scale, a zero-surplus under neutral scale economies, and a

surplus under diseconomies of scale. Mohring and Harwitz's contribution was to show that this remains true if part of the inputs in the production process (namely, the time invested to make a trip) are user-supplied under congested conditions.

The theorem appears highly relevant for practical policy making. Its application in practice would in the first place imply that the road operator seeks to achieve an efficient road system, in terms of optimal capacity and optimal pricing. Second, application would firmly reduce the need to use tax revenues from other sources for the financing of roads. This may improve efficiency further, because these other taxes are often distortionary. Third, it may help in overcoming problems of public acceptability of road pricing. The resulting scheme is likely to be perceived as 'fair' (only the users of a road pay for its capacity) and 'transparent' (there are no 'hidden' transfers surrounding the financing of roads). Finally, it may lead to improved transparency in political decisions on infrastructure expansion. It is easily demonstrated that if the neutral-economies-of-scale assumption is fulfilled, and other external costs are optimally priced, road capacity should be expanded when short-run optimal congestion pricing yields revenues per unit of capacity that exceed the unit (capital) cost of capacity.[1] The market would thus indicate whether or not expansion is socially warranted, which will generally help improving the transparency and credibility of cost-benefit analyses.

Trouble

But there are also problems. For example: roads are lumpy. They must have an integer number of lanes; π lanes won't do. The capacity of lanes can be varied by changing their widths, altering curves and making them more or less steep, but lanes must be wide enough to allow vehicles to pass. Still, nothing guarantees that the traffic level which satisfies equations (4) and (6) would have the capacity that an integer number of lanes would provide. If not, (6) is not satisfied and the remainder of the analysis breaks down. How big a threat is this to the practical applicability of the theorem? As indicated, because road design affects the capacity per lane, the problem may be somewhat smaller than it seems at first sight (when only thinking of "numbers of lanes"), as long of course as we are beyond capacities of one lane. Moreover, when an operator can pool deficits from oversized roads with surpluses from undersized ones, the relative problem will be smaller for full networks than for individual roads. And, when demand grows steadily over time, one can anticipate alternating periods of deficits and surpluses for individual roads, so that also pooling 'over time' would reduce the relative size of the problem, compared to what might appear from an instantaneous analysis. Nevertheless, as indicated, especially for smaller roads economies of scale may often dominate, and exact self-financing would not be consistent with optimal road design and pricing.

[1] To see why, observe that for a given demand function, both the short-run optimal congestion price (*i.e.* for a given capacity) and the road use per unit of capacity are decreasing in capacity. Short-run optimal toll revenues per unit of capacity therefore exceed the unit cost of capacity with a below-optimal total capacity, and fall short of it with an above-optimal total capacity.

Next, the assumption of neutral scale economies in road construction is essential to the conclusion that, on optimally designed and priced roads, toll revenues exactly cover capital costs. Sadly for the theorem, both rural and urban road construction may have increasing or decreasing returns to scale (*e.g.*, Mohring, 1976; Keeler and Small, 1977; Kraus, 1981; Small, Winston and Evans, 1989), so that exact self-financing need not apply in reality – even if capacity were continuous. The consequences will be explored numerically in Section 3 below; here we briefly address the backgrounds.

A rural road with one 12-foot lane in each direction is commonly regarded as having a capacity of about 2,000 vehicles an hour regardless of their directional division; on such roads, travelers in one lane must wait for both an adequate view of the other lane and a gap in its traffic. With four-lane roads, only a gap in one direction is necessary. Road expansion from two to four lanes therefore increases hourly capacity to about 2,000 vehicles per lane; doubling lanes quadruples road capacity.

Rural and to a lesser extent, urban road geometry may also often involve scale economies. A normal rural expressway has two 12-foot lanes in each direction with wide paved shoulders on each side. The driving lanes themselves account for less than half of its right of way and of the costs of the earth moving required to create it. Three-lanes in each direction would add 50% to its capacity but considerably less than 50% to its capital cost.

At the same time, urban expressways have many more interchanges and overpasses per mile than do their rural counterparts. Doubling the span of a bridge more than doubles its costs. Walls rather than earthen slopes form its boundaries. The excavation economies associated with increased lanes are, therefore, smaller for urban than rural roads and may even turn into diseconomies. Moreover, scale diseconomies could also arise from a rising supply price of urban land, especially in large cities where urban land is scarce (Small, 1999). For all of these reasons, scale economies are considerably smaller and may even turn to diseconomies for urban roads – where capacity expansion is often more relevant – than for rural roads. Small and Verhoef (2007) review a number of studies and conclude "Altogether, the evidence supports the likelihood of mild scale economies for the overall highway network in major cities. Scale economies are probably substantial in smaller cities in which one or two major expressways are important, and may disappear altogether in very large cities where expanding expressways is extraordinarily expensive due to high urban density" (p. 112).

And finally, for exact self-financing to hold, actually two neutral-scale-economies assumptions have to be fulfilled (Mohring and Harwitz, 1962, p. 85-86). One is what Small (1992) called "constant returns to scale in congestion technology": the fact that the travel time function can be written as $t(F/K)$. The other is "neutral scale economies in road construction": $\kappa=1$. In reality, what matters is the combined effect of these two elements: decreasing economies in the one respect can be compensated for by increasing economies in the other. As a matter of fact, units of capacity can always be chosen such that $\kappa=1$ is satisfied, namely by defining a measure of capacity that is proportional to (minimized) capacity cost. But

whether the combined effect implies neutral scale economies is, as just discussed, an empirical question for which the answer seems to vary over place and probably time.

