Papers Reg. Sci. 79, 307–332 (2000)

Papers in Regional Science

The implementation of marginal external cost pricing

Long run vs short run and first-best vs second-best

Erik T. Verhoef

Department of Spatial Economics, Free University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands (e-mail: everhoef@econ.vu.nl)

Received: 3 August 1998 / Accepted: 21 June 1999

Abstract. This article discusses a number of issues that will become increasingly important now that the concept of marginal external cost pricing becomes more likely to be implemented as a policy strategy in transport in reality. The first part of the article deals with the long-run efficiency of marginal external cost pricing. It is shown that such prices not only optimize short-run mobility, given the shape and position of the relevant demand and cost curves, but even more importantly, also optimally affect the factors determining the shape and position of these curves in the long run. However, first-best prices are a hypothetical bench-mark only. The second part of the article is therefore concerned with more realistic pricing options. The emphasis is on the derivation of second-best pricing rules. Four types of second-best distortions are considered: distortions on other routes, in other modes, in other economic sectors, and due to government budget constraints.

Key words: externalities, Pigouvian taxation, road transport, second-best

JEL classification: R41, R48, D62

* I would like to thank ERSA's Epainos Prize Jury for awarding me the 1996 Epainos Prize for a paper summarising my PhD-research at the Free University in Amsterdam. I am also grateful to Peter Nijkamp and Piet Rietveld for the superb supervision during my PhD-years, which led to the work summarised in the above-mentioned paper. In consultation with the editor the current article builds upon the prize-winning paper, but at the same time extends beyond it. The work presented in this article was carried out within the EU DG-VII project AFFORD (PL97-2258). Financial support by the EU is gratefully acknowledged. The views expressed in this article, however, represent the author's own, and are not necessarily shared by the EU, nor by the AFFORD consortium. Finally, the useful comments of two anonymous referees on an earlier draft version of this article are appreciated. The usual disclaimer applies.

1 Introduction

Pigou (1920) and Knight (1924) were probably the first to argue that from the viewpoint of economic efficiency, road users should be charged their marginal external costs. This concept of Pigouvian taxation has remained the leading principle in transport economic theory on road traffic externalities regulation (Button and Verhoef 1998). Much to the satisfaction of most transport economists, this principle now also seems to be gaining increasing political support; witness for instance the recent EU-report 'Towards Fair and Efficient Pricing in Transport' (EC, 1995). Now that the practical implementation of Pigouvian pricing principles becomes more likely, it also becomes increasingly important that the economic analysis of externality pricing be extended beyond the limiting boundaries of textbook models. It should deal as well with the complexities that will be encountered when actually applying the general idea in practice.

Certainly, the basic concept of marginal external cost pricing is straightforward: whereever efficient prices appear to be lacking, apply the price mechanism in the same way as it applies elsewhere, by setting appropriate taxes or user fees. When there is high demand, resulting in congestion, charges should be high to deter excessive road use during peak hours. When transport noise affects residential areas more strongly , higher charges should give a stronger incentive to reduce mobility levels, to drive at different times, on different routes, or to use more quiet cars, and so forth. Simple as this general idea may seem, the practical application may often be complicated. Even if we ignore more general implementation problems, such as the limited social and political acceptability and the technical feasibility of marginal external cost pricing, it can be expected that in reality, most of the implicit assumptions underlying the standard economic analysis, leading to the basic Pigouvian tax rule, will not be met. Instead, second-best situations are likely to be the rule rather than the exception in setting regulatory transport taxes.

The present article discusses some of the issues that will become particularly important when designing transport pricing policies in practice. We start with an exposition of the economic optimality of marginal external cost pricing in Sect. 2. An important distinction will be made regarding the short-run and long-run optimality of marginal external cost pricing, and in particular the latter will be considered in more detail. It will then be emphasised that the 'textbook' case is unfortunately nothing more than a hypothetical bench-mark. It is 'a bench-mark' because it is unique in simultaneously providing optimal short-run and long-run incentives for behavioural changes. Practical pricing schemes will therefore benefit, in particular in the long run, from being designed according to first-best principles as closely as possible. However, it is 'hypothetical' because in reality, transport charging will often be a matter of second-best pricing. This is caused by the highly unrealistic nature of the assumptions underlying the derivation of the standard first-best Pigouvian tax rule. Section 3 proceeds by discussing some important issues in second-best pricing. The discussion focuses on second-best tax rules. Such rules will be discussed for four typical second-best distortions

in road transport pricing: distortions on other routes, in other transport modes, in other economic sectors, and finally due to government budget constraints. Finally, Sect. 4 offers some concluding remarks.

2 Marginal external cost pricing: a basic exposition

Transport in general, and road transport in particular, causes a variety of external effects: real impacts on the welfare levels of other agents, which are not accounted for by those causing the effects due to the complete lack or at least non-optimal feature of prices.[1] Normally, in the context of transport, the following main external cost categories are distinguished: congestion, environmental effects, noise annoyance, and accidents. Road transport is generally identified as the most important inland transport mode in terms of external cost generation. For that reason, the discussion in this article will be cast in terms of road transport. It should be noted, though, that the principles discussed are often applicable to other transport modes as well.

2.1 Short-run optimality in road usage

An important distinction regarding the external costs of road transport is between 'intra-sectoral externalities' on the one hand, which are, like congestion and part of the external accident costs, posed upon one-another by road users, and 'inter-sectoral externalities' on the other, which are posed upon society at large. The latter include environmental externalities, noise annoyance, and another part of the external accident costs. This distinction may sometimes give rise to confusion on the question of exactly what is an externality. For instance, it is sometimes argued that congestion would not be an externality, because it is internal to the road transport sector (road users only hinder each other, and no-one else suffers). However, it is important to bear in mind that for a correct welfare analysis, the relevant level of disaggregation is of course the individual level. At least from a welfare economic viewpoint, therefore, both intra-sectoral and inter-sectoral externalities are Pareto-relevant. This also follows from the standard diagram of road transport externalities depicted in Fig. 1 (attributed to, for instance, Walters 1961).

Figure 1[2] shows how, due to the existence of intra-sectoral and environmental external costs, the unregulated free market outcome exceeds the Pareto optimal level of road mobility. The market equilibrium N^0 is at the intersection

[1] A precise definition of external effects, based on the writings of for instance Mishan (1971) and Baumol and Oates (1988) could be as follows: an external effect exists when an actor's (the receptor's) utility (or production) function contains a real argument whose actual value depends on the behaviour of anther actor (the 'supplier of the effect'), who does not take this effect of his behaviour into account in his decision making process. This definition guarantees that unpriced effects other than Pareto-relevant externalities are excluded, such as pecuniary externalities, barter trade, etc. A further discussion can be found in Verhoef (1996).

[2] The discussion of Fig. 1 draws on earlier expositions as given in, for instance, Verhoef (1996).

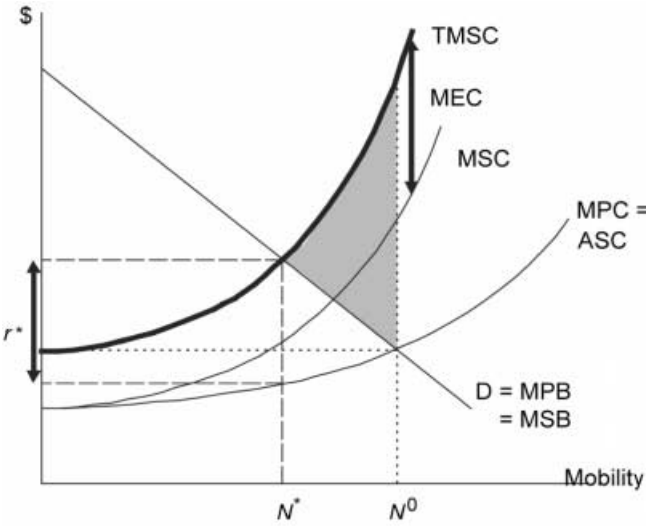

Fig. 1. The graphical representation of the bench-mark model of road transport externality regulation

of the demand curve, which is equal to the marginal private and social benefits (D = MPB = MSB),[3] and the marginal private cost curve (MPC). The latter is positively sloped because of intra-sectoral externalities, such as congestion. This reflects that the private costs of road usage increases with the level of road usage; that is, with the number of road users sharing the road together with the marginal road user. Because individual road users do not consider their own impact on average speed and safety when deciding whether to use the road, but rather take the congestion and safety levels as given, MPC may be equated to average social cost (ASC). Therefore, taking account of intra-sectoral externalities only, MSC represents marginal social costs. The fact that MSC is necessarily higher than ASC follows from the fact that the total costs are by definition equal to $N \cdot$ ASC, so that the marginal costs are equal to $\text{ASC} + N \cdot \partial \text{ASC}/\partial N$. [4]

When accounting for the fact that also inter-sectoral (marginal) external costs exist, such as environmental effects and noise annoyance, represented by MEC (where E stands for 'environmental'), TMSC gives the 'total marginal social costs' (TMSC is found by shifting MSC upwards by a distance equal to MEC). From the economic perspective, socially optimal road usage is therefore at N^*, where net social benefits, given by the area between the curves MPB and TMSC, is maximised, and the shaded welfare loss is avoided.

