PRICING, CAPACITY CHOICE AND FINANCING IN TRANSPORTATION NETWORKS

Paper presented at ERSA 2003, Jyväskylä (August 27–30)

Erik T. Verhoef* and Jan Rouwendal**
Department of Spatial Economics
Free University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
Phone: +31-20-4446094
Fax: +31-20-4446004
Email: everhoef@econ.vu.nl

This version: 25/03/03

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

Abstract

This paper explores the interrelations between pricing, capacity choice and financing in transportation networks. It builds on the famous Mohring-Harwitz result on self-financing of optimally designed roads under optimal congestion pricing, and specifically investigates its ins and outs in a network environment and under various types of second-best regulation. The paper develops a small network model, with endogenous car-ownership, in order to study these questions both from an analytical and a numerical viewpoint. It is for instance shown that application of the principle over an entire network may cause user prices to increase more strongly in initially mildly congested areas compared to heavily congested areas, and that a flat kilometre charge, provided accompanied with optimal capacity policies, may result in first-best efficiency gains.

*Affiliated to the Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam. The research of Erik Verhoef has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

**Affiliated to the Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam. Jan Rouwendal is also at Wageningen Agricultural University, Department of Economics.

1. Introduction

Traffic congestion is one of the daily recurring problems resulting from the high car-dependence in most modern societies. The economic approach to analyzing traffic congestion, and to suggesting ways of coping with it through public policy, can be summarized as viewing a congested road as a distorted market, on which travellers demand a service (the use of the road network), supply is defined by the capacity of the road(s) on the network, and where a distortion exists because travel time losses due to congestion constitute an externality: individual users do consider their own travel times when deciding whether or when to use the road, but typically do not take implied travel time losses for other users into account (see, among many others, Small 1992). As travel times on roads are simultaneously determined by – among other factors – the intensity of use and the capacity on offer (*e.g.*, the number of lanes), two archetype policies for coping with congestion can be distinguished immediately in such frameworks. The first involves measures that aim to affect the total levels of road use, or the distribution over time, place, or links, given the network's capacity. The second involves adjustments – usually increases – in capacity levels. Whereas in the practice of policy making, capacity expansion has been the dominant policy for most societies, much of the economic literature has focussed on demand management in general and road pricing in particular (*e.g.* Button and Verhoef, 1998), a choice motivated by the inefficiency of unregulated congested road use, and by the economic insight that when externalities are present, corrective pricing could lead to an improvement in efficiency.

Nevertheless, one of the most famous results in transport economics, due to Mohring and Harwitz (1962), establishes an important relation between demand management and capacity policies: under certain technical conditions, the revenues from optimal congestion pricing will be just sufficient for financing the costs associated with optimal capacity supply. The certain conditions entail that (1) capacity is adjustable in continuous increments, (2) capacity can be expanded at constant marginal cost, (3) trip costs are homogeneous of degree zero in usage and capacity. This 'self-financing' theorem can been shown to extend to full networks instead of single roads, to dynamic models (Arnott, De Palma and Lindsey, 1993), and in present-value terms when adjustment costs and depreciation are allowed (Arnott and Kraus, 1995). Empirical evidence suggests that conditions (2) and (3) may hold at least approximately in a range of circumstances (Small, 1992, 3.4, 3.5). Condition (1) typically does not hold for a single road because the number of lanes is discrete. But capacity can still be varied by widening lanes, or by resurfacing. And at the scale of a road network, capacity may be almost perfectly divisible (Lindsey and Verhoef, 2000). The theorem may therefore be highly relevant for practical policy making. Application of the theorem would in the first place help in achieving an efficient road system, both in terms of capacities and in terms of pricing. It furthermore firmly reduces the need of using tax revenues from other sources for the financing of roads. This may improve efficiency further because these taxes are often distortionary, and may furthermore help in overcoming problems of public acceptability of road pricing because the resulting scheme may be perceived as 'fair' (only the users of a road pay for the 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 can easily be demonstrated that if the technical assumptions are fulfilled, the capacity of a road should be expanded whenever currently optimal congestion pricing yields revenues per unit of capacity that exceeds the unit (capital) cost of capacity.[1] Comparable rules can be formulated for cases of discrete units of capacity – say, lanes – which, however, will always suffer from an 'integer problem'. 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.

One of the reasons why the self-financing principle has not yet been applied in practical policy making, despite these advantages, is undoubtedly that until recently the idea of optimal congestion pricing was much more an academic *curiosum* than a realistic policy option. However, increases in congestion levels, the decreasing scope and increasing costs for further capacity expansions in congested areas, and the development of technologies enabling electronic toll collection, have changed the potential practical viability of congestion pricing in road transport. Not in the least place because of the limited social acceptability of congestion pricing *per se*, and the expected positive impact that the use of revenues exclusively for the financing of roads on which the tolls are charged may have upon acceptability, it seems of great importance to investigate the applicability of the self-financing theorem for practical applications in greater detail.

This paper focuses on three aspects. In the first place, for the theorem to hold, it is necessary that first-best congestion pricing be applied on all links of the road network. Many practical policy proposals, however, foresee implementation of prices only on a limited number of links. Examples include toll cordons and pay-lanes. The implications for pricing *per se* under such second-best circumstances have been explored, for instance, in Verhoef (2002). This paper will address some possible implications of second-best congestion pricing for the applicability of the self-financing theorem. The second-best cases to be addressed include pricing on a sub-set of links only, pricing through undifferentiated kilometre charges, and pricing through fixed 'annual' vehicle taxes. These cases already indicate a second aspect to be explored, connected to the underlying assumption that link-based congestion pricing is the only tax instrument available. We will analyze whether and how insights may change when the regulator has, in addition, other tax instruments at its disposal. One of these is the annual tax, and a second is a flat kilometre tax – which under our simplifying assumptions will be equivalent to a fuel tax. Would application of the theorem take away the need of using these 'traditional' tax instruments; *i.e.*, would their optimal value be zero, or is there a need to take into account the fact that for instance reduced annual taxes may stimulate car-ownership which in turn will have an indirect negative impact upon congestion? What is the relation between revenues from pricing and costs of optimized capacity if only second-best tax instruments are available? How do the insights change when there is a benefit in raising taxes

[1] Too see why, simply observe that for a given demand function, both the short-run optimal congestion price (for a given capacity) and the road use per unit of capacity are both decreasing in the road's total capacity, so that short-run optimal toll revenues per unit of capacity will exceed the cost of a unit of capacity with a below-optimal total capacity.

as such, because other taxes are distortionary, or when there is an exogenous constraint on total tax revenues and capacity cannot freely be adjusted?

The paper is organized as follows. Section 2 below discusses some prior literature. Section 3 introduces the model and discusses the calibration of its parameters for the numerical version. Section 4 reports on the key findings. And finally, Section 5 concludes.

2. Prior literature

Mohring and Harwitz (1962) were the first to establish that, under appropriate conditions set out above, the revenues from optimal congestion tolling are exactly equal to the cost of providing the optimal capacity of the road. They in fact derived a more general result, namely that the ratio between toll revenues and capacity cost equals the elasticity of total capacity costs with respect to capacity. Although the Mohring-Harwitz result is now generally considered as one of the cornerstones of transportation economics,[2] it appears to have gone largely unnoticed at the time. Strotz (1965) reached similar results a few years later.[3] The initial lack of attention to the result may possibly be explained by doubts about its practical significance for transport policy. One may, for instance, doubt whether the cost function is appropriately specified, whether the static model of road congestion is appropriate, whether congestion tolls can be successfully introduced and, if so, if they can be sufficiently varied over time. These problems have received attention later on and below we discuss the main results of later studies.

Keeler and Small (1977) considered the implementation of the Mohring-Harwitz analysis to urban expressways from an empirical perspective. The homogeneity of degree zero of the transport cost function was not assessed empirically, but assumed on the basis of "considerable empirical evidence" (p. 3), and they measured the cost of capacity as the sum of construction cost, land acquisition cost and maintenance cost. They found no evidence of (dis)economies of scale in construction cost. This study therefore confirms Mohring and Harwitz conjecture that the conditions under which the self-financing result holds are plausible. Stahl (1981) developed an analytical model in which capacity cost depends on the number of road users. He states that cost will be exactly recovered under optimal pricing in this framework if the cost function is homogeneous of degree one,[4] but argues that available evidence suggests that these conditions are not satisfied in reality. He points in particular to "evidence of substantial increasing returns to scale in highway pavement thickness" (p. 18). However, Newbery (1988, 1989) reconsidered this issue 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

[2] Arnott and Kraus (1995) regard is as one of the two central results relating to first-best management of congestible facilities in general.