3. Some extensions

The self-financing result from our basic model suggests a very simple and clear relation between infrastructure charging and capacity costs: the degree of self-financing is equal to the elasticity of the capacity cost function. An important question is to what extent this result is a fluke, resulting from specific simplifying assumptions in the basic model, and to what extent it carries over to more elaborate settings. This section will consider a number of complications that were ignored above, but that will be relevant in practical applications. Our discussion will follow and sometimes draw from reviews as given by Lindsey and Verhoef (2000), De Palma and Lindsey (2005), and Small and Verhoef (2007).

Growing traffic

As economies grow and population increases, so, too, do the demands for road space. Continual infinitesimal expansion of a road would be intolerably expensive. Standard practice is to expand capacity to a level greater than that which would be optimal for a steady-state traffic level at the time expansion takes place. Traffic then grows to and then above the level which would be optimal for the expanded road's capacity. At some point, further expansion becomes in order. Consider a road authority in a growing economy that wants to design and to price its network so as to maximize the present value of its future user benefits minus user and road-authority costs. Setting marginal-cost congestion tolls would be an essential part of this optimization process. An interesting question then arises: as with roads in a steady-state economy, would such congestion tolls exactly cover the network's capital costs in a growing economy given constant returns to scale in road production? Arnott and Kraus (1998a) address this question. They find that the self-financing theorem remains valid in present value terms, provided the size of capacity additions is optimized conditional on the timing of investments. This is true whether or not capacity is added continuously or intermittently, and whether or not the timing of investments is optimal.

Heterogeneous users

The same authors address heterogeneity across users and find that, as long as every user faces an optimal charge, this does not undermine the self-financing theorem (Arnott and Kraus, 1998b). The important pre-condition is that marginal cost pricing applies to all users: when not all users are charged or when charges deviate from marginal cost pricing for other reasons (so that (4) above does not apply), the self-financing result generally breaks down.

Time-of-day dynamics

One of the more disturbing simplifications of the basic model in Section 2 is that congestion is assumed to be a static, stationary-state phenomenon. This is helpful in keeping our

discussion transparent, but rather unrealistic when looking at real-world traffic congestion. It is therefore important to verify whether the self-financing result remains intact when taking the time patterns of congestion and optimal congestion tolls into account. Arnott and Kraus (1998a) have shown that this is indeed the case, provided tolls can be varied optimally over time. A specific example of this result has been given for the so-called bottleneck model, first introduced by Vickrey (1969), and later analyzed in greater depth by Arnott, de Palma and Lindsey (1993).[2]

Network extensions

The self-financing result also continues to hold when extending the analysis from a single road or bottleneck to a full network. Yang and Meng (2002) show that self-financing will hold for every individual link in an optimally priced network, and therefore also for the network at large. As we shall see in Section 4 below, network effects do lead to a breakdown of the self-financing result if other parts of the network are *not* optimally priced.

Further extensions

Various other extensions have been considered in the literature.

Newbery (1989) for example considered self-financing in the face of durability choice and maintenance cost, and concluded that "if there are constant returns to scale in roads construction (for roads of given strength), and if there are strictly constant returns to road use (in the sense that heavy vehicles distribute themselves uniformly over road width), then the optimal road user charge (congestion charge plus road damage charge) will recover all road costs (maintenance and interest on capital)" (Newbery, 1989, p. 167).

Small (1999) considered variable input prices, relevant for urban land that may rise in price when demand for road construction increases. This matter makes explicit the distinction between "returns to scale" (a property of a production function) and "economies of scale" (a property of a cost function). Small shows that the sign of actual profits from highway operation under first-best marginal cost pricing will then still be determined by the degree of scale of the cost function (which differs from the degree of returns to scale of production with a rising supply curve for land). The critical condition for exact self-financing under marginal cost pricing thus involves the degree of economies of scale of the cost function, and not the degree of returns to scale of its underlying production function.

Conclusion

In general, we find that extensions to the simple model of Section 2 lead to important additional insights, but generally do not undermine the self-financing theorem as summarized in equation (10).

---

[2] Arnott and Kraus (1998a) consider growth in demand over calendar time. Arnott, de Palma and Lindsey (1993) consider systematic fluctuations in demand by time of day. Demand is intertemporally substitutable in Arnott, de Palma and Lindsey (1993), but not Arnott and Kraus (1998a).

4. Some fallacies in the interpretation of self-financing road infrastructure

As the foregoing illustrated, the Mohring-Harwitz theorem is a strong result, with important policy implications, that extends to various more realistic instances than the case for which it is typically illustrated in textbooks. Practical application would not only result in the use of optimal investment and pricing rules, but – provided the appropriate technical conditions are approximately fulfilled – also to a balanced budget for road operations, which in turn might have political and social advantages related to transparency and perceived fairness. At the same time, the theorem lends itself to fallacious interpretation. In this section, we will highlight three plausible mistakes that a public operator can make in interpreting the theorem, and we will assess the potential (welfare) implications of such misinterpretation using a small numerical example. The analysis bears resemblance to studies into the use of naïve cost-benefit investment rules for road infrastructure, as reviewed in, for example, Small and Verhoef (2007). We will study, in that order, (1) the case where the regulator mistakenly assumes the theorem to imply that under neutral scale economies all toll revenues should be reinvested in capacity (a mixing up of capital costs with investment costs); (2) the case where the regulator imposes a balanced-budget restriction when there are increasing or decreasing scale economies in capacity; and (3) the case where the regulator imposes a balanced-budget restriction when second-best pricing is appropriate due to unpriced congestion elsewhere in the network. We start with a brief discussion of the numerical model that we will be using.