[3] Significant external benefits of road transport are not likely to exist; the benefits are usually either purely internal or pecuniary in nature (Verhoef 1996). Hence, MPB and MSB are assumed to be identical in Fig. 1.

[4] Along the same line of reasoning, it is easy to show that MSC should also be *steeper* than ASC: $\partial \text{MSC}/\partial N = N \cdot \partial^2 \text{ASC}/\partial N^2 + 2 \cdot \partial \text{ASC}/\partial N > \partial \text{ASC}/\partial N$, because all terms are positive.

Although diagrams such as Fig. 1 are usually taken to represent the situation at a certain road on a certain time of day, the figure can also be seen as an abstraction for the more general road transport issue.

The optimal road price that would secure the realisation of the optimal mobility is depicted by r^*, which is equal to the level of the marginal external cost in the optimum. After imposition of this charge, drivers between N^* and N^0 – whose road usage is excessive from a social perspective since the social benefits do not outweigh the social costs – will not find it attractive to use the road anymore, since their benefits of road usage (MPB) then falls short of the sum of the marginal private cost (MPC) *plus* the charge r^*.

Figure 1 thus gives the basic theory of optimal transport pricing in a nutshell. A number of points are worth emphasising. The first is that the postulation of given demand and cost curves implies that essentially a short run-view is taken. We will therefore address long-run issues surrounding marginal external cost pricing in Sect. 2.2 below. The second point is that the bench-mark model presented in Fig. 1 relies on a number of rather essential but – unfortunately – unrealistic assumptions. These will be addressed explicitly in Sect. 2.3. Finally, it is important to emphasize that the first-best character of the optimal road charge r^* can actually be attributed to two distinct features:

- The charging *mechanism* itself is optimal: all road users face a charge exactly equal to their marginal external costs. This point may appear somewhat abstract in the present context with one link and basically identical users, but the relevance will become clear in the following where second-best charging mechanisms are considered.
- The *level* of the charge is optimal: the fee is set equal to the marginal external costs *in the optimum*. It can be noted that if, for instance, the charge were based on the marginal external costs in the non-intervention outcome N0, it would overshoot its target. It can also be noted that, due to the variability of the marginal external costs, the optimal value of the road charge cannot be determined unless the optimal level of road mobility is known.

2.2 Long-run optimality and marginal external cost pricing

An analysis such as represented in Fig. 1 presupposes that the various curves identified are stable. This implies that the analysis pertains to a certain time frame within which the various factors determining the shape and positions of these curves cannot change. Such a time frame is normally, somewhat tautologically, referred to as the short run. Evidently, it is an attractive property that the marginal external cost charge secures an optimal level of road usage in the short run. However, from a policy perspective, a probably even more important question concerns the long-run characteristics of such a measure. Is the optimality of pricing measures maintained in the long run, or are additional measures necessary to steer long-run developments in desirable directions? The answer is reassuring: also in the long run, marginal external cost pricing provides first-best incentives;

that is: they affect the factors behind the 'shapes and positions' of short-run curves as shown in Fig. 1 in an optimal way.

The long-run optimality of marginal external cost pricing in general has been discussed in the context of entry-exit behaviour of firms by, for instance, Spulber (1985). We will here provide illustrations of this general result – which is basically an application of the so-called 'envelope theorem' – for road transport, using relatively simple but hopefully illuminating models, which are designed only to demonstrate the general point that marginal external cost pricing provides optimal incentives not only in the short run, but also in the long run. Before presenting these models, it is worthwhile explaining what we mean with 'long-run factors', by mentioning some examples for each of the relevant curves shown in Fig. 1:

1. *Factors affecting the shape and position of the demand curve for transport (D)* may include locational choices of firms and households. The demand for transport, very often, is a derived demand, depending on differences in spatial distributions of the supply and demand of goods, production factors (for instance labour in the context of commuting), and services. When considering peak demand, also issues like the flexibility in working and shopping hours could be mentioned as an important factor determining the shape of the demand function for peak traffic (in particular, its elasticity).
2. *Factors affecting the shape and position of the marginal inter-sectoral external cost curve* (MEC; not shown explicitly in Fig. 1, but implied by the vertical distance between TMSC and MSC) are those factors that determine the emissions of pollutants or noise per vehicle kilometre. These factors will often be related to the vehicle technology used and the driving style.
3. *Factors determining the shape and position of the marginal intra-sectoral external cost curve* (not shown explicitly in Fig. 1 but implied by the vertical distance between MSC and MPC) are usually related to the capacity and the quality of the infrastructure, in particular as far as the congestion externality is concerned.
4. *Factors determining the shape and position of the marginal private cost curve* (MPC) include the same factors as mentioned above for congestion, but also factors like fuel-efficiency, and indeed any factor influencing the private costs of road usage. Hence, also here, vehicle technology will be an important factor.

Now, in order to demonstrate the long-run optimality of marginal external cost pricing, we will consider three simple models: one dealing with factors behind the demand curve, one with factors behind the marginal environmental cost curve, and finally, one involving optimal investments in road infrastructure under conditions of congestion.

2.2.1 Optimal locational choice in the presence of transport externalities

To illustrate the optimal incentives that marginal external cost pricing in transport give in terms of locational choices, consider the following simple model. Suppose that there are N individuals, who can select a residence in either area A or area B. All individuals have identical individual (inverse) demand functions $DTR(n)$ for making trips to a third area, say the city centre (CBD), where n is the number of trips made by that person. The distance between A and CBD is F times as large as between B and CBD, so both the private costs C^P and the environmental costs C^E are F times as large (there is no congestion). However, area A is generally considered to be more attractive to live in for other reasons (otherwise, area A would of course be an irrelevant alternative). Because we consider the long run, it is assumed that dwellings are offered in both areas R according to a not-perfectly-inelastic local supply curve S_R. There are no externalities other than the environmental effects of transport. Dwellings are, apart from their location, homogeneous, and it is assumed that dwellings are supplied efficiently; that is, the supply S_R coincides with the marginal social costs.

The locational benefits of living in B are normalised to zero, and $D^{LOC}(X)$ is subsequently used to give the 'excess benefits' of living in A. Hence, $D^{LOC}(X)$ represents the inverse demand for location in area A rather than B, and hence the marginal willingness to pay to live in A (rather than B) for the X'th individual (with $0 < X \leq N$). We can then derive that in any long run equilibrium, with X individuals living in A and $N - -X$ in B, the 'generalised cost difference' between living in A and B must be $D^{LOC}(X)$. If this generalised cost difference is smaller than $D^{LOC}(X)$, more people would be attracted to A; otherwise, the opposite occurs. This generalised cost difference between living in A and B is in the present model given by the price difference between dwellings, *plus* the difference in net private benefits due to individually optimised mobility to the CBD, *given* the locational choice and *given* the prevailing type of transport regulation.