[3] Strotz' analysis was based on Mohring and Harwitz (1962). This publication is in his reference list and he attributes the similarity between his first parable and the Mohring-Harwitz framework to the help provided by Mohring in formulating it.

[4] Stahl (1981) p. 18. Stahl's model also contains external effects. These external effects should be homogeneous of degree zero in the number of road users and capacity in order to have self-financing of capacity under optimal pricing and capacity choice.

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) even if there are substantial economies of scale in road construction" (Newberry, 1989, p. 167). The result requires that there are constant returns to scale in constructing roads with a given strength at different capacities. If this is the case, economies of scale in strengthening roads are unimportant. The optimal policy response to road damage externalities is a flat charge per 'equivalent standard axle load' and Newbery (1988) established a self financing property of this tax that is analogous to the Mohring-Harwitz result with respect to the congestion externality. Small, Winston and Evans (1989) considered issues of congestion and road damage in a simultaneous multiproduct framework and conclude that there are diseconomies of scope, which suggests that it may be worthwhile to consider separation of cars and trucks.

Another aspect of the cost of infrastructure capacity is that the market for land is probably not perfectly competitive. When land has to be bought for construction of a new road, the authority that makes the investment is usually the largest participant in the market. It is therefore probable that the supply curve for land, one of the inputs for the production of road capacity, is increasing. Berechman and Pines (1991) provided a general equilibrium analysis that suggested that in this case the use of 'imputed profits,' i.e. profits that are based on shadow prices that do not take into account the monopsony situation but are solely based on characteristics of the production function, would be appropriate. Small (1999) showed that self-financing is possible under the usual conditions if actual (as opposed to imputed) profits are used.

The Mohring Harwitz result was obtained originally in the context of the static model of road congestion pioneered by Pigou and Knight. In more recent years attention of transport economists shifted somewhat to the bottleneck model developed by Vickrey (1969). Arnott, de Palma and Lindsey (1993) showed that the self-financing result also holds for this type of model when road users are homogeneous. A remarkable aspect of their analysis is that it holds "*independent of the form of the pricing system employed*. If a road system should be self-financing when a sophisticated tolling system is employed, it should also be self-financing when only a flat parking fee is applied" (p. 173, italics in original).

A further issue concerns heterogeneity of road users. Arnott and Kraus (1995) considered the bottleneck model with heterogeneous users and concluded that marginal cost pricing (and the associated self financing property) would still be feasible with an anonymous congestion toll if the heterogeneity concerns unobservable differences (*e.g.* in the value of time) and if tolling is unconstrained (*i.e.* the toll can vary arbitrarily over time). Mohring (1970) showed, in the context of the static model, that under a single toll constraint, implying that the same toll should be charged at peak and off-peak periods, the self-financing result does not hold. Bichsel (2001) showed that the self financing result does not hold if there are two groups of road users that use the road at different times and the toll is restricted to be uniform.

Finally, we wish to mention two studies that are not directly related to the question of self-financing, but that do bear a close relation with the issues studied in this paper. A first study is De Borger (2001), who – like we do – develops a model for analyzing transport

pricing in which car ownership is endogenized, and is explained simultaneously with the demand for trips or kilometres. Although De Borger's model refers to an unspecified externality that could be congestion, he does not incorporate road capacity into his model, and therefore does not consider its financing. Another study that is related to the present paper is Mayeres and Proost (1997), who study optimal tax and public investment rules in the context of an applied general equilibrium model with congestion. However, they do not refer to the Mohring-Harwitz result, and provide no information about the ratio between capacity cost and toll revenues in the optimal situation. One of the interesting aspects of their analysis is that the optimal congestion taxes and congestion levels are almost unaffected by the degree of inequality aversion used by the regulator (p. 277). This suggests that distributional concerns are of minor importance in the design of an optimal policy concerning congestion and road infrastructure.

3. The model

Given the research questions identified in Section 1, the model to be developed should have a number of characteristics. First, it should pertain to a road network of a size larger than a single road, in order to meaningfully consider second-best charges confined to a sub-set of links only. Secondly, in order to study the role of fixed annual vehicle ownership taxes and the possible effects of a switch to a policy regime based on the self-financing theorem, car-ownership should be endogenized, in other to prevent such fixed taxes from appearing a perfect lump-sum tax. Given these requirements, we seek to develop a model as simple and transparent as possible. The numerical version of the model should therefore much more be seen as a mathematical system that allows us to study the main questions in a consistent equilibrium setting, than as an attempt to represent any realistic network – although its calibration deploys some empirical evidence. Section 3.1 will now first describe the analytical modelling framework, and Section 3.2 proceeds with the numerical model.

3.1. The analytical framework

The presentation of the analytical framework is subdivided into the demand side in Section 3.1.1, the supply side in 3.2.2, equilibrium in 3.2.3, and social welfare indicators in 3.2.4.

3.1.1. Demand side

It is assumed that for every origin-destination (OD) pair in the network, there is a set of potential users. A potential user can contribute to the demand for only one OD-pair at most – which is a restrictive assumption in general, but probably more acceptable in the context of commuting: most people have only one residential and one work location. The model does, however, not account for group switching: a user cannot switch to another OD-pair due to policy intervention. An individual's demand function for road use, conditional on car-ownership, is not perfectly inelastic. If the generalized price of road use (including monetized travel times and variable taxes) increases, the number of road trips per unit of time decreases, which may for instance reflect that people would more often take an alternative transport

mode, travel outside the peak considered, or work from home[5] – but not change residential or work location. Individuals within a group are identical with the exception of one characteristic: their relative inclination to road use. This heterogeneity is introduced in order to prevent that car-ownership as a function of policy variables would be constant and positive over a certain range, and constant and equal to zero outside that range.

For simplicity, it is assumed that an individual i's inverse demand function for road use is linear:

$$D_i = \bar{d} - d_i \cdot q_i \tag{1}$$

where subscripts for OD-pairs are suppressed, q_i refers to the equilibrium quantity number of trips demanded, $-d_i$ gives the slope of the individual's inverse demand function which varies across individuals to reflect heterogeneity in terms of the inclination to road use, and $\bar{d}$ gives the intercept which is assumed to be equal across individuals for simplicity: heterogeneity is such that if $d_i = 0.5 \cdot d_j$, individual i consumes twice as many trips as individual j for all generalized cost levels, as long as both own a car. From (1), it is namely straightforward to derive that with a generalized price level for road use, p, the consumer would choose:

$$q_i = \frac{\bar{d} - p}{d_i} \tag{2}$$

and would hence enjoy a 'gross' consumer surplus (not accounting for the costs of vehicle ownership) of:

$$CS_i^G = \tfrac{1}{2} \cdot q_i \cdot \left(\bar{d} - p\right) = \frac{\left(\bar{d} - p\right)^2}{2 \cdot d_i} \tag{3}$$

The price level p thus reflects the price associated with the use of the road conditional on car-ownership. It is assumed to be equal to the sum of the monetized travel time – the travel time t multiplied by the common single value of time vot – plus the congestion charges τ encountered, plus the kilometre charge τ_{km} multiplied by the length of the trip. By setting the unit of distance equal to the distance that can be travelled in one unit of time under uncongested conditions, so that the trip length becomes equal to the value of the free-flow travel time t^{fftt}, we can thus write:

$$p = vot \cdot t + \tau + \tau_{km} \cdot t^{fftt} \tag{4}$$

(note that we define t, τ, and t^{fftt} over the entire trip in (4), and thus ignore route choice and summations over links to avoid notational clutter).

Apart from the 'variable' price p, a travelling individual will incur fixed costs due to the ownership of a car: p_f, which is the sum of per unit of time resource costs c_f and the fixed 'annual' tax on ownership, τ_f ('annual' in quotation marks, because we will calibrate the numerical model so as to describe a single morning peak), which are assumed to be equal across individuals:

$$p_f = c_f + \tau_f \tag{5}$$

[5] All such alternatives are implicitly assumed to be efficiently priced.