4.1 A numerical model

We use a numerical model that considers static congestion for homogeneous travelers between a single origin-destination pair connected by a single road (at least in the first two applications of the model). Demand is iso-elastic, with the elasticity with respect to generalized price *p* equal to η, and the associated inverse demand function is:

$$D(F) = \delta \cdot F^{1/\eta}, \tag{11}$$

where *D* is marginal willingness-to-pay, *F* is traffic flow, and δ a parameter.

The generalized price *p* is again the sum of average cost *c* and a toll τ, where *c* is now specified according to the widely used BPR (Bureau of Public Roads) function:

$$p = c + \tau = \alpha \cdot t_f \cdot \left(1 + \beta \cdot \left(\frac{F}{K}\right)^{\chi}\right) + \tau, \tag{12}$$

where α is the value of time, t_f the free-flow travel time, *K* the road's capacity, and β and χ are parameters. Note that *c* only contains time costs.

We ignore road maintenance and depreciation. The capital cost is also iso-elastic and is given by:

$$C_K = \gamma \cdot \frac{K_0}{(K_0)^{\kappa}} \cdot K^{\kappa}, \tag{13}$$

where κ is the elasticity of capital cost with respect to capacity, and γ the average unit price of capacity evaluated at a base-level of capacity K_0 (note that the middle term consists of parameters only and could therefore easily be avoided by redefining γ; it is included only for ease of calibration).

Total benefit can be determined as the area below the inverse demand function, so that social surplus S, our measure for welfare, can be written as:

$$S = \underbrace{\delta \cdot \int_0^F x^{1/\eta} \mathrm{d}x}_{User\ benefit} - \underbrace{F \cdot \left(\alpha \cdot t_f \cdot \left(1 + \beta \cdot \left(\frac{F}{K} \right)^{\chi} \right) \right)}_{(Variable)\ user\ cost} - \underbrace{\gamma \cdot \frac{K_0}{(K_0)^{\kappa}} \cdot K^{\kappa}}_{(Fixed)\ capital\ cost}. \tag{14}$$

We choose the following parameters. The BPR parameters β and χ are set equal to 0.15 and 4, respectively; their conventional values. The free-flow travel time t_f is set at 0.5, so we consider a 60 km road if the speed limit is 120 km/hr. The base capital cost elasticity κ is 1. We seek a representative unit price of capacity γ for a three lane (one-directional) highway, which we assume to involve K_0=4500, so that a conventional traffic lane would correspond to K=1500. This implies a doubling of travel times at a use level of around 2400 vehicles per lane per hour. This is roughly in accordance to the flow at which, empirically, travel times double for a single highway lane and the maximum flow on a lane is reached (*e.g.* Small, 1992, Fig. 3.4, p. 66). A maximum flow, however, is not defined for BPR functions. The average unit price of capacity at capacity level K_0, γ, is set equal to 7 (all monetary costs are in Euros). With a unit of time of one hour, this parameter ought to reflect the hourly capital costs. To derive a value from empirical construction cost estimates, an assumption has to be made on whether the model aims to represent stationary traffic conditions throughout a day, or during peak hours only. Our parameterization concerns the latter. The value of 7 was then derived by dividing the estimated average yearly capital cost of one highway lane kilometre in The Netherlands (€ 0.2 million)[3] by 1100 (220 working days[4] times 5 peak hours per working day; assuming two peaks) and next by 1500 (the number of units of capacity corresponding with a standard highway lane), and finally multiplying by 60 (the number of kilometres corresponding with a free-flow travel time of half an hour). We set the value of time α at 7.5, in line with the "official" Dutch value. On the demand side, we use an elasticity η of –0.35. To create a reasonable reference equilibrium, where demand F is such that the travel time is twice the free-flow travel time t_f for the base capacity of K=4500, we finally set δ=7.97·10^{11}.

[3] With an infinitely-lived highway without maintenance and an interest rate of 4%, this implies construction costs of € 5 mln per lane-km, or € 8 mln per lane-mile. This order of magnitude is well in line with figures presented in Litman (2006) for the US, who quotes widely diverging estimates that suggest that the median construction cost per lane mile would be in the range of $ 5 – 10 mln, while more than a third would exceed $ 10 mln.

[4] A rule of thumb in The Netherlands is that there are some 220 regular work days per year (44 weeks) on which "normal" travel conditions occur. 8 weeks are much quieter because of holidays, Christmas breaks, *etc.*

	K	τ	F	D (= p)	c	C_K	ω
Equilibrium	4 500	0	7 231	7.5	7.5	31 500	0
Optimum	5 085	5.58	6 380	10.72	5.14	35 593	1

Table 1: Numerical model: base equilibrium and optimum

For this parameterization, Table 1 shows the base equilibrium as well as optimal levels of the most relevant endogenous variables. Most of this table's content is self-explanatory. The final column, though, gives the efficiency measure ω that we will use. It is for a certain equilibrium defined as the surplus gain in that equilibrium compared to the base equilibrium, divided by the surplus gain in the first-best optimum compared to the base equilibrium. The indicator is therefore naturally 0 in the base equilibrium, and 1 in the optimum.

4.2 Naïve interpretation I: mixing up capital cost and investment cost

The first fallacy we consider concerns the mixing up of capital cost ("To what extent do the yearly toll revenues cover yearly interest cost?") with investment cost ("To what extent should we reinvest toll revenues in additional road capacity?"). This mistake may seem terribly naïve to the trained economist, but may in fact not be so far-fetched in the practice of policy making, where investments are financed from public funds that are raised through taxation, and no interest is paid (at least not directly) on capital invested in public infrastructure. In fact some current proposals for road pricing in The Netherlands contain the qualification that toll revenues be used for road investments.