The total social welfare in the system can then be written out as the sum of the net benefits of location behaviour, and the net benefits of transport (given the location chosen):

$$\begin{aligned} W = & \int_0^X D^{LOC}(x)\,dx - \int_0^X S_A(x)\,dx - \int_0^{N-X} S_B(x)\,dx + \\ & X \cdot \left(\int_0^{n_A} D^{TR}(x)\,dx - n_A \cdot F \cdot \left(C^P + C^E\right) \right) + \\ & (N - X) \cdot \left(\int_0^{n_B} D^{TR}(x)\,dx - n_B \cdot \left(C^P + C^E\right) \right) \end{aligned} \tag{1}$$

where n_R gives the number of trips made by an inhabitant of area R. We can find the overall optimum by taking the first derivatives of (1) with respect to X,

n_A and n_B. This yields:

$$\begin{aligned}\frac{\partial W}{\partial X} \;=\; & D^{LOC}(X) - S_A(X) + S_B(N-X) + \\ & \left(\int_0^{n_A} D^{TR}(x)dx - n_A \cdot F \cdot \left(C^P + C^E\right)\right) - \\ & \left(\int_0^{n_B} D^{TR}(x)dx - n_B \cdot \left(C^P + C^E\right)\right) = 0\end{aligned} \tag{2a}$$

$$\frac{\partial W}{\partial n_A} = X \cdot \left(D^{TR}(n_A) - F \cdot \left(C^P + C^E\right)\right) = 0 \tag{2b}$$

$$\frac{\partial W}{\partial n_B} = (N-X) \cdot \left(D^{TR}(n_B) - \left(C^P + C^E\right)\right) = 0 \tag{2c}$$

However, for given locally differentiated transportation taxes r_A and r_B, individuals will act according to the following equations:

$$\begin{aligned}& D^{LOC}(X) - (S_A(X) - S_B(N-X)) + \\ & \left(\int_0^{n_A} D^{TR}(x)dx - n_A \cdot (F \cdot C^P + r_A)\right) - \\ & \left(\int_0^{n_B} D^{TR}(x)dx - n_B \cdot (C^P + r_B)\right) = 0\end{aligned} \tag{3a}$$

$$D^{TR}(n_A) - F \cdot C^P - r_A = 0 \tag{3b}$$

$$D^{TR}(n_B) - C^P - r_B = 0 \tag{3c}$$

Equation (3a) describes individually optimising locational choice, taking into account that an individual will act so as to maximise the net private benefits of transport given the location chosen, and Eqs. (3b) and (3c) show the selection of the individually optimising number of trips, given the choice of a location, and given the prevailing transportation taxes.

Comparing (2b) and (3b), and (2c) and (3c), we first find that for both types of trips the optimal transport taxes should be equal to the marginal external costs, exactly as depicted in Fig. 1:

$$r_A = F \cdot C^E \tag{4a}$$

$$r_B = C^E \tag{4b}$$

If we then substitute these taxes into (3a), it is easy to see that the incentives to locate in either area A or B are then exactly according to the social optimality condition (2a). Hence, the long-run decision of where to reside will then be made in line with overall economic efficiency, and no further regulation regarding locational decisions is necessary. At the same time, it can be seen that when the optimal transportation taxes r_R are not used and are set equal to zero, we

will not only have inefficiently high mobility levels for both types of transport given the location of people, but in addition also an inefficiently high number of residents in area *A*, which 'boosts' also the demand for the relatively polluting type of mobility in the sense that the demand curve for trips of type *A* is 'too much outward rotated', and for type *B* 'too much inward rotated'.

This admittedly simple model thus illustrates how optimal transport taxes not only optimise transport *given* the shape and position of the demand curves, but also create optimal incentives to change those aspects of behaviour that affect the actual shape and position of these demand curves in the long run.

Finally, it can be mentioned that this type of result carries over to more complex settings. An interesting example can be found in Oron et al. (1973), who consider optimal location in a continuous-space monocentric city with congested roads. Although they are unable to give a conclusive analytical answer to the question of whether too much sub-urbanisation occurs without optimal congestion taxes, they do show that optimal congestion charges are necessary for decentralising efficient locational choices.

2.2.2 Optimal environmental technology choice in the presence of transport externalities

According to the same sort of principles, we can spell out an equally simple model that demonstrates the mechanism of optimal user charges affecting the position and shape of the MPC and MEC curves implied in Fig. 1 in an optimal manner. To keep things simple, we again assume that there is no congestion, and that all trips have the same length (a formulation with a demand for vehicle-kilometres instead of trips would in fact be exactly the same as the present one). We further assume that, by obtaining more expensive cars, road users can improve the energy efficiency, ε, above some default level ε_0, and hence have lower private costs (through a lower fuel input per kilometre travelled) as well as lower emissions and hence external environmental costs per kilometre travelled. The (per vehicle) marginal cost of such improvements is given by the function $C^{\varepsilon}(\varepsilon)$. These improvements are fixed costs with respect to the number of trips. Denoting the environmental costs per trip as $C^{E}(\varepsilon)$ and the private costs as $C^{P}(\varepsilon)$, the (inverse) demand curve for road use as $D(n)$ where n gives the number of trips per individual, and the number of (identical) individuals as N, we can write out the following social welfare function:

$$W = N \cdot \left(\int_0^n D(x)dx - n \cdot \left(C^P(\varepsilon) + C^E(\varepsilon) \right) - \int_{\varepsilon_0}^{\varepsilon} C^{\varepsilon}(x)dx \right) \tag{5}$$

The social optimum requires the selection of an optimal ε and n according to:

$$\frac{\partial W}{\partial \varepsilon} = N \cdot \left(-n \cdot \left(\frac{dC^P(\varepsilon)}{d\varepsilon} + \frac{dC^E(\varepsilon)}{d\varepsilon} \right) - C^{\varepsilon}(\varepsilon) \right) = 0 \tag{6a}$$

$$\frac{\partial W}{\partial n} = N \cdot \left(D(n) - \left(C^P(\varepsilon) + C^E(\varepsilon)\right)\right) = 0 \tag{6b}$$

Road users, when being informed on the nature of the user charge being made optimally dependent on the technology chosen, so that it can be written as $r(\varepsilon)$, will invest in energy efficiency improvements up to the point where the marginal private costs of doing so become equal to the marginal private benefits in terms of reduced private costs and reduced charges for road usage. Hence, individual road users act so as to set:

$$-n \cdot \left(\frac{dC^P(\varepsilon)}{d\varepsilon} + \frac{dr(\varepsilon)}{d\varepsilon}\right) - C^{\varepsilon}(\varepsilon) = 0 \tag{7a}$$

Given the choice of a technology, they will choose a level of mobility according to:

$$D(n) - C^P(\varepsilon) - r(\varepsilon) = 0 \tag{7b}$$

Again, comparing (7a) and (6a), and (7b) and (6b), it turns out that the first-best pricing rule:

$$r(\varepsilon) = C^E(\varepsilon) \tag{8}$$

simultaneously optimises the choice of technology as well as the level of mobility, *given* the technology chosen. Since the technology in this example affects, simultaneously, the marginal private costs and the marginal external costs, we have illustrated how the long-run decisions – now in terms of technology choice – are 'automatically' optimised using the marginal external cost pricing rule.

2.2.3 Optimal investments in infrastructure capacity under conditions of congestion

The last example concerns the choice of optimal road capacity in the presence of congestion. This problem differs somewhat from the two foregoing ones in the sense that in this case, the long-run decision (i.e., the choice of infrastructure capacity) is normally made by a different actor (i.e., the government) than the short-run decisions (i.e., the choice of using the infrastructure, which is of course made by the potential road users). Still, we wish to consider this example, also because it underlines the long-run optimality of marginal external cost pricing in an elegant and policy relevant way.