It is assumed that individual i will own a car, reflected with the dummy δ_i taking on the value of 1, if the gross consumer surplus from its use, in equation (3), is at least equal to p_f:

$$\delta_i = \begin{cases} 1 \text{ if } CS_i^G \geq p_f \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

It is thus assumed that the vehicle is used exclusively for travelling in the peak and on the network considered. The net consumer surplus for individual i can then be written as:

$$CS_i^N = \begin{cases} CS_i^G - p_f \text{ if } \delta_i = 1 \\ 0 \text{ otherwise} \end{cases} \tag{7}$$

From (6) and (3), the critical value of d_i, d^*, can be derived for which individual i is indifferent between not owning a car, and owning it and using it optimally:

$$d^* = \frac{\left(\bar{d} - p\right)^2}{2 \cdot p_f} \tag{8}$$

Under these assumptions, the existence of a continuous and smooth aggregate inverse demand function requires that users form a continuum defined over d_i. The type of density function assumed for d_i will have potentially significant effects for the results in the numerical model. One way of selecting an appropriate density function recognizes explicitly that each type will imply a different ratio between what can be called the 'short-run aggregate demand elasticity' with respect to p, holding car-ownership fixed, and the 'long-run aggregate demand elasticity', accounting for changes in car ownership. A plausible, constant, ratio of 2 between these two measures is found for a density function:

$$n(d_i) = \nu \cdot \sqrt{d_i} \tag{9}$$

where $n(d_i)$ denotes the density of users with a slope of their individual inverse demand equal to d_i, and ν is a parameter. With this density function, the aggregate demand function can be written as:

$$\begin{aligned} Q &= \int_0^{d^*} \nu \cdot \sqrt{d_i} \cdot \frac{\bar{d} - p}{d_i} \mathrm{d}d_i \\ &= \nu \cdot \left(\bar{d} - p\right) \cdot \left[2 \cdot \sqrt{d^*} - 2 \cdot \sqrt{0}\right] \\ &= 2 \cdot \nu \cdot \frac{\left(\bar{d} - p\right)^2}{\sqrt{2 \cdot p_f}} \end{aligned} \tag{10}$$

implying a long-run elasticity equal to:

$$\varepsilon = -4 \cdot \nu \cdot \frac{\left(\bar{d} - p\right)}{\sqrt{2 \cdot p_f}} \cdot \frac{p}{Q} \tag{11}$$

The perceived 'short-run' aggregate demand function, holding d^* fixed, can be distilled from the middle line in equation (10) and reads:

$$\widetilde{Q} = \nu \cdot \left(\bar{d} - p\right) \cdot 2 \cdot \sqrt{d^*} \tag{12}$$

with d^* treated as fixed, implying a short-run elasticity equal to:

$$\begin{aligned}\tilde{\varepsilon} &= -2 \cdot v \cdot \sqrt{d^*} \cdot \frac{p}{Q} \\ &= -2 \cdot v \cdot \frac{\left(\bar{d} - p\right)}{\sqrt{2 \cdot p_f}} \cdot \frac{p}{Q}\end{aligned} \tag{13}$$

which is, indeed, half the value of ε.

Finally, the inverse aggregate demand function can be found after some manipulation of (10), and reads:

$$D = \bar{d} - \frac{\sqrt[4]{p_f} \cdot \sqrt{Q}}{\sqrt[4]{2} \cdot \sqrt{v}} \tag{14}$$

Note that (10) and (14) fully capture car ownership decisions. The aggregate consumer surplus measures that can be derived from these functions correspond to the net consumer surplus as defined in (7). This is most easily verified by observing that the aggregate equilibrium consumer surplus that can be calculated from $\tilde{Q}$ in (12) – or from its inverse – is a summation, over all users, of the gross surplus in (3). The demand function Q in (10), and its inverse D in (14), only incorporates demand from an individual i below a certain price p_i at which his or her gross consumer surplus is just equal to the total fixed costs p_f. For any equilibrium, the difference between the consumer surpluses that can be calculated from (10) and (12) is therefore exactly equal to the fixed costs incurred by all users using the road in that equilibrium.

3.1.2. Supply side

The model considers static, steady-state congestion on a network, for which the travel time on a link depends on the equilibrium flow on that link alone (there are no direct link interactions), as well as on its capacity. Cost functions are used that are consistent with the technical assumptions underlying the self-financing theorem. For link travel time functions, this is for instance the case with the well-known Bureau of Public Roads (BPR) function, which implies in the notation of the current model that for link l the generalized price reads:

$$p_l = vot \cdot t_l^{fftt} \cdot \left(1 + b \cdot \left(Q_l \cdot cap_l^{-1}\right)^k\right) + \tau_l + t_l^{fftt} \cdot \tau_{km} = c_l + \tau_l + t_l^{fftt} \cdot \tau_{km} \tag{15}$$

where b and k are parameters typically set at 0.15 and 4, respectively; cap_l is a measure for the link's capacity; Q_l gives the equilibrium use level for the link; and c_l is the generalized travel cost for link l. The generalized price is homogeneous of degree zero in use and capacity, as required.

For the cost of providing capacity, constant returns to scale require that the unit price of capacity is constant for a link, so that the total cost for a link can be written as:

$$C_l^{cap} = cap_l \cdot c_l^{cap} \tag{16}$$

3.1.3. Network equilibrium

As customary, the demand and supply side are brought into equilibrium accounting for the network structure typically characterizing road use in reality. We will be deploying a standard deterministic Wardropian equilibrium concept. This means that in equilibrium, users from a

given OD-pair will only use minimum generalized price routes provided the equilibrium generalized price is below the reservation price ($\bar{d}_j$ in equation (14), with *j* denoting OD-pairs), and that there are no routes available with a strictly lower equilibrium generalized price level. A general formal mathematical treatment is suppressed, as it would duplicate standard expositions as given in, among many others, Verhoef (2002), while introducing much notational clutter. A simple case is presented in Section 3.2. It should be noted here, however, that when using inverse aggregate demand functions as in (14) for every OD-pair, the equilibrium obtained simultaneously describes equilibrium use of the network given car ownership, and car ownership given the network equilibrium. Furthermore, it is emphasized that the application of the Wardropian equilibrium principle implies that only one single level of p_j will prevail in equilibrium for an OD-pair *j*. However, the shares of travel time costs and tolls in the equilibrium trip price may of course vary between different used routes for a given OD-pair (*e.g.* when a pay-lane and parallel untolled lanes are simultaneous in use).

3.1.4. Social welfare

For the definition of social welfare, social surplus measures will be used. The structure of the model allows a distinction between a number of welfare indicators, which will be defined in this sub-section. First, consistent with the discussion in Section 3.1.1, one can distinguish for each OD-pair *j* between gross and net consumer surplus:

$$CS_j^G = \int_0^{Q_j} \tilde{D}_j(x) - p_j \mathrm{d}x \tag{17}$$

$$CS_j^N \equiv CS_j^G - N_j \cdot p_f = \int_0^{Q_j} D_j(x) - p_j \mathrm{d}x \tag{18}$$

where $\tilde{D}_j$ is the inverse of the function $\tilde{Q}_j$ as defined in (12), Q_j is to be interpreted as the equilibrium demand, N_j is the number of car owners for OD-pair *j*, and p_f is assumed not to vary over OD-pairs.

When considering the fact that the prices in the model potentially all could incorporate tax payments, equation (18) implies a 'variable' social surplus – meaning that road capital costs are not included – of:

$$S^v = \sum_{j=1}^{J} \left(CS_j^N + \phi \cdot N_j \cdot \tau_f \right) + \sum_{l=1}^{L} \phi \cdot \left(\tau_l + t_l^{fftt} \cdot \tau_{km} \right) \tag{19}$$

where ϕ gives the 'shadow price of public funds' (which is assumed to be exogenous and constant), which is equal to unity when tax revenues are valued equally high as consumer surplus, and may for instance exceed unity when tax revenues are used to reduce other distortive taxes, or be set below unity when these revenues are used in an inefficient manner. Note that a total number of *J* OD-pairs and *L* links is assumed to apply for the network; the total number of potential and active routes – overall, and per OD-pair – need not be specified in this general treatment.

Finally, the overall welfare measure to be employed, *W*, can now be defined as:

$$W = S^v - \phi \cdot \left(\sum_{l=1}^{L} cap_l \cdot c_l^{cap} \right) \tag{20}$$

It is thus assumed that the same shadow price of public funds applies to the capital cost of infrastructure provision, which reflects our assumption that the infrastructure is publicly owned and operated.

3.2. A numerical model

A numerical model will be used to study the research questions identified in Section 1. The model consists of a small static network with 3 links (1–3) and 3 OD-pairs linking three nodes (A-C), and is depicted in Figure 1 below. Travellers for OD-pairs AB and AC have two routes to choose from (choosing either link 1 or 2), while all users for OD-pair BC only use link 3. In all exercises, only equilibria will prevail with both links 1 and 2 used, and with all OD-pairs having a positive demand. An admittedly unrealistic assumption, inherent to the model's static nature but necessary to obtain individuals' vehicle ownership decisions consistent with their behaviour on the network, will be that a vehicle is owned solely to be used on this small network and during the period that is implicitly described by the static model.