In a neutral-scale-economies environment, where optimal roads are exactly self-financing, it would be harmless in our model for overall efficiency to impose the constraint that toll revenues should be used to finance the capital costs. However, it is certainly not harmless to impose that all revenues should be reinvested in additional capacity. The easiest way to see where and how the two principles would diverge is to imagine starting with an optimal road initially, in an otherwise stationary environment. Optimal policies then entail constancy of toll and capacity for all future periods, with revenues covering the constant interest on invested capacity. The naïve policy, in contrast, would use revenues to expand capacity in the next period, so that capacity will grow over time as long as road use and toll are positive.

Our first simulation illustrates the consequences. We assume that the regulator saves up all toll revenues during a year (with no interest revenues) t, and uses these to expand capacity at the beginning of the next year, t+1. We assume that the short-run toll is optimal at each moment, to avoid clouding of results by introducing further inefficiencies from non-optimal pricing. In our calculations, we assume that the construction cost per unit of capacity is 25 times as high as the yearly capital cost, which under our assumptions corresponds with an interest rate of 4%. We start with an optimal road in year 1 and trace the development of key variables over the next 50 years.

Figure 1 displays the results. As expected, capacity (upper-left panel) rises over time as toll revenues continue to be collected. Although the optimal short-run toll (upper-right panel) falls over time as expanding capacity reduces congestion, the BPR function will always produce positive optimal tolls for any flow larger than zero. With capacity set optimal in period 1, it is no surprise that ω equals 1 initially, but falls over time afterwards, as capacity deviates further from the optimal level. Before too long, in year 19, ω falls below zero, indicating that the untolled base equilibrium produces a higher social surplus than the tolled equilibrium with excess capacity. The further drop in ω illustrates how the negative impact of this naïve policy upon social surplus becomes worse, the longer the policy is maintained. Finally, the lower-right panel displays the "correctly" calculated profits, Π (*i.e.*, that use capital cost, not investment cost). With a zero surplus in period 1, and rising capacity costs and falling toll levels afterwards, these profits fall over time, indicating deficits. This confirms the claim in Verhoef and Rouwendal (2004) that under short-run optimal pricing and under neutral scale economies, an above-optimal capacity will produce a deficit.[5] All in all, there are good reasons, based on theory and simulation, to discourage regulators from pursuing this particular type of naïve investment policy.

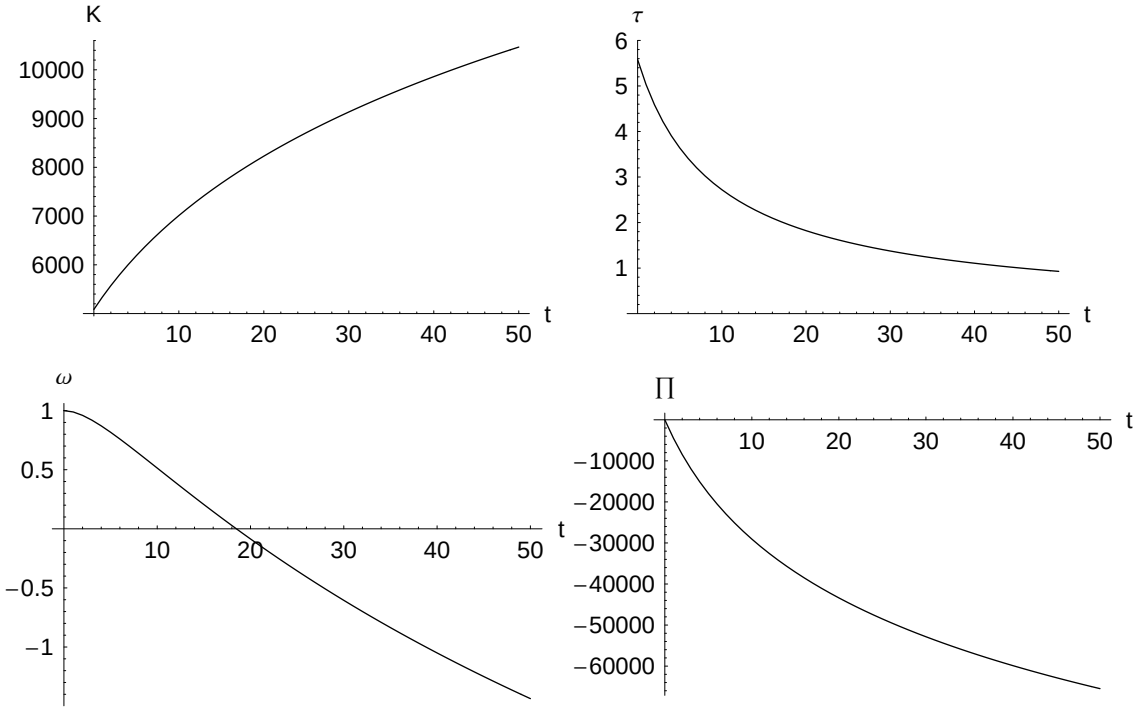

Figure 1: Re-investing all toll revenues: time paths of capacity (upper-left), short-run optimal toll (upper-right), relative efficiency (lower-left) and profit (lower-right)

[5] A below-optimal capacity, not actually considered in Figure 1, produces a surplus. Footnote 1 explains why.