Consider a single road, and denote its capacity K. The average user costs of making a trip are denoted as $C^P(N,K)$, where N gives the number of users. Let $\partial C^P/\partial N \geq 0$ represent congestion (at a given capacity), and $\partial C^P/\partial K \leq 0$ mitigation of congestion through capacity expansion. The marginal social cost of capacity expansion is given by the function $C^K(K)$. We can then write out the following social welfare function:

$$W = \int_0^N D(x)dx - N \cdot C^P(N,K) - \int_0^K C^K(x)dx \tag{9}$$

The social optimum requires the selection of an optimal N and K according to:

$$\frac{\partial W}{\partial K} = -N \cdot \frac{\partial C^P(\cdot)}{\partial K} - C^K(K) = 0 \tag{10a}$$

$$\frac{\partial W}{\partial N} = D(N) - C^P(\cdot) - N \cdot \frac{\partial C^P(\cdot)}{\partial N} = 0 \tag{10b}$$

For a given capacity and with road pricing, road users will choose a level of mobility according to:

$$D(N) - C^P(\cdot) - r = 0 \tag{11}$$

Comparing (11) and (10b) gives us the standard first-best congestion charge:

$$r = N \cdot \frac{\partial C^P(\cdot)}{\partial N} \tag{12}$$

On the basis of (10a) and (12), we can now derive one of the more famous analytical results in transport economics, namely that the revenues from optimal road pricing are, under certain conditions, just sufficient to cover the cost of the optimal supply of road infrastructure capacity (Mohring and Harwitz 1962). These conditions involve constant returns to scale in congestion technology, so that $C^P(N, K)$ can be written as $C^P(N/K)$, and constant returns to scale in capacity extension ($C^K(K)$ is constant). Writing $N/K = R$ (ratio), we find:

$$\frac{\partial C^P(\cdot)}{\partial N} = \frac{\partial C^P(\cdot)}{\partial R} \cdot \frac{\partial R}{\partial N} = \frac{\partial C^P(\cdot)}{\partial R} \cdot \frac{1}{K} \tag{13a}$$

$$\frac{\partial C^P(\cdot)}{\partial K} = \frac{\partial C^P(\cdot)}{\partial R} \cdot \frac{\partial R}{\partial K} = \frac{\partial C^P(\cdot)}{\partial R} \cdot \frac{-N}{K^2} \tag{13b}$$

Using (13a) and (12), we can then write the total revenues from congestion pricing as:

$$T = N \cdot N \cdot \frac{\partial C^P(\cdot)}{\partial N} = \frac{N^2}{K} \cdot \frac{\partial C^P(\cdot)}{\partial R} \tag{13c}$$

Substituting (13b) into (10a) and multiplying both sides by K finally yields:

$$\frac{N^2}{K} \cdot \frac{\partial C^P(\cdot)}{\partial R} - K \cdot C^K(K) = 0 \tag{13d}$$

The first term again gives the total revenues from congestion pricing, and the second term the total costs of infrastructure capacity supply in case $C^K(K)$ is constant. In other words, the government will then have a balanced budget, and outlays on infrastructure capacity expansion can be exactly covered by the revenues from optimal pricing. It should be noted that the fact that (13d) implies a surplus (deficit) for the government in case of decreasing (increasing) returns to scale in capacity expansion is of course not something specific to transport – this property holds for optimal pricing in any market (Varian 1992).

The purpose of this last example is to highlight a slightly different type of long-run optimality result from short-run marginal cost pricing. Here, the main

message involves the optimal level of infrastructure capacity that would result from investing the revenues from optimal pricing.

Clearly, the models presented above are rather abstract. The main purpose of these models, however, is to demonstrate the general principle that the pricing rule that is optimal in the short run, with given demand and cost curves, also provides optimal incentives in terms of long-run behavioural issues, that determine the shape and position of these curves. This result surely is not merely an interesting academic side-issue. It is of utmost importance for the evaluation of second-best pricing mechanisms, which will very often lack such 'convenient' properties, and therefore often require additional types of measures to compensate for the implied 'lost incentives'. This will be discussed in further detail in Sect. 2.3 below.

Another remark that can be made regarding the three illustrative models just presented is, that it was assumed that all other relevant markets operate efficiently. These other markets include the housing market in the first case and the automobile market in the second. If this were not the case, additional second-best considerations would of course enter and complicate the analysis. Clearly, these can be considered in the present models, but were ignored above deliberately, for reasons of exposition. In Sect. 3 below, however, such issues will be considered in more detail, by investigating the impacts of such distortions on second-best transportation taxes. It should be mentioned already, though, that generally such other market failures could be dealt with more efficiently by direct intervention at the source of the distortion, rather than by applying appropriately corrected second-best pricing rules.

2.3 Marginal external cost pricing: a hypothetical but crucial bench-mark

Given the short-run and long-run optimality of marginal external cost pricing outlined above, the question rises why such evidently attractive instruments have not, or only sparsely, been used in the practice of policy making. Apart from issues related to the limited social feasibility of pricing instruments, a different sort of explanation for this paradox may be the fact that reality is often a lot more complicated than the simple world assumed in Fig. 1. This, in turn, may seriously complicate the determination and application of optimal road charges in reality. It is instructive to explicitly list the most important implicit assumptions underlying the model depicted in Fig. 1:

1. There is complete certainty and perfect information on all benefits and costs (including external costs) of road usage;
2. Road users are completely homogeneous, and only (possibly) differ in terms of their marginal willingness to pay to use the road;
3. The demand curve is stable over time, so that a static approach is valid;
4. The road system is a one-link network;
5. The (spatio-)economic system within which this transport network operates is otherwise in a first-best optimum, without any uncorrected market failures like external effects, market power, distortionary taxes, and so forth.

Clearly, other than possibly in a few analysts' minds, these assumptions will never be met. Unfortunately, however, once these assumptions are relaxed, things can become much more complicated than is suggested by the basic analysis of Fig. 1. This is especially so because when more realism is added to the model, an increasingly complicated optimal charge scheme will result, where optimal user charges will vary according to many dimensions. In particular, because the first-best principle that optimal user charges should be equal to marginal external costs caused remains valid, these charges should vary along with variations in marginal external costs caused by individuals. Recalling that transport externalities include a large variety of effects – congestion, emissions, noise annoyance, accidents – optimal individual charges should therefore vary at least according to the following dimensions:

1. the vehicle (technology) used,
2. the actual state of this vehicle,
3. the kilometrage,
4. the time of driving,
5. the place of driving,
6. the actual route chosen,
7. the driving style.

Only then can the feature of road charging provide optimal incentives to change behaviour in both the short run and the long run that possibly carry over to real-life situations. Technically speaking however, such a situation can only be realised if one would apply some 'Big Brother' type of electronic road charges, using very sophisticated technologies that can monitor the actual emissions, the place and time of driving, the driving style, and the prevailing traffic conditions; and that allows the regulator to adjust the charge accordingly (see Johansson-Stenman and Sterner 1998, for a thoughtful evaluation of the pros and cons of such systems). Even disregarding the social acceptance of such technologies, possibly seriously intruding the drivers' privacy, even from a purely technological viewpoint such systems are not likely to be introduced on a significant scale in the foreseeable future. Still, the use of electronic congestion charges at an increasing number of sites throughout the world (see Small and Gomez-Ibañez 1998, for an overview) can of course be seen as a major step into this direction.

As a result, one will often have to rely on imperfect substitutes to the first-best scheme when implementing pricing policies in reality. From a theoretical viewpoint, this implies that we will have to accept second-best pricing measures. Apart from the fact that such second-best instruments will generally provide only imperfect incentives in terms of affecting mobility in the short run, a probably even more important shortcoming will be that the 'convenient property' of first-best charges affecting also all long-run decisions in an optimal manner, will be partially or even entirely lost. As a result, there will often be a need for complementary policies affecting long-run behavioural aspects underlying transport decisions in a socially desirable manner. In other words, the acceptance of second-best pricing measures generally implies that 'policy packaging' should be

at the core of the policy design. Specifically, it is very likely that separate policies should be used to affect short-run decisions concerning mobility behaviour, and long-run decisions concerning factors determining the long-run position and shape of the relevant demand and cost curves identified above.