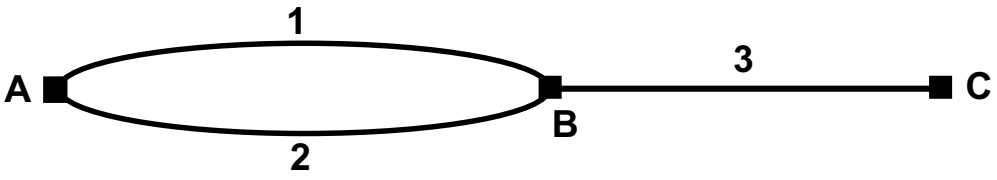


Figure 1. The network used for the numerical model

The equilibrium conditions for such interior equilibria on this simple network read:

$$c_1(Q_{AB}^1 + Q_{AC}^1) + \tau_1 + t_1^{fftt} \cdot \tau_{km} - D_{AB}(Q_{AB}^1 + Q_{AB}^2) = 0 \tag{21a}$$

$$c_2(Q_{AB}^2 + Q_{AC}^2) + \tau_2 + t_2^{fftt} \cdot \tau_{km} - D_{AB}(Q_{AB}^1 + Q_{AB}^2) = 0 \tag{21b}$$

$$\begin{aligned} &c_1(Q_{AB}^1 + Q_{AC}^1) + \tau_1 + t_1^{fftt} \cdot \tau_{km} + c_3(Q_{AC}^1 + Q_{AC}^2 + Q_{BC}) \\ &\quad + \tau_3 + t_3^{fftt} \cdot \tau_{km} - D_{AC}(Q_{AC}^1 + Q_{AC}^2) = 0 \end{aligned} \tag{21c}$$

$$\begin{aligned} &c_2(Q_{AB}^2 + Q_{AC}^2) + \tau_2 + t_2^{fftt} \cdot \tau_{km} + c_3(Q_{AC}^1 + Q_{AC}^2 + Q_{BC}) \\ &\quad + \tau_3 + t_3^{fftt} \cdot \tau_{km} - D_{AC}(Q_{AC}^1 + Q_{AC}^2) = 0 \end{aligned} \tag{21d}$$

$$c_3(Q_{AC}^1 + Q_{AC}^2 + Q_{BC}) + \tau_3 + t_3^{fftt} \cdot \tau_{km} - D_{BC}(Q_{BC}) = 0 \tag{21e}$$

where Q_{AB}^1 identifies the use for OD-pair AB on link 1 (and similarly for OD-pair AC and link 2). An equilibrium for given tax levels is found by solving the above set of equations. Because of the linear dependence in the system, a sixth equation $Q_{AB}^1 / Q_{AC}^1 = Q_{AB}^2 / Q_{AC}^2$ is added to distribute travellers from OD-pairs AB and AC proportionally over links 1 and 2.

Table 1 shows the base-case parameters and policy variables. The demand side parameters were set so as to obtain an equilibrium in which links 1 and 2 are relatively heavily congested, with travel times just exceeding twice the free-flow travel times (see also

Table 2), and link 3 is only mildly congested, while equilibrium short-run demand elasticities are in a plausible range between –0.3 and –0.4. All links have a free-flow travel time of half an hour. A value of *cap* equal to 1750, for the BPR cost function, implies a doubling of travel times at a use level of around 2800 vehicles 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. The latter, however, is not defined for BPR functions.

OD-pair AB	*OD-pair AC*	*OD-pair BC*	*Link 1*	*Link 2*	*Link 3*	*General*
$\bar{d}_{AB} = 30$	$\bar{d}_{AC} = 50$	$\bar{d}_{BC} = 17.5$	$t_1^{fftt} = 0.5$	$t_2^{fftt} = 0.5$	$t_3^{fftt} = 0.5$	$vot = 7.5$
$v_{AB} = 15$	$v_{AC} = 3.5$	$v_{BC} = 17.5$	$c_1^{cap} = 6$	$c_2^{cap} = 6$	$c_3^{cap} = 6$	$c_f = 10$
			$b_1 = 0.15$	$b_2 = 0.15$	$b_3 = 0.15$	$\phi = 1$
			$k_1 = 4$	$k_2 = 4$	$k_3 = 4$	
			$\tau_1 = 0$	$\tau_2 = 0$	$\tau_3 = 0$	$\tau_f = 0$
			$cap_1 = 1750$	$cap_2 = 1750$	$cap_3 = 1750$	$\tau_{km} = 0$

Table 1. Base parameters and policy variables

The hourly unit prices of capacity of € 6 were determined by dividing the estimated average yearly capital cost of one highway lane kilometre in The Netherlands (€ 0.2 million) by 1100 (220 working days times 5 peak hours per working day; assuming two peaks) and next by 1750 (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). The calibration procedure thus implicitly assumes that small changes in capacities induce no effects on travel times outside the peak hours considered in the model, and that off-peak travel (absent from the model) can indeed be fully ignored when optimizing capacities. The BPR parameters *b* and *k* have their conventional values. A value of time of € 7.5 corresponds to average estimates for The Netherlands. The fixed cost of car ownership of € 10 implies, when the car is only used for commuting, that the yearly fixed costs of capital and depreciation would be € 2200, which seems a reasonable order of magnitude for an average car. The shadow price of public funds is set at unity in the base case. Finally, all taxes are set equal to zero for the base-case, 'no toll', equilibrium.

OD-pair AB	*OD-pair AC*	*OD-pair BC*	*Link 1*	*Link 2*	*Link 3*	*General*
$Q_{AB} = 3379$	$Q_{AC} = 2267$	$Q_{BC} = 1345$	$Q_1 = 2823$	$Q_2 = 2823$	$Q_3 = 3612$	$W = 17\,909$
$N_{AB} = 1264$	$N_{AC} = 1438$	$N_{BC} = 294$	$p_1 = 7.56$	$p_2 = 7.56$	$p_3 = 4.39$	$C^v = 58\,616$
$p_{AB} = 7.56$	$p_{AC} = 11.95$	$p_{BC} = 4.39$	$t_1 = 1.01$	$t_2 = 1.01$	$t_3 = 0.59$	$C^f = 29\,954$
$\varepsilon_{AB} = -0.67$	$\varepsilon_{AC} = -0.63$	$\varepsilon_{BC} = -0.67$				$C^{cap} = 42\,000$
						$TR = 0$
						$G = -42\,000$

Note: t_l indicates travel time on link l; C^v the total generalized (variable) travel costs; C^f the total costs of vehicle ownership; C^{cap} the total capacity costs; *TR* total tax revenues; and *G* the government budget

Table 2. Base case: key characteristics

Table 2 summarizes the key characteristics of the base equilibrium. It is clear that the more heavily used links 1 and 2 make up the more congested part of the network (each link has a free-flow travel time of 0.5). Furthermore, the government faces a deficit of € 42 000.

4. Simulation results

This section presents the results of the modelling exercises that were carried out in order to answer the various questions raised in the introduction. The section starts with a brief discussion of the technical approach followed for finding second-best optima.

4.1. An extended algorithm for finding second-best optima in transportation networks

Despite the small scale of the model, it includes eight potential policy variables: three link tolls, a kilometre charge, a fixed vehicle ownership charge, and three link capacities. Most policy regimes addressed below only study a sub-set of these. Nevertheless, the task of finding a second-best optimum can be cumbersome, especially if the number of policy variables available exceeds two. Due to the interactions between the policy variables, optimization for instance through a procedure in which each policy variable is optimized given the level of the other variables will not converge to a second-best optimum after one round, and may thus become very time consuming. At the same time, the model's dimensions are too large to allow for the derivation of insightful analytical optimality conditions that would help in easily finding a second-best optimum. Moreover, these conditions would of course be different for each sub-set of available policy instruments considered.

This problem was handled by developing a general algorithm for finding second-best optima. The algorithm is a generalization of the algorithm presented in Verhoef (2002), who considers the problem of finding second-best optimal toll levels for a congested network under the constraint that not all links can be tolled – which is a special case of the type of second-best problems considered in this paper.

The backbone of the algorithm is formed by the following Lagrangian:

$$\Lambda = W + \sum_{r=1}^{R} \lambda_r \cdot constr_r \tag{22}$$

where r denotes the relevant routes – defined as routes that for an OD-pair j offer the lowest possible generalised price and hence that can be used in the equilibrium – in the network during an iteration (see below), λ_r denote route-specific Lagrangian multipliers, and $constr_r$ is a constraint that equates marginal benefits for the relevant OD-pair to the generalized price for the route considered – as given by the left-hand sides of equations (21) for the current network.