4.3 Naïve interpretation II: imposing self-financing under non-neutral scale economies

A second type of naïve interpretation would start from the political and social advantages that a balanced-budget regime might bring in terms of transparency and perceived fairness, and would strive for balanced budgeting even when the capital cost elasticity κ is unequal to unity. It is a way to impose a hard budget constraint so that costs could be contained. Again this is not an unlikely situation. It may occur whenever the primary motivation for road tolling is the financing of infrastructure, as it seems to have been the case for example for the Norwegian toll rings and for various applications in the US (*e.g.*, Small and Verhoef, 2007, Ch. 4.3). We investigate this situation by tracing the impacts of varying κ upon the model results of interest when a balanced budget is imposed as a constraint.

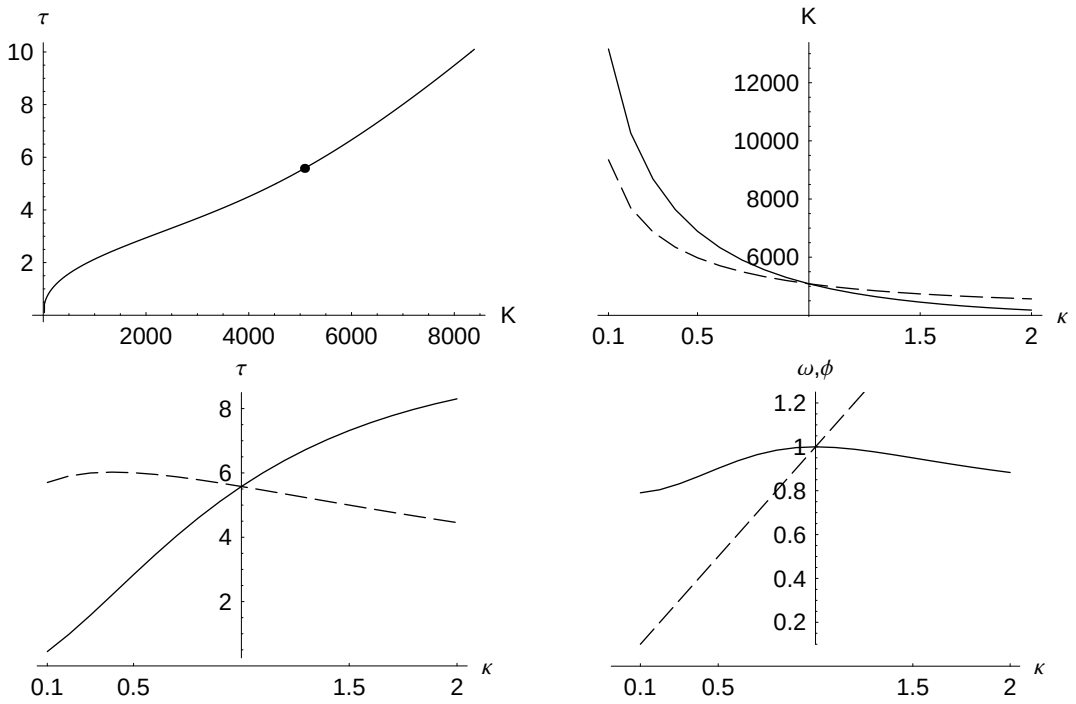

Figure 2: Ignoring capital cost elasticity: zero-profit contour for κ=1 (upper-left), optimal (solid) and second-best (dashed) capacity (upper-right), optimal (solid) and second-best (dashed) toll (lower-left), relative efficiency (solid) and degree of self financing under first-best policies (dashed) (lower-right)

In doing this, we first deal with the question of which zero-profit capacity-toll combination the regulator chooses for a given κ. This combination is namely not uniquely defined. The upper-left panel of Figure 2 illustrates this for the (base) case of κ=1, by showing the zero-profit contour in the K-τ space (the optimum shown in Table 1 is represented by the dot). Note that the iso-elastic demand function with η=–0.35 secures that any capacity can be financed fully as long as the toll is sufficiently high; hence the shape of the contour. We again aim to avoid further distortions from clouding the analysis and assume that, for every κ

unequal to 1, the regulator sets second-best levels for K and τ, so as to maximize social surplus under the constraint that the budget be balanced.

For an elasticity $\kappa<1$, there are economies of scale and we expect a deficit for first-best policies; for an elasticity $\kappa>1$ we expect diseconomies of scale to produce a surplus. For $\kappa<1$, the second-best (zero-profit) capacity K is therefore below the first-best capacity (see the upper-right panel in Figure 2) and the second-best toll is above the first-best toll (lower-left panel). These patterns are reversed for $\kappa>1$, with first-best and second-best tolls and capacities naturally coinciding for neutral scale economies, at $\kappa=1$ in the centre of the diagrams.

The Mohring-Harwitz theorem predicts the degree of self-financing ϕ, defined as the ratio of toll revenues over capacity cost under first-best toll and capacity setting, to be equal to the elasticity of capital cost with respect to capacity κ. The dashed line in the lower-right panel shows ϕ as a function of κ, and confirms that this result is indeed reproduced in our numerical model. A quite different question is how large the efficiency loss would be from imposing self-financing when κ is unequal to 1. The pattern of ω by κ, in the same panel, confirms the intuitive notion that the relative efficiency loss increases with the divergence of κ from 1. But the relative efficiency loss ω is found to be much smaller than the deviation of the degree of self-financing ϕ from 1, reaching values near 0.8 for the two extreme values of κ considered in Figure 2, 0.1 and 2. In other words, whereas the relative deficit or surplus from first-best optimal pricing and capacity setting depends relatively strongly on the capital cost elasticity κ, the relative social 'loss' of maintaining self-financing when κ is unequal to 1 is far less sensitive in our model – provided, of course, self-financing is achieved by setting the second-best toll and capacity, as assumed in Figure 2. Although exact self-financing under first-best policies thus breaks down for $\kappa\neq1$, the social sacrifice to be made for maintaining exact self-financing, if desired for other reasons, may not be too large.