For reasons of space, the specific complications of policy packaging will not be discussed in any detail in this article. However, one general remark concerning the use of such second-best policy packages should be made. This concerns the very important, but yet often neglected issue of the implied informational and organisational burden that the deviation of marginal external cost pricing implies for the regulator. In particular, it should be noted that the optimal incentives created by the first-best policy naturally imply that decisions regarding these long-run issues can be left to the market – unless, of course, other market failures would exist (but these, in turn, could then normally be dealt with more efficiently using measures directly aimed at mitigating those failures). Hence, there is actually no need for the regulator to obtain any information other than the optimal value of the actual marginal external costs – which in itself will often be complicated enough. Things become quite different, however, when second-best policies are used, and the regulator feels that some of the implied imperfect or perhaps even entirely lacking long-run incentives should be compensated for using policies directly aiming at the relevant long-run issues. The regulator should then formulate expectations about the *(second-best) optimal* long run developments determining the position and shape of the demand and cost curves for transport (i.e., how would they have been affected when first-best policies were used, and what is the second-best optimal shape and position, given the fact that first-best pricing cannot be used?), and subsequently specify policies that could achieve the implied targets in a socially cost-effective way. The implied informational and organisational burden will of course be enormous, and the question remains whether a regulator would ever be able to collect all information present in the relevant markets, and process it as efficiently as an optimal market would do.

Therefore, apart from the unavoidable welfare losses that result by definition from the use of second-best instruments compared to first-best instruments, it is very likely that additional welfare losses due to ‘government failures’ will then further reduce the social benefits from regulation. As just argued, such government failures may in the first place result from the strongly increased informational needs, necessary for applying second-best instruments optimally. A second reason, however, would be that second-best tax rules will generally be much more complicated than the relatively simple first-best ‘tax-equals-marginal-external-costs’ rule. This is the topic of the next section.

3 Applying second-best policies in practice: The use of second-best tax rules

As argued above, under the rather stringent assumptions that first-best conditions pertain elsewhere in the economic system and perfectly flexible regulatory policies exist for coping with road transport externalities, there would be little scope

for improving on the standard Pigouvian solution to the problem of external costs of road transport, as depicted in Fig. 1. These assumptions are, however, usually not satisfied. Second-best problems have, accordingly, received ample attention in the recent literature on road pricing. For instance, Wilson (1983), and d'Ouville and McDonald (1990) studied optimal road capacity with suboptimal congestion pricing; Braid (1989) and Arnott et al. (1990) and Laih (1994) consider uniform or step-wise pricing of a bottleneck. Arnott (1979) and Sullivan (1983) look at congestion policies through urban land use strategies. Two classic examples on second-best regulation in road transport are Lévy-Lambert (1968) and Marchand (1968), studying optimal congestion pricing with an untolled alternative, an issue that was recently discussed also by Braid (1996), Verhoef et al. (1996), and Liu and McDonald (1998, 1999). Glazer and Niskanen (1992) as well as Verhoef et al. (1995) have studied second-best aspects related to parking policies, and Mohring (1989) considered fuel taxation.

An essential joint conclusion from these studies is that, when applying second-best regulatory tools, economic efficiency requires these instruments to be applied according to different *rules* than that apply for the first-best benchmark policy. This links in with the two last points raised in Sect. 2.1, where the optimality of marginal external cost pricing was attributed to two separate features: the charging mechanism is perfect, and the charge is set at the optimal level. In second-best regulation, the second-best optimal *tax rules* (if a tax is used) should account for the imperfection of the instrument itself, in order to use it in a welfare optimising manner; given, of course, the persistence of the second-best aspects.

In this section, this point will be clarified, by again giving a number of examples. Four models will be presented, reflecting four major types of second-best distortions that may occur in reality. The first model deals with distortions on other routes within the same mode, the second one with distortions in other transport modes, the third one with distortions elsewhere in the economy, and the fourth one with distortions due to government budget constraints. As one can imagine, second-best tax rules can become rather complicated as models become more realistic. Therefore, the second-best tax rules will be considered only for the most simple model settings possible. This is not because these simple models are considered representative for any real-world situation, but simply because this allows us to concentrate on the basic economic issues at hand, in analytically tractable models. For each of the four models, more realistic extensions can be constructed rather straightforwardly, based on the same methodology as used below.

3.1 Second-best tolling with distortions on other routes

The impact of distortions on other routes in a road network on second-best tolls can be illustrated by considering the classic two-route problem (Lévy-Lambert 1968). This entails a two-link network, connecting a joint origin-destination pair.

Road users distribute themselves over both routes according to the rule that marginal private costs, including tolls if there are any, should be equalised over both routes. Now, if there is congestion, it is easy to show that if the regulator can levy a charge on both routes, optimal congestion charges as derived in (12) would apply for both routes:

$$r_i = N_i \cdot \frac{d\,C_i^P(N_i)}{d\,N_i} \tag{14a}$$

where the subscript i denotes the particular route considered. However, if the regulator is for some reason only capable of putting tolls into effect on one route only (say route T), and has to leave the other route (U) untolled, it would be incorrect to apply the first-best tax rule (14a) as if first-best conditions apply throughout the network. Instead, for this particular problem, the following second-best tax rule for route T can be derived:

$$r_T = N_T \cdot \frac{d\,C_T^P(N_T)}{d\,N_T} - N_U \cdot \frac{d\,C_U^P(N_U)}{d\,N_U} \cdot \frac{-\dfrac{d\,D(N)}{d\,N}}{\dfrac{d\,C_U^P(N_U)}{d\,N_U} - \dfrac{d\,D(N)}{d\,N}} \tag{14b}$$

It would take too far to discuss this tax rule in great detail here: a comparable problem is discussed below, and the interested reader may in addition refer to Verhoef et al. (1996), where the derivation of (14b) is also provided. It is, however, worth pointing out that for the specification of the second-best one-route toll, one has to take account of the specific situation on the other, untolled route, as well as of the prevailing demand structure (in particular the demand elasticity, or more precisely: the slope of the demand curve). This has to do with the spill-overs that regulation on route T imply for the driving conditions on route U, and with the fact that one single tax aims to control two variables affecting the overall efficiency: the overall level of demand, and the route split. Note that the expression is composed of a term reflecting the marginal external cost on the tolled route, and a second (negative) term representing those on the untolled route, weighted with a fraction that may vary between 0 and 1. Also, observe that the second-best toll may therefore be negative.

In general, this type of second-best problem may actually often be 'self-imposed' by the regulator. In particular, when electronic charging mechanisms are used, it may be considered inefficient to apply charges on all links, due to the high fixed costs of installing the necessary equipment. Hence, the regulator may choose to have toll-points installed only on a few key links in the network. The second-best tolling problem resulting from such situations has recently been studied for general networks by Verhoef (1998).

3.2 Second-best tolling with distortions in other modes

A second important assumption underlying the applicability of standard Pigouvian taxes is that alternative transport modes are efficiently priced. The validity

of this assumption is of course often questionable, to say the least. In particular, it may often be the case that public transport services, for instance due to subsidies, are inefficiently priced from an overall social welfare perspective. The resulting impacts for second-best tax rules in private transport can be derived in a manner that, from an analytical viewpoint, is comparable to the two-route problem considered above. This reflects that, in a way, one could interpret public transport as an inefficiently priced alternative route.

In order to study this problem, let us again consider a simple model. To keep the analysis simple, and manageable using continuous functions, the following assumptions are made. Consider the short run, so that only variable costs matter. The generalised variable costs for public transport, as experienced by its users, are made up of two components: the price of the ticket P^T, and a term C^T reflecting the valuation of the average (per passenger) travel time in public transport. The total short-run social costs of public transport are given by the total variable costs made by the operator, $TVCTO$, plus the travel time costs C^T times the number of users N^T. There is therefore neither congestion, nor a 'Mohring-effect'[5] (Mohring 1972) present in public transport. We use, as before, $C^P(N^P)$ to denote the average generalised costs for road usage, where N^P gives road usage. Finally, there is one shared demand function for transport, $D(N)$, where $N = N^T + N^P$. Hence, mode choice in this simple model results from generalised private cost differences.[6] Finally, we wish to take account of the fact that the public transport operator may have some market power. In particular, he is not a price-taker and can, for instance, change the price depending on the average costs. This will be reflected below by the very general formulation that the ticket price P^T may depend on the level of usage N^T. The exact pricing rule used needs, for the present purpose, not be made explicit.

Under these assumptions, the second-best congestion toll for road transport r can be found by solving the following Lagrangian, showing that the objective is to maximise the difference between total benefits and total costs, under

[5] The Mohring-effect is the reverse of congestion, reflecting the positive externality that public transport users create for each other through the increased frequency that is (in the long run) associated with increased usage.