The algorithm can then be summarized as follows:

step 0: set starting values for the available policy variables

step 1: calculate the network equilibrium given the exogenous values of the available policy variables

step 2: solve the system of linear equations that is defined by:

1. $\partial\Lambda/\partial Q_r - \partial\Lambda/\partial\lambda_r = 0$ for all relevant routes r, evaluated in the network equilibrium determined in step 1
2. $\partial\Lambda/\partial\pi_k = 0$ for all policy instruments π that are available for optimization, evaluated in the network equilibrium determined in step 1

in the variables λ_r for all relevant routes and all available policy instruments π_k (note that with R relevant routes and K available instruments, this gives a system of R+K equations in R+K unknowns)

step 3: update the values for the policy instruments by setting π_k equal to a weighted sum of its previous value and the newly predicted value in *step 2*

step 4: check for convergence of all available π_k; terminate if convergence is reached, otherwise return to step 1

This algorithm bypasses the problem that the full problem, requiring a solution not only in terms of all available π_k and relevant λ_r but also in terms of all relevant Q_r, defines a set of $2\cdot$R+K simultaneous non-linear equations in $2\cdot$R+K unknowns, the solution to which generally cannot be found with the mathematical software used. The algorithm was found to perform rather efficiently, although the appropriate weighting procedure for 'old' and 'newly predicted' levels of π_k was not easily determined. A pragmatic trial-and-error approach was employed, where the trade off concerned speed of convergence on the one hand, and instability of the convergence process on the other. Instability was in particular relevant for sets of policy instruments including both taxes and capacities. With weights of 50% for newly predicted tolls, and 5% of newly predicted capacities, most exercises converged within a few minutes on a standard PC. All second-best optima found were subsequently checked by varying all available policy instruments by plus and minus 5%, keeping the other instruments at their predicted second-best optimal level. This yielded lower values for the objective in all cases, and the relative reductions (objective values in all cases exceeding 99.5% of the predicted second-best optimal value) indicated flatness of the objective function around the predicted second-best optimum. Both findings indicate that the algorithm, indeed, performs conform its purpose, and finds second-best optima as required.

4.2. First-best configuration

Mathematically speaking, the first-best optimum is a special case of the broader set of second-best optima, and the algorithm described in Section 4.1 could be used equally well to identify it. (A peculiarity of the first-best optimum is that all multipliers λ_r for the relevant routes will individually be equal to zero; see also Verhoef 2002). Table 3 gives the key characteristics of the first-best optimum. The table shows that we make – at least – one unrealistic assumption for the calculation of the first-best optimum, and that is that capacity can be adjusted as if it were a continuous variable, and what is more, that capacity investments are reversible so that all links indeed can obtain a lower capacity in the first-best optimum than in the base case (with a different base case, the latter assumption of course could have been avoided easily).

We make these assumption because it enables us to indeed consider the overall optimum for the system considered. The results seem to make it worthwhile making these assumptions.

OD-pair AB	*OD-pair AC*	*OD-pair BC*	*Link 1*	*Link 2*	*Link 3*	*General*
Q_{AB} = 80.1%	Q_{AC} = 62.9%	Q_{BC} = 33.5%	Q_1 = 73.2%	Q_2 = 73.2%	Q_3 = 51.9%	W = 188%
N_{AB} = 71.7%	N_{AC} = 49.8%	N_{BC} = 19.3%	cap_1 = 97.0%	cap_2 = 97.0%	cap_3 = 44.0%	C^{cap} = 70.5%
p_{AB} = 131%	p_{AC} = 166%	p_{BC} = 226%	p_1 = 131%	p_2 = 131%	p_3 = 226%	TR = 29 620
			t_1 = 65.9%	t_2 = 65.9%	t_3 = 114%	G = 0.00
			τ_1 = 4.93	τ_2 = 4.93	τ_3 = 4.93	τ_f = 0
						τ_{km} = 0

Note: Percentages denote values relative to the base equilibrium

Table 3. First-best optimum: key characteristics

In the first place, the results confirm that the self-financing theorem carries over to full networks. The government budget equals zero in the optimum, a result that both holds at the level of the full network, and for every individual link (which cannot be verified in Table 3).

Next, in the optimum, two taxes appear redundant. The first is τ_{km}, which is not surprising as any non-zero level could be corrected for by adjusting the link tolls accordingly, and obtaining the same network equilibrium. The other is τ_f. This can be understood by noting that, provided road use is taxed optimally, prospective car owners face the optimal incentive of purchasing a car when confronted with the resource costs c_f. The marginal user will, after paying optimal road taxes for the optimized use level, enjoy a gross surplus that is just equal to the resource cost of owning a vehicle. All other users' ownership adds to social welfare – their gross benefit exceeds the sum of the private and external costs of owning and using the car – whereas all non-users would enjoy a gross surplus from individually optimized use that would fall short of this sum.

Third, the results show that compared to a somewhat arbitrary but not unrealistic base case where congestion is not evenly spread over the network and excess capacity in some areas exist, the implementation of joint pricing an capacity policies may increase the generalized price of transport more strongly in the originally mildly congested areas (link 3 in our model) than in the heavily congested areas (links 1 and 2). The reason is that the optimal capacity reduction is greater. When such capacity reduction is not possible in practice, a comparable result would of course be approached in the long run, as the capacity on link 3 would not be increased for a much longer time, while congestion and optimal tolls would increase over time, if demand over the entire network would grow steadily over time.

Fourth, an interesting feature is that the optimal congestion tax obtains the same value on all links in the network. With the value of time, the length of the links, and cost of capacity units equalized, this result in fact follows immediately from the constant returns to scale assumption. But it is worthwhile emphasizing that the concern over second-best congestion pricing, suffering from the inability of toll differentiation over place, may loose its relevance in the long run – provided capacities are optimized throughout the network, congestion occurs

throughout the network, and values of time and costs of construction are indeed constant over space (see also Section 4.3 below).

A final issue worth addressing is that for all OD-pairs, the relative reduction of car ownership exceeds that of road use. This can be explained by the fact that the users priced off the road will be those that have the highest value of d_i, and hence are those with the smallest inclination to use the road. The sensitivity of car-ownership is of course (much) greater than what would seem realistic at first sight, and it is therefore important to emphasize the main responsible modelling features. One is that the model ignores the possibility that a vehicle could be used for trips other than for the individual's OD-pair and/or outside the peak considered. Another feature is that the time-frame considered is the very long run, in which car-ownership is fully adjusted. This is reflected by the assumption that p_f can be fully saved by terminating car-ownership, and by the fact that no second-hand car market is modelled. Under these two assumptions, any driver priced off the road will indeed give up car-ownership, and because these are the drivers with a relatively low inclination to use the road, the result arises. Finally, the car ownership effects are of course magnified by the unrealistic assumption that in the benchmark equilibrium, all tax levels are zero. If, for instance, column 7 in Table 4 below were taken as the benchmark, with the same road capacities but (optimized) annual taxes in place, a policy change towards first-best regulation would have been found to lead to an increase in total car ownership (of 7%), instead of the significant decrease reported in Table 3. Combined with the initially low congestion level on link 3 – which, given the assumed value of c^{cap} indicates considerable initial overinvestment in its capacity – these factors explain the drastic optimal reductions in car ownership.

4.3. Some second-best policies

The results discussed above demonstrate that under first-best conditions, the self-financing result carries over to full networks, and to situations where car ownership is endogenized – provided prospective car owners make a rational decision on whether or not to own a vehicle. The result has potentially far reaching policy implications, as it opens the way to an efficient and transparent system of simultaneously financing and regulating road networks. However, as discussed in the introduction to this paper, the assumption of first-best policies being feasible may often be considered as rather hypothetical. The question then arises to what extent the result would apply under different types of second-best regulation. Table 4 shows some key results for various policy scenarios that were investigated in order to obtain some further insight into this question. Column (1) repeats the key results for the first-best policies discussed above, and the other columns depict those for what we consider to be the most realistic alternative second-best configurations.

Column (2) presents the results for link-specific tolls while keeping capacities fixed. The optimal toll for link 3 is in that case lower than under first-best policies, which is consistent with the fact that congestion will be less severe because capacity will not be reduced. The fact that the tolls on links 1 and 2 are higher than under first-best policies may then be surprising, as also for these links capacities will be reduced under first-best regulation, albeit slightly. The explanation is that under first-best regulation, the optimal trip price on link

3 becomes much higher, which discourages trips by users from OD-pair AC, and hence reduces congestion on links 1 and 2 compared to the second-best case (2). The one-but-last row shows that with unchanged capacity costs and the implementation of pricing, the government deficit will of course shrink (by 65%). The final row shows that the implementation of pricing alone will lead to an efficiency gain equal to 61% (indicated with the index $\omega = 0.61$) of what can be achieved under first-best regulation.