This is a surprising result, and it is important to assess how sensitive it is to the key assumptions in our numerical model. Figure 3 shows that the pattern seems robust with respect to two parameters that warrant particular attention. These are the demand elasticity η, taking on a relatively low absolute level of 0.1 in the upper-left panel and a relatively high absolute value of 0.75 in the upper-right panel (see for example Goodwin, 1992, for a review of demand elasticity estimates); and the power coefficient of the travel time function χ, taking on a low level of 1 in the lower-left panel and a high value of 10 in the lower-right one.[6] The dashed diagonals confirm that the self-financing theorem of equation (10) remains valid independent of the parameterization; the solid lines show that the relative social cost of imposing exact self-financing, provided it is achieved in a second-best way, appears to be limited quite generally.

[6] For the sensitivity analysis for η, the parameter δ was adjusted to obtain the same base-equilibrium (in terms of flow and travel time); for the sensitivity analysis for χ, this was done by adjusting β.

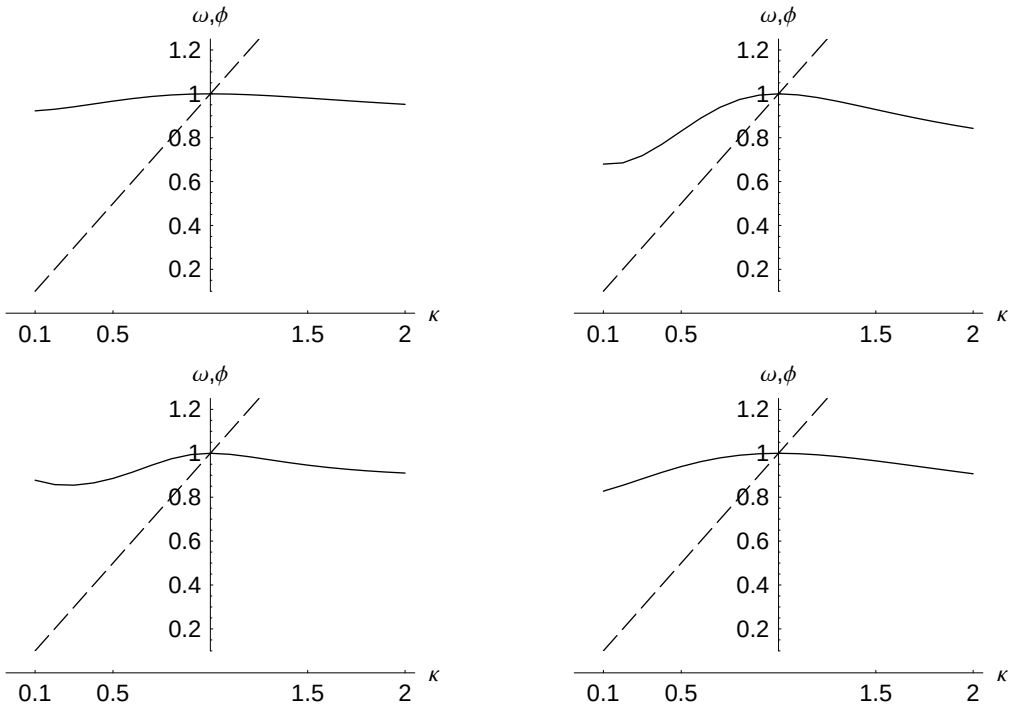

Figure 3: Ignoring capital cost elasticity: sensitivity analysis of relative efficiency (solid) and degree of self financing under first-best policies (dashed) for low demand elasticity (upper-left), high demand elasticity (upper-right), low convexity of travel time function (lower-left), and high convexity of travel time function (lower-right)

4.4 Naïve interpretation III: imposing self-financing under second-best network conditions

A final naïve interpretation we wish to highlight concerns the case where the regulator ignores that the Mohring-Harwitz theorem applies to full networks only when all its links are priced optimally. Actual applications of road pricing that are motivated to finance infrastructure invariably concern situations where not all links of the network are optimally priced, so also this case appears to be relevant in practice. We illustrate the implications for a simple extension of our single-road model, namely one where an unpriced parallel road (denoted *U*) is available, and a toll and capacity can be set only for a substitute tolled road (*T*).

This is a modest extension of the classic two-route problem studied by, *inter alios*, Lévy-Lambert (1968), Verhoef *et al.* (1996), and Liu and McDonald (1998). Important conclusions from these studies are (1) that the second-best toll is below the marginal external cost on the tolled road *T* in order to optimally trade off the toll's positive impact upon congestion on the tolled road *T* against its negative impact upon congestion on road *U* (see also equation (15a) below); and (2) that the efficiency gains from second-best tolling (as measured by ω) will generally be modest.

Verhoef (2007) derives that in the more general case, where both the toll and the capacity for road *T* can be optimized, the second-best toll rule remains the same as in the classic problem (with a fixed capacity) just discussed, while the optimal investment rule

presented in (6) for first-best optimization remains valid for road T also when an unpriced congested alternative is available. Specifically, the second-best optimum requires:

$$\tau^T = F^T \cdot \frac{\partial c^T}{\partial F^T} - F^U \cdot \frac{\partial c^U}{\partial F^U} \cdot \left(\frac{-\frac{\mathrm{d}D}{\mathrm{d}F}}{\frac{\partial c^U}{\partial F^U} - \frac{\mathrm{d}D}{\mathrm{d}F}} \right), \tag{15a}$$

and:

$$-F^T \cdot \frac{\partial c^T}{\partial K^T} - \frac{\mathrm{d}C_K^T}{\mathrm{d}K^T} = 0, \tag{15b}$$

where superscripts denote roads, $F \equiv F^U + F^T$, and $\mathrm{d}D/\mathrm{d}F$ denotes the slope of the (single) inverse demand function.