[6] One could of course also model this choice using finite cross-elasticities of demand, but that is firstly an unnecessary complication for the present purpose, and secondly quite restrictive too, in the sense that this cross-elasticity assumes a high degree of homogeneity of users, and a relatively low degree of substitutability between private and public transport. Taking the view that the eventual good demanded is the move from A to B, a trip by private or by public transport would be perfect substitutes. Probably, the 'correct' way of modelling mode choice would allow for individually differentiated levels of generalised costs attached to the use of public transport, reflecting taste differences such as non-monetary costs associated with privacy, comfort, reliability, and so forth. Continuous functions can then no longer be used in the optimisation procedure, since there needs not be a perfect correlation between willingness to pay for trips, and individual generalised costs for using public transport. However, since these differences between individual generalised costs are fully internalised, a second-best tax rule comparable to the one given below is then still likely to result. To see this, observe that different individual generalised costs attached to using public transport would equally directly appear in the first-order conditions (17) and (20) below, also when the choices of 'discrete' individuals were to be optimised. Hence, the simplified procedure followed here certainly needs not necessarily imply that the result is flawed.

the restrictions caused by individually optimising behaviour, equating marginal benefits to marginal private generalised costs for both modes:[7]

$$\begin{aligned}\Lambda = & \int_0^{N^P+N^T} D(x)dx - N^P \cdot C^P(N^P) - N^T \cdot C^T - TVC^{TO}(N^T) \\ & +\lambda_P \left(C^P(N^P) + r - D(N^P + N^T)\right) \\ & +\lambda_T \left(C^T + P^T(N^T) - D(N^P + N^T)\right)\end{aligned} \tag{15}$$

yielding the following first-order conditions:

$$\frac{\partial \Lambda}{\partial N^P} = D - C^P - N^P \cdot \frac{d\,C^P}{d\,N^P} + \lambda_P \cdot \frac{d\,C^P}{d\,N^P} - (\lambda_P + \lambda_T) \cdot \frac{d\,D}{d\,N} = 0 \tag{16}$$

$$\frac{\partial \Lambda}{\partial N^T} = D - C^T - \frac{d\,TVC^{TO}}{d\,N^T} + \lambda_T \cdot \frac{d\,P^T}{d\,N^T} - (\lambda_P + \lambda_T) \cdot \frac{d\,D}{d\,N} = 0 \tag{17}$$

$$\frac{\partial \Lambda}{\partial r} = \lambda_P = 0 \tag{18}$$

$$\frac{\partial \Lambda}{\partial \lambda_P} = C^P + r - D = 0 \tag{19}$$

$$\frac{\partial \Lambda}{\partial \lambda_T} = C^T + P^T - D = 0 \tag{20}$$

Using (17), (18), and (20), it can be shown that:

$$\lambda_T = \frac{\dfrac{d\,TVC^{TO}}{d\,N^T} - P^T}{\dfrac{d\,P^T}{d\,N^T} - \dfrac{d\,D}{d\,N}} \tag{21}$$

Using (16), (18), (19), and (21), the following second-best toll can then be derived:

$$r = N^P \cdot \frac{d\,C^P}{d\,N^P} - \left(\frac{d\,TVC^{TO}}{d\,N^T} - P^T\right) \cdot \frac{-\dfrac{d\,D}{d\,N}}{\dfrac{d\,P^T}{d\,N^T} - \dfrac{d\,D}{d\,N}} \tag{22}$$

First of all, the reader may verify the similarity with (14b). Observe also that the sign of (22) is ambiguous. Also the interpretation of (22) is similar to that of (14b). The first term shows the direct impact of the toll on congestion on the road itself. The second term reflects that in the second-best optimum, account should also be taken of a possibly non-optimal price in public transport. The term between the large brackets represents the difference between the marginal social costs of using public transport and the ticket price. Evidently, if public transport is efficiently priced, this term vanishes, showing that the standard Pigouvian toll suffices for the regulation of road use if the alternative mode is managed

[7] The reverse of this problem, namely the second-best optimal price for public transport with unpriced road traffic congestion, was studied by Henderson (1977, pp. 153–157).

according to first-best standards. Note that the correction factor increases with the extent to which public transport prices are distorted.

As is shown by the fraction behind the term between the large brackets, the extent to which this distortion affects the second-best road price r depends also on the elasticity of the total demand for transport, and the sensitivity of public transport prices to its usage, both evaluated in the second-best optimum.

In the one extreme, where the demand is perfectly elastic, the second-term vanishes, reflecting that the usage of public transport cannot be affected with the road price. The same extreme results if the public transport system is operating at its capacity, and the price P^T is used by the operator to keep out excessive demand exceeding the capacity. Then, the use of public transport is given and determined by its capacity. In both cases, the road price can be set according to the first-best rule, since the use of public transport cannot be affected.

In the other extreme, where either the demand is perfectly inelastic or the public transport price is insensitive to it usage, the second-best road price becomes equal to the difference between the marginal external congestion costs on the road, and the extent to which the marginal social costs of public transport exceed the ticket price. With inelastic demand, this reflects that the total usage of both modes together is given, and the road price should be used so as to equate the marginal social costs for both modes in the second-best optimum. With insensitive public transport prices, it also reflects that the overall level of transport demand is given, but now by the intersection of the price-line P^T and the demand curve D. Also then, the distribution of this given number of users over both modes in the second-best optimum should of course be such that the marginal social costs are equalised.

For intermediate cases, the interpretation of the correction term can be given by considering the joint impacts of the effects just discussed for extreme situations.

This admittedly simple model is at least sufficient to demonstrate how in a second-best situation, where alternative modes are not efficiently priced, the standard Pigouvian tax rule is no longer optimal. Instead, the second-best tax rule to be used then depends on (and reflects) the distortions occurring also in the other transport modes. Clearly, also this model could be made much more realistic – at the price of increasing complexity. The general conclusion just given, however, would not be affected. Such more realistic formulations will often no longer have analytically tractable solutions as (22), and may therefore be solvable only using numerical procedures for models with explicit demand and cost functions.

Finally, it can be noted that the Lagrangian multipliers λ reflect the 'shadow price of non-optimal pricing'. Such multipliers are typical for second-best problems. These multipliers cause the second-best optimum to differ from the first-best situation, where also the alternative mode is optimally priced. In particular, note that if P^T could be chosen freely by the regulator in the optimisation procedure, we would find $\partial \Lambda / \partial P^T = \lambda_T = 0$.

Second-best tolling with distortions in other economic sectors

A next important assumption underlying the direct applicability of the standard Pigouvian tax rule is that all other economic sectors, somehow connected to the transport sector, should operate under first-best conditions themselves. This is perhaps even more unlikely to hold true in reality than the two previously considered assumptions, of first-best conditions prevailing in alternative routes and modes. In particular, given the fact that most (if not all) economic sectors require at least some transportation for their operation, the assumption actually requires, for instance, the absence of market power and the absence of unpriced environmental pollution throughout the economy. Needless to say that this will normally not be the case.

One can again illustrate the basic economic issues involved in a simple model. Let us consider the case of freight transport. To underline the second-best character of the problem, consider two polluting economic sectors (A and B), and assume that their production processes are polluting, causing constant average external costs C_A^E and C_B^E. Next, observe that the demand for freight transport is a derived demand, which is closely connected to the demand (and supply) structure for the transported good itself. In particular, the transportation of a good is (normally) a necessary step in the process of bringing the demand and supply physically together, and accomplishing a transaction. In the below model, it is therefore assumed that every unit of good traded requires a transport movement. Defining the units of both goods such that the transport effort for one unit requires the same unit transport service, the equilibrium demand for transport is simply equal to the sum of equilibrium quantities traded, Q_A and Q_B (note that we have a non-spatial model, so that all trips have equal length).