Column (3) shows the results for the mirror case, where capacities can be adjusted while tolls are kept equal to zero. The associated second-best capacities exceed those in the base-case for links 1 and 2, but falls short of it for link 3, which seems plausible given the relatively heavy initial congestion on links 1 and 2, and the mild congestion on link 3. All capacities exceed the first-best capacities, which is consistent with (but not fully explained by) the observation that when starting from the first-best optimum, a marginal reduction in toll levels will lead to an upward adjustment in optimal capacities, because the positive direct effect (the reduction in travel costs for the users) will then always dominate the negative indirect effect (the negative net social benefits from the induced increase in road use) (*e.g.* Arnott and Yan, 2000). The welfare gain from this policy amounts to 37% of first-best gains, and naturally, the government runs a deficit.

	(1) τ_1, τ_2, τ_3, cap_1, cap_2, cap_3	**(2)** τ_1, τ_2, τ_3	**(3)** cap_1, cap_2, cap_3	**(4)** τ_{km}, cap_1, cap_2, cap_3	**(5)** τ_{km}	**(6)** τ_f, cap_1, cap_2, cap_3	**(7)** τ_f	**(8)** τ_2	**(9)** τ_2, cap_2	**(10)** τ_2, cap_1, cap_2
τ_1	4.93	5.47	0	0	0	0	0	0	0	0
τ_2	4.93	5.47	0	0	0	0	0	2.70	1.47	5.07
τ_3	4.93	1.21	0	0	0	0	0	0	0	0
τ_{km}	0	0	0	9.86	7.48	0	0	0	0	0
τ_f	0	0	0	0	0	8.04	7.27	0	0	0
cap_1	97.0%	100%	134%	97.0%	100%	110%	100%	100%	100%	0%
cap_2	97.0%	100%	134%	97.0%	100%	110%	100%	100%	165%	217%
cap_3	44.0%	100%	66.9%	44.0%	100%	57%	100%	100%	100%	100%
G	0%	35%	100.3%	0%	37%	56%	73%	85%	103%	48%
ω	1	0.61	0.37	1	0.53	0.71	0.37	0.15	0.29	0.60

Notes: Percentages denote values relative to the base equilibrium

ω is the index of relative welfare improvement: the increase in W compared to the base equilibrium, as a fraction of the increase in the first-best optimum

Table 4. Various second-best optima: key characteristics

The relative welfare gains from pricing (2) and capacity adjustments (3), ω=0.61 and ω=0.37 respectively, nearly sum up to unity. This suggests nearly perfect additivity of welfare gains from both policies for the numerical model (as opposed to sub- or super-additivity, where the

gains from joint implementation would be below or above the gains from implementation in isolation). Nevertheless, if capacity choice is irreversible, the nearly perfect additivity does of course not mean that there would be little lost when the policies are implemented sequentially rather than simultaneously. Apart from the obvious fact that a sequential implementation implies sub-optimal welfare gains before both instruments are optimized, the maximum welfare gains from capacity adjustments in isolation require these to be set at levels above optimal levels. It is therefore typically not optimal to set capacities at second-best optimal levels, as indicated in column (3) in Table 4, if subsequent pricing is anticipated. An advantage of starting with pricing would be that prices can be adjusted more easily than capacities – at least from a technical viewpoint.

One of the unanticipated results concerns the second-best policy described in column (4), in which capacities can be adjusted and a flat kilometre charge can be applied. This policy results in a first-best efficiency gain, and the government budget is perfectly balanced. As explained above, with the value of time and unit costs of capacity equalized over the network, speeds and first-best charges per kilometre will become equalized over the network, too – at least as long as speeds are homogeneous of degree zero in use level and capacity. It is indeed easily checked that the value of τ_{km} in column (4) implies link tolls equal to the values of τ_1, τ_2 and τ_3 in column (1) (the length of each link is 0.5). Columns (5) and (2), in contrast, illustrate that if capacities cannot be adjusted and only τ_{km} can be set, efficiency will typically fall short of that under differentiated tolls ($\omega = 0.53$ versus $\omega = 0.61$). A technical or political constraint on differentiation of (per kilometre) tolls thus need not lead to efficiency losses, provided capacities can be optimized.

However, as soon as values of time or capacities are not constant over the network, this equivalence breaks down. An interesting question then becomes how much will be lost by imposing the constraint that the per kilometre charges be equal over the network in such a case. In order to shed some light on this issue, the simulations were re-run for the case where c_3^{cap} was lowered from € 6 to € 3 (the results are not shown in Table 4). Capacity is thus considerably cheaper in the less congested area, which is not unlikely in reality as land prices may be lower. First-best regulation then entails an optimal toll of € 2.83 (as opposed to € 4.93 under base-case parameters) for link 3, with a capacity equal to 67% of the initial value (as opposed to 44%). The optimal tolls for links 1 and 2 of course remain unaltered at € 4.93, but their optimal capacities now increase to 103% of the initial value (as opposed to 97%), due to the increased demand for OD-pair AC following the cost reduction on link 3. The variation in optimal toll levels indicates that a flat kilometre charge in combination with capacity adjustments will no longer succeed in replicating the first-best. However, the relative efficiency of this second-best policy – involving τ_{km} = € 8.10, cap_1 and cap_2 at 107% of the initial values, and cap_3 at 63% – nevertheless realizes 98% of the achievable welfare gains under this adjusted parameterization. Therefore, even with capacity costs varying over a network, an inability to differentiate per-kilometre tolls over a network may induce far smaller efficiency losses than perhaps anticipated if capacities can be adjusted. Moreover,

self-financing may still nearly hold: the government deficit is dwarfed to 3.6% of the initial deficit in this run.

Columns (6) and (7) in Table 4 consider the effects of fixed annual taxes, τ_f, with and without simultaneous capacity adjustments. An intuitive result is that optimal capacities in column (6) are higher than under first-best regulation, which is explained by the absence of user charges and the resulting higher equilibrium use levels. For the same reason, the relative efficiency of τ_f alone in column (7), with ω = 0.37, falls short of that of use charges in columns (2) and (5) (ω = 0.61 and ω = 0.53, respectively). Also with capacity adjustments, in column (6), ω = 0.71 is of course below unity, its value in columns (1) and (4). At the same time, the government deficit remains larger than under use charges. In absence of capacity adjustments, this is the joined result of a lower equilibrium number of car owners and lower average taxes per user (recall that users typically make more than one trip in equilibrium; see *e.g.* Table 2). With capacity adjustments, a third factor is the higher second-best optimal capacity. Both from the perspective of efficiency and government finance therefore, optimized annual taxes appear less attractive than optimized use taxes in our model under the chosen parameterization.

Finally, columns (8)-(10) concern pay-lanes (link 2), with or without capacity adjustments. The relatively low second-best optimal and limited efficiency gains (ω = 0.15) in case the capacity is not adjusted – and therefore fixed at 50% of the total capacity between nodes A and B – is in agreement with earlier findings (*e.g.* Verhoef, Nijkamp and Rietveld, 1996; Verhoef and Small, 1999). It derives from the congestion spill-overs created on the untolled parallel road, which pushes down the optimal second-best toll level and the associated efficiency gains. Column (9) shows that when also the capacity of the pay-lane is optimized, the associated efficiency gains increase. In the numerical example, the capacity is increased by 65%, and ω nearly doubles (to 0.29). The government budget, however, worsens: the tolls collected on the pay lane fall short of the additional capacity costs after optimal expansion. This is of course not necessarily the case: if the initial capacity coincidentally would have been (nearly) optimal, the simultaneous implementation of a second-best toll and optimization of the capacity would have led to an improvement of the government budget (assuming a positive second-best toll; see Verhoef, Nijkamp and Rietveld, 1996, for the possibility of negative second-best tolls for pay-lanes).

A related question is whether full financing of an entire network is possible when only pay-lanes are in place. Our numerical results confirm the intuitive answer that this would generally not be the case, and that substantial deficits are to be expected. The reason is that typically, many users will not pay a toll at all (on the free lanes), and those that do pay a toll are charged below their 'direct' marginal external costs. Nevertheless, full financing may in exceptional cases occur coincidentally. An example could be constructed on the basis of column (9) in Table 4, which shows the final second-best case where both cap_1 and cap_2 can be optimized simultaneously with τ_2. Optimization would lead to a complete removal of link 1, with link 2 expanded to 108.5% of the initial joint capacity of links 1 and 2 together (hence to 217% of link 2's initial capacity). This result in the first place nicely illustrates the

inefficiency of the aforementioned inefficiency of leaving some 'parallel' capacity unpriced: optimality requires the size of this part of the capacity to be reduced to zero, in exchange for increases in priced capacity. As far as financing is concerned: under this regime, the toll revenues on link 2 exceed its capacity costs with 3.5% (the toll exceeds the marginal external congestion costs on link 2 because the external costs of users from OD-pair AC on link 3 are partly recovered through the charge on link 2). This suggests that with a very small link 1 and very low unit capacity costs for link 3, a balanced or even a small positive government budget would theoretically be possible under second-best optimal pricing and capacity choice, provided a certain share of users of unpriced links is present on the pay-lane, and capacity costs for unpriced links are sufficiently low. Clearly, the practical relevance is probably negligible, but the theoretical point is that exact self-financing, or even a budget surplus, on a full network with some unpriced links and second-best optimal tolls and capacities is not by definition impossible.