Because the second-best toll for this particular problem is below the marginal external cost, whereas the optimal investment rule for the toll road is not affected in functional form, the existence of unpriced parallel (congested) capacity causes the self-financing rule to break down. The degree of self-financing will be below the elasticity of capital cost, implying that under neutral economies of scale, a deficit will result for second-best toll and capacity choice (*e.g.*, Verhoef 2007). Imposing a balanced-budget constraint under such circumstances would reduce maximum achievable social surplus to a yet lower level. In our final analyses, we compare the associated second-best/zero-profit results to the "conventional" second-best results (*i.e.*, without a zero-profit constraint) for varying levels of unpriced capacity K_U and assuming, as we did before, that the government makes one naïve misinterpretation only: self-financing is believed to be appropriate, but otherwise K_T and τ_T are set so as to maximize social surplus. The main results are shown in Figure 4.

The lower-right panel shows that the second-best policy indeed produces a deficit when unpriced capacity K_U is greater than zero. These deficits increase in K_U up to the point where K_U equals the second-best capacity that would be chosen in absence of pricing (K_U=5891), a level we shall refer to as K_U^* in what follows. For $K_U > K_U^*$, it is uneconomical to supply additional capacity when it is unpriced. When optimal capacity decreases in toll, as is true in our model but not necessarily in general,[7] it is therefore also uneconomical to supply additional capacity when it is priced. The second-best optimal capacity K_T is then zero, as shown also at $K_U = K_U^* = 5891$ in the upper-left panel of Figure 4, so that the deficit also drops to zero. The upper-right panel of Figure 4 shows that the second-best optimal toll falls when K_U rises (naturally starting at the first-best level for K_U=0 and K_T=5085, the first-best level of capacity), which reflects that spill-overs upon road U become increasingly important as its capacity rises. Relative efficiency, finally, falls from 1 at K_U=0 to around 0.5 at K_U^*, which is under our parameterization the relative efficiency gain that can be achieved from second-best optimal investment without pricing new capacity. The falling pattern between these two

[7] See also Wheaton (1978), Wilson (1983), and d'Ouville and McDonald (1990).

points reflects that efficiency rises monotonously with the size of priced capacity (and falls with the size of unpriced capacity). The continuation beyond K_U^* reflects that efficiency of course falls when K_U further exceeds K_U^*.

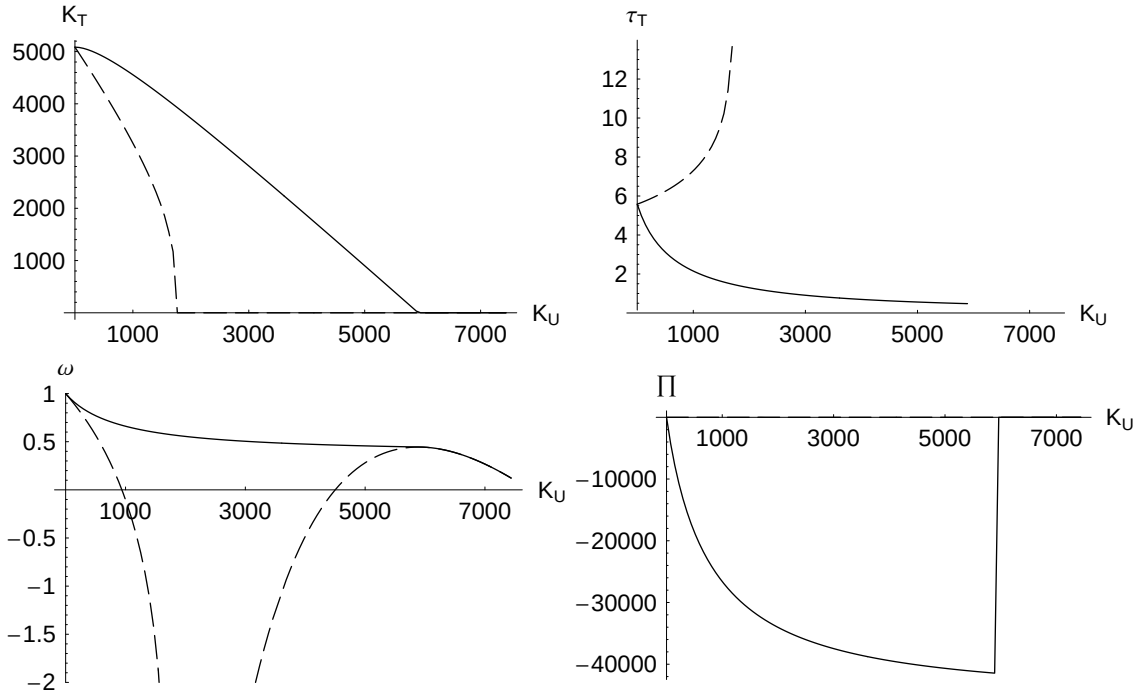

Figure 4: Ignoring network spill-overs: capacity (upper-left), toll (upper-right), relative efficiency (lower-left) and profit (lower-right) for second-best (solid) and second-best/zero-profit (dashed) policies