Assume again that no congestion occurs, and that the constant average private and external costs of transport can be written as C^P and C^E, respectively. Denote the demand and supply curves for both goods (i) as D_i and S_i, and assume that apart from the externality, both markets operate efficiently, with prices reflecting marginal social costs. The average transportation costs C^P will thus drive a wedge between the marginal benefits D and the marginal production costs S (see also the restrictions in the Lagrangian below). Finally, assume that only regulatory transport taxes r are available (otherwise, we would not have a second-best problem). The following Lagrangian then represents the second-best optimisation problem:

$$\begin{aligned}\Lambda \;=\; & \int_0^{Q_A} D_A(x)\,dx - \int_0^{Q_A} S_A(x)\,dx - Q_A \cdot C_A^E \\ & + \int_0^{Q_B} D_B(x)\,dx - \int_0^{Q_B} S_B(x)\,dx - Q_B \cdot C_B^E \\ & - \left(C^P + C^E\right) \cdot (Q_A + Q_B) + \lambda_A \left(S_A(Q_A) + C^P + r - D_A(Q_A)\right) \\ & + \lambda_B \left(S_B(Q_B) + C^P + r - D_B(Q_B)\right)\end{aligned} \tag{23}$$

which has the following first-order conditions (where primes denote derivatives):

$$\frac{\partial \Lambda}{\partial Q_A} = D_A - S_A - C_A^E - C^P - C^E + \lambda_A \cdot \left(S_A' - D_A'\right) = 0 \tag{24}$$

$$\frac{\partial \Lambda}{\partial Q_B} = D_B - S_B - C_B^E - C^P - C^E + \lambda_B \cdot \left(S_B' - D_B'\right) = 0 \tag{25}$$

$$\frac{\partial \Lambda}{\partial r} = \lambda_A + \lambda_B = 0 \tag{26}$$

$$\frac{\partial \Lambda}{\partial \lambda_A} = S_A + C^P + r - D_A = 0 \tag{27}$$

$$\frac{\partial \Lambda}{\partial \lambda_B} = S_B + C^P + r - D_B = 0 \tag{28}$$

Substitution of (27) in (24), and (28) in (25) yields:

$$\lambda_A = \frac{C_A^E + C^E - r}{S_A' - D_A'} \tag{29}$$

$$\lambda_B = \frac{C_B^E + C^E - r}{S_B' - D_B'} \tag{30}$$

As in the previous model, these multipliers reflect the shadow price of non-optimal pricing, now in the two goods markets. These multipliers are for both goods increasing in the difference between the marginal external costs, of production and transportation together, and the regulatory tax. Equations (26), (29) and (30) finally imply the following second-best transportation tax:

$$r = C^E + \frac{\dfrac{C_A^E}{S_A' - D_A'} + \dfrac{C_B^E}{S_B' - D_B'}}{\dfrac{1}{S_A' - D_A'} + \dfrac{1}{S_B' - D_B'}} \tag{31}$$

The second-best tax rule shows that, in addition to the 'first-best component' reflecting the marginal external costs of transport itself, a term is added which reflects the marginal external costs caused by production in the two sectors. More precisely, a weighted average of these marginal external costs is included in r, where the weight reflects the sensitivity of the equilibrium output to price distortions: if either the demand or the supply for a sector is fully inelastic, the associated term vanishes. This is of course rather intuitive: due to the inelasticity, the emissions cannot be affected, and the best thing to do for the regulator is to set the tax such that emissions from transport and from the other sector are optimised. Note also that if we happen to find $C_A^E = C_B^E$, the first-best outcome can be reproduced, since the road tax then simply includes also the external costs of production. Because every good produced is also transported, and all shipments are assumed to be equally long, a tax on transport alone is then in fact indistinguishable from the set of first-best taxes on both production and transport.

The tax rule in Eq. (31) thus shows that distortions in other economic sectors will generally affect second-best transportation taxes. Clearly, the economically optimal way of dealing with such distortions would be to use regulatory taxes directly targeted at the sectors involved, and to apply first-best tax rules throughout the economy. The purpose of the above analysis, however, is merely to illustrate that if such a tax system, for some reason, does not exist, or if taxes are not used optimally, the second-best tax rule for transport will be affected accordingly. Simply ignoring the distortions elsewhere in the economy is then non-optimal, and would lead to regulatory taxes for transport that can be improved upon. As is illustrated in Verhoef et al. (1997), in a spatial analysis of the above problem, the naïve use of standard Pigouvian taxes may then in some cases even be counter-productive, in the sense that positive taxes for transport could lead to a reduction in social welfare.

3.3 Second-best tolling with distortions due to government budget constraints

A final, somewhat different type of distortion we would like to consider here concerns the case where the government's budget constraint somehow enters the optimisation procedure. The standard procedure used for finding optimal taxes normally assumes that the marginal utility of funds is constant over actors. It has been argued, however, for instance by Ochelen et al. (1998), that this needs not always be the case. In particular, if a government uses the tax revenues from regulatory transport pricing to reduce distortive taxes on, for instance, labour, a double dividend can possibly be reaped (on this double-dividend hypothesis, see for instance Bovenberg and De Mooij 1994; and Bovenberg and Goulder 1996). Such a higher social value of tax revenues is often modelled using a 'shadow price of public funds', λ_P. This denotes the additional reward that is given to each unit of tax revenues. Note that $\lambda_P > 0$ denotes the case where it is assumed that toll revenues are used by the government in a way that enhances economic efficiency; $\lambda_P < 0$ would denote the situation where the government uses the revenues in a less efficient way than consumers would do. Applying this procedure in a simple model of transport with an environmental externality only, and maintaining the notation used before, the following Lagrangian can be set up:

$$\Lambda = \int_0^N D(x)\,dx - N \cdot \left(C^P + C^E\right) + \lambda_P \cdot r \cdot N + \lambda \cdot \left(C^P + r - D(N)\right) \qquad (32)$$

The following first-order conditions apply:

$$\frac{\partial \Lambda}{\partial N} = D - C^P - C^E + \lambda_P \cdot r - \lambda \cdot D' = 0 \qquad (33)$$

$$\frac{\partial \Lambda}{\partial r} = \lambda_P \cdot N + \lambda = 0 \qquad (34)$$

$$\frac{\partial \Lambda}{\partial \lambda} = C^P + r - D = 0 \tag{35}$$

Substitution of (34) and (35) into (33) yields the following tax rule:

$$r = \frac{C^E - \lambda_P \cdot N \cdot D'}{1 + \lambda_P} = \frac{C^E \cdot \left(1 + \lambda_P \cdot \frac{N \cdot -D'}{C^E}\right)}{1 + \lambda_P} \tag{36}$$

where the second formulation merely facilitates interpretation. It is interesting to note that Eq. (36) shows how, even if $\lambda_P > 0$, the implied tax certainly needs not exceed the standard Pigouvian rule $r = C^E$ applying with constant marginal external cost. The reason is that the sub-goal of revenue maximisation may require an upward or a downward adjustment of the tax, depending on the elasticity of demand. This is caused by the fact that a marginally higher tax rate on the one hand increases the tax revenue per road user, but on the other hand decreases the number of road users. In the extreme of a perfectly inelastic demand, additional taxes revenues can be generated without affecting demand, and the sub-goal of externality regulation then becomes completely unimportant (observe, however, that the assumed constancy of λ_p will of course become less realistic as total tax revenues approach infinity). With a relatively elastic demand, however, (when $-D'$ approaches zero from above) a downward adjustment on the standard Pigouvian tax rule is called for, since in that case a lower tax rate is associated with higher revenues. In particular, we find:

$$sign\ \left(r - C^E\right) = sign\ \left(\frac{N \cdot -D'}{C^E} - 1\right) \tag{37}$$

Hence, if either the demand is relatively inelastic or the marginal external cost relatively low, a tax rate exceeding C^E will be found.

It is evident that the four second-best tax rules derived above are more complicated than the first-best rules. Generally, one finds from the literature on second-best taxation increasingly complex policy rules for increasingly imperfect instruments. In other words, when the charging mechanism itself is no longer perfect, it becomes in addition more difficult to apply this instrument in an optimal way given its inherent distortions. Therefore, additional welfare losses, due to a larger probability of not using the instrument in the optimal manner, are likely to reduce the efficiency of the second-best instrument even further as compared to the first-best bench-mark. Apart from the 'information argument' mentioned in Sect. 2.3, also for this reason, therefore, the probability of government failures in addition to market failures thus increases when second-best instruments are used.