4.4. Absence of congestion on some roads

For self-financing under optimal regulation of a full network to hold, it is required that in the optimum, congestion exists and hence optimal tolls are positive on all roads that have positive unit capacity costs. In reality, this may not always be true, for instance when the assumption of capacity as a continuous variable becomes binding and some minimum capacity exists (*e.g.* a single lane) for which no congestion occurs in the optimum, while the costs of supplying this capacity are positive. To illustrate the break-down of self-financing under these circumstances, we ran a simulation in which link 3 was assumed to be uncongested by definition, and an arbitrary capacity and associated costs were assigned to this link. The result was that under first-best regulation, as expected, self-financing still holds exactly for links 1 and 2 (with $\tau_1 = \tau_2 = 4.93$ of course still holding). The optimal τ_3 however becomes equal to zero because of the absence of congestion on link 3, yielding a government deficit exactly equal to the assumed capacity costs for the uncongested link 3.

4.5. An above-unity shadow price of public funds ϕ

A second reason why first-best regulation might not lead to exact self-financing, even though the cost functions satisfy the required technical assumptions, would be when the shadow price of public funds ϕ is different from unity. The use of an exogenous ϕ to reflect that tax revenues may be used, for instance, to lower existing distortionary taxes elsewhere in the economy (for instance on labour), can of course be critisized on various grounds: the exact value of ϕ would depend on the question of exactly how the tax revenues are used, its value will in reality not be constant, and partial equilibrium models such as ours typically ignore many aspects that would affect the 'true' value of ϕ, and that may often make it lower than what might be expected on intuitive grounds (relevant mechanisms to consider include the so-called tax-interaction, complementarity, and tax-shifting effects; see *e.g.* Lindsey and Verhoef, 2001). Such shortcomings are important to bear in mind, but do not mean that an exogenous ϕ could not be used to study the impacts of general tax revenue raising objectives upon optimal pricing and capacity choices for road networks.

If ϕ is set above (below) unity, first-best regulation will of course lead to a budget surplus (deficit). A second consequence is that the optimal annual tax τ_f will no longer generally be equal to zero. The reason is that the implied second objective of raising (or avoiding) tax revenues as efficiently as possible would typically require all available taxes to be adjusted, so as to minimize the overall distortions introduced (τ_{km}, however, remains redundant for the same reason as above). The sign of τ_f, however, will not generally be that of $\phi - 1$. An increase in τ_f namely induces two relevant effects. The first is the direct effect of increasing the revenues from those users who remain in possession of a vehicle, which would suggest the sign would be the same. The second is the effect of reducing the tax revenues from annual and use taxes from those users that give up car ownership in response to the higher annual tax. This would work in the opposite direction. Especially because use and annual taxes can be set simultaneously, the latter effect may become important: starting from $\tau_f = 0$ and positive use taxes, a marginally lower annual tax attracts more users, and for instance by raising the use taxes such that the same individual remains the marginal car owner (approximately requiring a marginal raise in use taxes such that his total tax sum remains unchanged) would mean extra revenues from the non-marginal users (who drive more and therefore lose more on additional use taxes than that they gain from the lower annual tax).

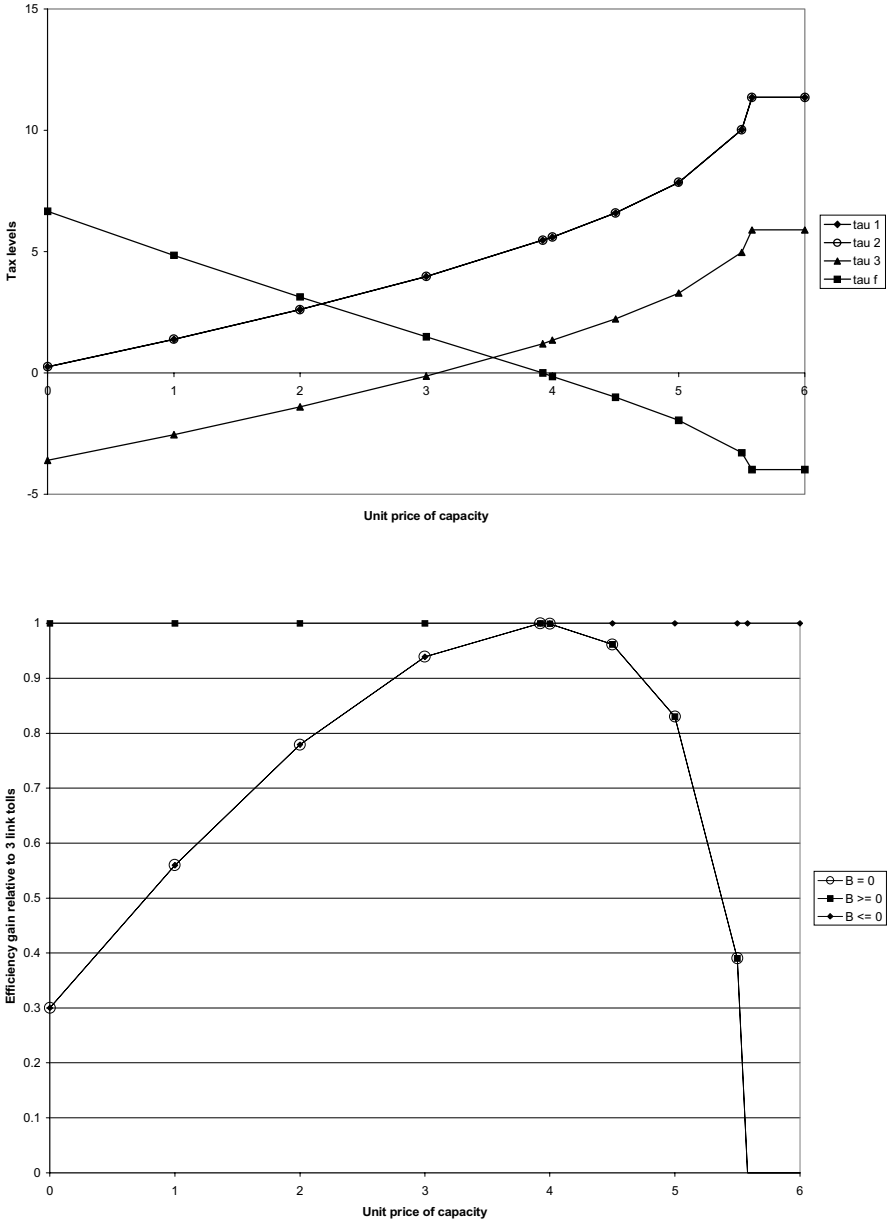
Which of these effects will dominate in reality is of course an empirical matter, largely depending on the elasticity of car-ownership with respect to annual taxes. Our numerical model provides an illustration of the possible dominance of the second effect, which is consistent with the relatively high elasticity of car ownership discussed earlier. At $\phi = 1.1$, we find a modest negative optimal $\tau_f = -0.56$, the budget losses of which are more than compensated for through the optimal use taxes $\tau_1 = \tau_2 = 6.10$ and $\tau_3 = 5.42$ (all use taxes were equal to 4.93 under base parameters).[6] Optimal capacities, in contrast, are smaller than under base parameters (89% of initial capacities for links 1 and 2, 40% for link 3), which is not only consistent with the lower use levels following the higher use taxes, but also reflects that capacity costs, too, are weighed with ϕ, and hence have become higher.