To meet the zero-profit constraint, the toll should exceed the second-best optimal level, causing second-best/zero-profit capacity to be below the second-best level, as illustrated by the two upper panels. This of course implies a lower level of use on the second-best/zero-profit road than on the second-best road, which in turn causes the maximum level of K_U for which a balanced budget alternative is feasible ($K_U^\#$ in the sequel) to be smaller than K_U^*, the maximum level for which it is efficient to supply additional capacity. The deviations of these capacity and toll levels from the second-best optimal values cause ω to be lower, with negative values certainly not impossible. The right segment of the dashed ω-curve in the lower-left panel considers K_U exceeding $K_U^\#$, and therefore involves no road T being actually offered. The welfare effects underlying the pattern of ω over this range stem solely from the variation in the unpriced capacity K_U. Not surprisingly, then, this segment reaches its maximum at K_U^*, where it is in fact equal to ω for second-best regulation because both schemes involve an unpriced road of capacity K_U^*.

Over a significant range of K_U, therefore, the additional welfare loss from imposing self-financing – over the inherent welfare loss from second-best pricing compared to first-best

tolling – appears to be substantial. The reason is that self-financing requires a relatively high toll, which in turn aggravates the inherent inefficiency of congestion spill-overs upon the unpriced road.

4.4 Conclusion

The numerical model predicts that naïve interpretation of the Mohring-Harwitz rule may lead to substantial welfare losses. These were found in particular for the mixing-up of capital cost with investment cost, and for the imposition of a balanced-budget constraint under second-best network conditions. The losses from the imposition of a balanced-budget constraint when the elasticity of capital cost with respect to capacity is unequal to unity were, in contrast, found to be surprisingly small.

5. Conclusion

After 45 years, the self-financing theorem of Mohring and Harwitz has become one of the landmark results in transport economics, and one that has potentially important implications for real policies – especially now that road pricing appears to become an increasingly realistic option at many locations. This paper reviewed some of the literature showing that the theorem remains valid in more general settings than how it was originally derived and presented. We also showed that a naïve interpretation of the result, unfortunately, may lead to considerable social welfare losses. The economists' advice would therefore be to apply the theorem, but to do so with care.

References

Arnott, R., A. de Palma and R. Lindsey (1993) "A structural model of peak- period congestion: a traffic bottleneck with elastic demand" *American Economic Review*, **83(1)**, 161-179.

Arnott, R. and M. Kraus (1998a) "Self-financing of Congestible Facilities in a Growing Economy". In: D. Pines, E. Sadka and I. Zilcha (eds.) (1998) *Topics in Public Economics: Theoretical and Applied Analysis* Cambridge University Press, Cambridge UK, pp. 161-184.

Arnott, R. and M. Kraus (1998b) "When are anonymous congestion charges consistent with marginal cost pricing?" *Journal of Public Economics* **67** (1) 45-64.

De Palma, A. and C.R. Lindsey (2005) "Relation between pricing, toll revenues and investment". Task Report 2.1, Project REVENUE, DG-TREN Fifth Framework Programme.

Goodwin, P.B. (1992) "A review of new demand elasticities with special reference to short and long run effects of price changes" *Journal of Transport Economics and Policy* **26** 155-169.

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

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

Lévy-Lambert, H. (1968) "Tarification des services à qualité variable: application aux péages de circulation" *Econometrica* **36** 564-574.

Lindsey, C.R. and E.T. Verhoef (2000) "Congestion modelling". In: D.A. Hensher and K.J. Button (eds.) (2000) *Handbook of Transport Modelling, Handbooks in Transport 1* Elsevier / Pergamon, Amsterdam, pp. 353-373.

Litman, T. (2006) "Smart transportation investments: reevaluating the role of highway expansion for improving urban transportation". Unpublished paper, VTPI, Victoria.

Liu, L.N. and J.F. McDonald (1998) "Efficient congestion tolls in the presence of unpriced congestion: a peak and off-peak simulation model" *Journal of Urban Economics* **44** 352-366.

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

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

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

d'Ouville, E.L. and J.F. McDonald (1990) "Optimal road capacity with a suboptimal congestion toll" *Journal of Urban Economics* **28** 34-49.

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

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

Small, K.A. (1999) "Economies of scale and self-financing rules with noncompetitive factor markets" *Journal of Public Economics* **74** 431-450.

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

Small, K.A. and E.T. Verhoef (2007) *The Economics of Urban Transportation* London: Routledge.

Verhoef, E.T. (2007) "Second-best road pricing through highway franchising" *Journal of Urban Economics* **62** 337-361.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1996) "Second-best congestion pricing: the case of an untolled alternative" *Journal of Urban Economics* **40** 279-302.

Verhoef, E.T. and J. Rouwendal (2004) "Pricing, capacity choice and financing in transportation networks" *Journal of Regional Science* **44** (3) 405-435.

Vickrey, W.S. (1969) "Congestion theory and transport investment" *American Economic Review* **59** (Papers and Proceedings) 251-260.

Walters, A.A. (1961) "The theory and measurement of private and social cost of highway congestion" *Econometrica* **29** 676-699.

Wheaton, W.C. (1978) "Price-induced distortions in urban highway investment" *Bell Journal of Economics* **9** 622–632.

Wilson, J.D. (1983) "Optimal road capacity in the presence of unpriced congestion" *Journal of Urban Economics* **13** 337-357.

Yang, H. and Q. Meng (2002) "A note on 'Highway pricing and capacity choice in a road network under a build-operate-transfer scheme'" *Transportation Research* **36A** 659-663.