It should be emphasised that this phenomenon of more complicated policy rules for imperfect instruments is of course not restricted to tax instruments. For instance, in the context of Fig. 1, if the regulator would aim at accomplishing a reduction in road usage by physical traffic restraints (for instance an odd/even number plate scheme, or car-free Sundays), it would not be optimal to impose a restriction consistent with the optimal reduction in mobility with optimal charges:

$N^0 - N^*$. The reason is that with such a physical measure, it is not at all evident that it is the mobility representing the lowest social benefits, so to speak between N^* and N^0, will be affected. If the restriction is purely non-discriminating between mobility with a higher or lower benefit, one would use the instrument up to the level where the *expected* net benefits (benefits minus private costs) of the marginally affected traffic is equal to the marginal reduction in external costs. This, necessarily implies a smaller second-best optimal reduction than would be realised using the optimal tax instrument.

This brings us to the last point to be mentioned here, namely that in general, the second-best optimal reduction in external costs will be smaller when less perfect instruments are used. The intuition is simple: the marginal social costs of reducing the externalities will be higher, because of the policy-induced distortions. Hence, one will sooner get to the point where the marginal social costs of reducing the external costs becomes equal to the marginal social benefit of doing so (i.e., the marginal social value of reductions in the externality).

4 Conclusion

This article discussed some important issues in the operationalisation of marginal cost pricing in transport. The discussion was mainly directed to road transport, but many of the principles discussed carry over quite easily to other transport modes as well. The main conclusions are as follows.

Marginal external cost pricing is a first-best bench-mark policy, because it simultaneously provides optimal incentives both in the short run (that is, given the shape and position of the relevant cost and demand functions) and – probably even more importantly – also in the long run, by optimally affecting those factors that determine the shape and position of the relevant demand and cost functions. However, this bench-mark policy is hard to implement in reality, because of a variety of technical, political, social, psychological and institutional barriers. Realistic second-best alternatives will normally only cover parts of the first-best incentives, and will therefore often have to be combined in packages, such that the complete range of incentives is eventually covered. This normally involves instruments covering short-run behaviour, long-run demand factors, and long-run supply-side related factors.

Apart from the increased informational needs implied for the regulator, second-best instruments also require the application of second-best policy and tax rules in order to be used optimally, which are usually far more complex than the standard first-best Pigouvian rule, in which the regulatory tax is equated to the marginal external costs. For both reasons, therefore, there is a large risk of additional government failures, adding to unavoidable welfare losses arising from the second-best nature of the instruments themselves. Therefore, the first-best bench-mark should not be ignored in the process of policy making, for the reason that it is ‘only a hypothetical policy’. Instead, this aticle has made a strong case for using it as a focal point in the design of policy packages.

References

Arnott RJ (1979) Unpriced transport congestion. *Journal of Economic Theory* 21: 294–316

Arnott R, de Palma A, Lindsey R (1990) Economics of a bottleneck. *Journal of Urban Economics* 27: 11–30

Baumol WJ, Oates WE (1988) *The theory of environmental policy*. 2nd ed. Cambridge University Press, Cambridge

Bovenberg AL, De Mooij RA (1994) Do environmental taxes yield a double dividend? *American Economic Review* 84: 1085–1089

Bovenberg AL, Goulder LH (1996) Optimal environmental taxation in the presence of other taxes: general equilibrium analysis. *American Economic Review* 86: 985–1000

Braid RM (1989) Uniform versus peak-load pricing of a bottleneck with elastic demand. *Journal of Urban Economics* 26: 320–327

Braid RM (1996) Peak-load pricing of a transportation route with an unpriced substitute. *Journal of Urban Economics* 40: 179–197

Button KJ, Verhoef ET (1998) *Road pricing, traffic congestion and the environment: Issues of efficiency and social feasibility*. Edward Elgar, Aldershot

Coase RH (1960) The problem of social cost. *Journal of Law and Economics* 3 (Oct.): 1–44

Dupuit J (1844) On the measurement of the utility of public works. In: Murphy D (ed.) (1968) *Transport*. Penguin, London

EC (1995) *Green paper towards fair and efficient pricing in transport: Policy options for internalising the external costs of transport in the European Union*. Commission of the European Communities, Directorate-General for Transport, Brussels

Glazer A, Niskanen E (1992) Parking fees and congestion. *Regional Science and Urban Economics* 22: 123–132

Henderson JV (1977) *Economic theory and the cities*. Academic Press, New York

Johansson-Stenman O, Sterner T (1998) What is the scope for environmental road pricing? In: Button KJ, Verhoef ET (1998) *Road pricing, traffic congestion and the Environment: Issues of efficiency and social feasibility*. Edward Elgar, Aldershot

Knight FH (1924) Some fallacies in the interpretation of social cost. *Quarterly Journal of Economics* 38: 582–606

Laih C-H (1994) Queuing at a bottleneck with single- and multi-step tolls. *Transportation Research* 28A: 197–208

Lévy-Lambert H (1968) Tarification des services à qualité variable: application aux péages de circulation *Econometrica* 36 (3–4): 564–574

Liu LN, McDonald JF (1998) Efficient congestion tolls in the presence of unpriced congestion: a peak and off-peak simulation model. *Journal of Urban Economics* 44: 352–366

Liu LN, McDonald JF (1999) Economic efficiency of second-best congestion pricing schemes in urban highway systems. *Transportation Research* 33B: 157–188

Marchand M (1968) A note on optimal tolls in an imperfect environment. *Econometrica* 36 (3–4): 575–581

Mishan EJ (1971) The postwar literature on externalities: an interpretative essay. *Journal of Economic Literature* 9: 1–28

Mohring H (1972) Optimisation and scale economies in urban bus transportation. *American Economic Review* 62: 591–604

Mohring H (1989) The role of fuel taxes in controlling congestion. *Transport Policy, Management and Technology Towards 2001: Proceedings of the Fifth World Conference on Transport Research* (Yokohama) 1: 243–257

Mohring H, Harwitz M (1962) *Highway Benefits*. Northwestern University Press, Evanston Il

Ochelen S, Proost S, Van Dender K (1998) *Optimal pricing for urban road transport externalities*. Mimeo, Centre for Economic Studies, KULeuven, Leuven

Oron Y, Pines D, Sheshinski E (1973) Optimum *vs* equilibrium land use pattern and congestion toll. *Bell Journal of Economics* 4: 619–636

d'Ouville EL, McDonald JF (1990) Optimal road capacity with a suboptimal congestion toll. *Journal of Urban Economics* 28: 34–49

Pigou AC (1920) *Wealth and welfare*. Macmillan, London

Small KA, Gomez-Ibañez JA (1998) Road pricing for congestion management: the transition from theory to policy. In: Button KJ, Verhoef ET (1998). *Road pricing, traffic congestion and the environment: Issues of efficiency and social feasibility*. Edward Elgar, Aldershot

Spulber DF (1985) Effluent regulation and long-run optimality. *Journal of Environmental Economics and Management* 12: 103–116

Sullivan AM (1983) Second-best policies for congestion externalities. *Journal of Urban Economics* 14: 105–123

Varian HR (1992) *Microeconomic analysis*, 3rd ed. Norton, New York

Verhoef ET (1996) *The economics of regulating road transport*. Edward Elgar, Cheltenham

Verhoef ET (1998) *Second-best congestion pricing in general static transportation networks with elastic demands.* Mimeo, Free University Amsterdam

Verhoef ET, Nijkamp P, Rietveld P (1995) The economics of regulatory parking policies: the (im-)possibilities of parking policies in traffic regulation. *Transportation Research* 29A (2): 141–156

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

Verhoef ET, van den Bergh JCJM, Button KJ (1997) Transport, spatial economy and the global environment. *Environment and Planning* 29B: 1195–1213

Walters AA (1961) The theory and measurement of private and social cost of highway congestion. *Econometrica* 29 (4): 676–697

Wilson JD (1983) Optimal road capacity in the presence of unpriced congestion. *Journal of Urban Economics* 13: 337–357