4.6. Budget constraints

Apart from the already mentioned theoretical objections against the use of a shadow price of public funds ϕ as a means to introduce budgeting issues in the analysis, a practical objection could be that government budgets may often be allocated exogenously to the details of the policies implemented. A final point that we would therefore like to address concerns the impacts of an exogenously determined budget, where a balanced budget appears a natural choice, rather than having it resulting endogenously from the optimization of the available instruments. Clearly, a balanced-budget constraint will not be binding in our model if all instruments can be used – otherwise one simply finds the same optimum as discussed in Section 4.2 (given that we set $\phi = 1$). We will therefore consider the case where capacities are given (at their base-case levels), and the regulator can set the use tolls and annual tax so as to

[6] If the demand for road use is sufficiently elastic, it is conceivable that also use taxes would fall if ϕ is increased.

maximize efficiency under a budget constraint. Apart from an equality constraint that the budget be balanced ($B = 0$), we will also consider the inequality constraints $B \geq 0$ where the no deficit is allowed (perhaps motivated by considerations of inter-sectoral fairness), and $B \leq 0$ where no surplus is allowed (for instance motivated by considerations fairness for road users). A convenient way of presenting the main results is by varying c_{cap} such that both cases are covered where an unconstrained optimization through the setting of τ_1, τ_2, τ_3, and τ_f (the levels of which are independent of c_{cap}) would lead to budget surpluses, and where it would lead to deficits. Figure 2 presents the results.



Notes: At $p_{cap} = 3.924$, optimal use of τ_1, τ_2 and τ_3 (and τ_f optimized at 0) results in a zero budget
Beyond $p_{cap} = 5.581$, B = 0 is not attainable even under profit maximizing use of τ_1, τ_2, τ_3 *and* τ_f

Figure 2. Relative efficiency of second-best link tolls τ_1, τ_2 and τ_3 and annual taxes τ_f, for given capacities and under three types of budget constraints

The upper panel shows the optimal levels of the four taxes under the equality constraint, and the lower panel shows the efficiency gain relative to that under unconstrained optimization through the setting of τ_1, τ_2, τ_3, and τ_f (with capacities fixed) as presented in column (2) of Table 4. This index will be indicated with an index $\omega^{\#}$ below, and is calculated by applying $\phi = 1$ to government revenues and expenses.

At $c_{cap} = 3.924$, an unconstrained use of τ_1, τ_2, τ_3, and τ_f happens to result in a balanced budget. In this case, τ_f is optimized at zero, for the same reasons as in the full optimum discussed in Section 4.2. The concave pattern of $\omega^{\#}$, with the maximum at $c_{cap} = 3.924$, clearly shows that the welfare losses of imposing a balanced-budget constraint rise more than proportionally with the (absolute) value of the surplus or deficit that would occur under unconstrained pricing. For higher levels of c_{cap}, additional revenues are required to obtain a balanced budget, which is realized most efficiently in our model by subsidizing vehicle ownership (τ_f is negative), while simultaneously raising the link tolls above their unconstrained optimal levels – both to an extent increasing in c_{cap}. The interpretation is the same as for a shadow price of public funds ϕ exceeding unity, as discussed in Section 4.5 above. As the constraint $B \leq 0$ is not binding in this region, its $\omega^{\#}$ remains equal to unity.

Exactly the opposite results are found for $c_{cap} < 3.924$. What is noteworthy in this region is the fact that the toll on the relatively uncongested link 3, τ_3, becomes negative for sufficiently low values of c_{cap}. Furthermore, a constraint that total tax revenues should be zero, on the left-hand end of Figure 2, does not mean that $\omega^{\#}$ becomes zero. The constraint namely does not mean that all taxes be individually equal to zero, and some efficiency gains remain possible ($\omega^{\#} = 0.3$).

Finally, on the right-hand end of the diagrams, beyond $c_{cap} = 5.581$, a balanced budget is not even attainable under profit maximizing pricing, and the diagrams assume that profit maximizing pricing is in place in this region regardless of the exact value of c_{cap}. Consequently, if capacities are fixed but were set sufficiently far off optimal levels, self-financing may even become infeasible. What can not be read easily from the lower panel is that profit maximizing pricing in fact induces a (small) efficiency loss compared to no pricing: $\omega^{\#}$ has become equal to –0.00013. The closeness to 0 is coincidental.

5. Conclusions

This paper has developed a simple road network model with endogenous car ownership to study various aspects of Mohring-Harwitz (1962) result that will become relevant for a practical application of the principle. A number of findings stand out.

A first one is that optimal per-kilometre congestion tolls and optimal speeds become equal over the network, provided capacities are optimized throughout the network, congestion occurs throughout the network, values of time and costs of construction are constant over the network, and the function that relates travel times to the ratio of use over capacity is the same over the entire network. As a result, a flat kilometre charge in conjunction with optimal capacity policies is capable of reaching the optimum, and concerns over the inability of this

instrument to differentiate over space may become less relevant in the longer run when capacities can be optimized. Clearly, from a dynamic perspective, the ability of toll differentiation over time will remain an important requisite for optimal congestion tolling mechanisms. Our numerical results suggest that, although a flat kilometre charge becomes 'truly' second-best as soon as for instance unit capacity costs vary over the network, the associated welfare losses may remain very small provided capacities (and the kilometre charge) remain optimized: a 50% reduction in unit capacity costs in the lightly congested area resulted in a reduction of only 2% in the efficiency gains from flat kilometre charges compared to differentiated tolls (both with capacities optimized).

As a corollary, a second result of interest is that the implementation of the principle may very well lead to a stronger increase in trip costs as experienced by drivers in initially lower congested areas than in more heavily congested areas. This may be at odds with intuitive expectations, and reflects that an initially low congestion level may often equally well be interpreted as an initial excess capacity.

Thirdly, under second-best pricing on a sub-set of links only, self-financing no longer necessarily occurs. The numerical results however demonstrate that the efficiency from pay-lanes may increase significantly if their capacities are optimized. (This is even more so the case if also the capacities of untolled lanes can be optimized simultaneously, and reductions are possible; however, the consequence is that the untolled lanes are then eliminated, and the term 'pay-lane' would be misplaced.) Furthermore, although it is very likely that a government deficit would result under second-best optimal pay-lane policies, a balanced budget (or even a surplus) cannot be excluded on theoretical grounds. This requires, however, that a certain amount of users from an untolled serial link (upstream or downstream) use the pay-lane.

Next, if roads have some minimum technical capacity (such as the cheapest possible lane), self-financing under optimal pricing over a full network may break down as some roads with positive capacity costs may have optimal tolls equal to zero. And, if capacities cannot be adjusted, imposing a balanced-budget constraint on pricing may lead to substantial efficiency losses, that increase more rapidly as optimal short-run pricing implies larger (absolute) deficits or surpluses.

Finally, unless a non-unitary shadow price of public funds applies, optimal pricing involves a zero fixed ('annual') tax in our model (in which no externalities from car-ownership *per se*, such as through parking, are present). Optimal road pricing provides optimal incentives for car ownership decisions, provided vehicles are optimally priced. However, when also tax revenue raising objectives are relevant, optimal ownership taxes become a relevant instrument, although the objective of raising revenues as efficiently as possible may be served better by a negative annual tax (accompanied with increases in use taxes) than by a positive one.

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 (1995) *Self-financing of Congestible Facilities in a Growing Economy* Department of Economics, Boston College.

Arnott, R. and A. Yan (2000) "The two-mode problem: second-best pricing and capacity" Working paper, Boston College.

Bichsel, R. (2001) "Should Road Users Pay the Full Cost of Road Provision?" *Journal of Urban Economics* **50** 367-383.

Button K.J. and E.T. Verhoef (eds.) (1998) *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* Edward Elgar, Cheltenham.

De Borger, B. (2001) "Discrete choice models and optimal two-part tariffs in the presence of externalities: optimal taxation of cars" *Regional Science and Urban Economics* **31** 471-504.

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.

Lindsey, C.R. and E.T. Verhoef (2001) "Traffic congestion and congestion pricing". In: D.A. Hensher and K.J. Button (eds.) (2000) *Handbook of Transport Systems and Traffic Control, Handbooks in Transport 3* Elsevier / Pergamon, Amsterdam, pp. 77-105.

Mayeres, I. and S. Proost (1997) "Optimal tax and public investment rules for congestion type of externalities" *Scandinavian Journal of Economics* **99** 261-279.

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

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

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

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

Small, K.A., C. Winston and C.A. Evans (1989) *Road Work* Brookings Institution.

Strotz, H. (1965) "Urban transportation parables" in: J. Margolis (ed.) *The Public Economy of Urban Communities* Resources for the Future, pp. 127-169.

Verhoef, E.T. (2002) "Second-best congestion pricing in general networks: heuristic algorithms for finding second-best optimal toll levels and toll points" *Transportation Research* **36B** 707-729.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1996) "Second-best congestion pricing: the case of an untolled alternative" *Journal of Urban Economics* **40** (3) 279-302.

Verhoef, E.T. and K.A. Small (1999) "Product differentiation on roads: second-best congestion pricing with heterogeneity under public and private ownership" Discussion paper TI 99-066/3, Tinbergen Institute, Amsterdam-Rotterdam.