Rouwendal, Jan; Verhoef, Erik T.

Working Paper

Second-best Pricing for Imperfect Substitutes in Urban Networks

Tinbergen Institute Discussion Paper, No. 03-085/3

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Rouwendal, Jan; Verhoef, Erik T. (2003) : Second-best Pricing for Imperfect Substitutes in Urban Networks, Tinbergen Institute Discussion Paper, No. 03-085/3, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/86089

TI 2003-085/3
Tinbergen Institute Discussion Paper

Second-best Pricing for Imperfect Substitutes in Urban Networks

Jan Rouwendal
Erik T. Verhoef

Department of Spatial Economics, Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam, and Tinbergen Institute.

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

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

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

Please send questions and/or remarks of non-scientific nature to driessen@tinbergen.nl.
Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

Second-best pricing for imperfect substitutes in urban networks

Jan Rouwendal[1,2] **and Erik T. Verhoef**[2]

Department of Spatial Economics
Free University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands

JEL Classification codes: R41, R48, D62.

Keywords: traffic congestion, road pricing, second best policies, two mode problem

Abstract

This paper considers second-best pricing as it arises through incomplete coverage of full networks. The main principles are first reviewed by considering the classic two-route problem and some extensions that have been studied more recently. In most of these studies the competing routes are assumed to be perfect substitutes, which is probably not the case for most parallel roads in reality, and even less likely for the case where competing connections represent different transport modes. In this paper a modelling framework in which the alternatives are imperfect substitutes is developed and numerical results for two roads and two modes are presented. In the model, trip generation and trip distribution are distinguished in a way that is consistent with economic theory. The model is used to consider situations in which one route or mode cannot be tolled. Simulation results show that, for the chosen parameter values, there is a substantial difference between the effectiveness of policies in which the capacities have to be taken as given, and those in which capacity of at least one mode can be changed. A striking feature of the policy in which the capacities of both modes/routes and the railway fare/toll on one road can be used as policy instruments is the existence of two equilibria for a range of values of β. In one equilibrium there are substantial numbers of users of both modes, whereas in the other use of one mode is negligible.

[1] Also at Wageningen University. Economics Of Consumers and Households, Hollandseweg 1, 6706 KN Wageningen.
[2] Affiliated to the Tibergen Institute, Roeterstraat 31, 1018 WB Amsterdam.

INTRODUCTION

The basic economic intuition behind marginal cost pricing is deceptively simple. When prices are equal to marginal social costs, consumers would expand consumption (be it road trips or any other good) up to the point where the benefits of the last unit consumed have become equal to these marginal social costs. The implied equality of marginal benefits and marginal social costs secures that the social surplus, defined as the difference between total benefits and total costs, is maximised. This social surplus is often considered as an appropriate indicator for social welfare, and its maximisation is a condition for economic efficiency to prevail.

As explained in Chapter 1 of this volume, the 'price' for the use of a road would in the first place include the costs borne directly by the road user, such as fuel expenses and the value of the travel time spent on the road – the marginal private costs. Besides, a toll may be charged for the use of the road, which also adds to the price. Marginal cost pricing in road transport therefore means that the toll should reflect the marginal external costs: the marginal costs of road use other than the marginal private cost. For a congested road, this includes the value of time losses (and travel time uncertainty) imposed on other road users, as well as the value of emissions, noise, and accident risks created. Such a tax, equal to marginal external costs, is often referred to as a 'Pigouvian tax', after its spiritual father Arthur Pigou (1920).

Based on this intuition, transport economists and engineers alike have for long advocated the use of road pricing schemes for the management of road traffic congestion (*e.g.* Pigou, 1920; Knight, 1924; Walters, 1961; Vickrey, 1963; *etc.*). Even if road capacities can be adjusted to cope with congestion, capacity policies alone will typically not be sufficient to reach an optimal outcome, and the use of congestion

pricing remains warranted.[1] As pricing not only succeeds in reducing demand *per se*, but in doing so also in maintaining exactly those trips for which the benefits, as reflected by the willingness to pay, are highest, it will typically lead to higher welfare gains than what could be realised by alternative, non-price-based demand management schemes, such as number plate policies as currently in operation in Athens and Mexico City. This explains much of the analysts' preference for this type of policy. Moreover, the principle of marginal cost pricing securing efficiency carries over to more complex situations, including congestion pricing on full networks (rather than a single road), heterogeneous users, and time-varying congestion. This suggests that the practical application of the principle of optimal congestion pricing in reality, although yielding additional practical complications, would not create additional conceptual difficulties, compared to the basic mechanisms as identified in textbook models. However, this would – unfortunately – be true only under rather stringent and often unrealistic assumptions. Two of these unrealistic assumptions are discussed next.

A first one concerns optimality of the pricing instrument, which should allow the road regulator to differentiate prices such that every individual road user indeed faces a toll exactly equal to the marginal external costs he or she causes. This requires tolls to be differentiated, at least, by time of day, route followed and hence the length of the trip, type of vehicle (*e.g.*, small versus large cars or passenger cars versus trucks) and its state of maintenance (for pollution), driving style (for pollution, noise and accident risks), *etc.* It also requires that these issues can be monitored perfectly by the regulator. Although emerging technologies for electronic charging are likely to allow

[1] Note, however, that the reverse is also true: when capacity can be chosen, pricing alone will typically not result in an optimal outcome unless initial capacities 'happen' to be optimal.

for a more sophisticated toll differentiation in the near future, it remains to be seen to which extent such complex pricing will indeed be put into practice.

A second assumption concerns efficiency in all markets related to the transport market under consideration, where a market is 'related' to the transport market if its equilibrium is indirectly affected by transport policies, and its equilibrium is inefficient if, again, prices are not equal to marginal social costs. The existence of such related markets is the rule rather than the exception because transport demand is often a derived demand, where the 'consumption' of transport is often no goal in itself but serves to enable a market transaction between spatially separated suppliers and demanders of a certain good or service. For example, commuting typically does not yield any benefit in itself, but serves to allow people to supply labour services at a different location, the work place, than their residences. Likewise, freight traffic enables a transaction between a supplier of goods and demanders, final consumers or firms demanding intermediates, who are separated in space. Whenever these related markets do not function properly, the simple policy rule of setting taxes equal to marginal external costs will no longer be a truly optimal choice. Instead, transport taxes should optimally deviate from this rule, so as to correct for the inefficiency in the related market without of course sacrificing too much efficiency in the transport market itself. Important reasons why prices in related markets may deviate from marginal social costs include, amongst others, the existence of market power, unpriced externalities (such as environmental pollution), or distortionary taxes (for instance on the labour market in the case of commuter traffic). Inefficiencies are therefore likely to be the rule rather than the exception.

Despite the attention that first-best pricing has traditionally received in most of the literature on transport pricing, it will be clear from the above that second-best pricing

is in fact often the most relevant case from a practical perspective. In recent years, there has consequently been an upsurge of studies in second-best congestion pricing. Lindsey and Verhoef (2001) provide a recent review of this literature, and discuss second-best pricing resulting from network issues, heterogeneity of users, dynamic constraints on toll flexibility ('step tolls'), uncertainty in travel times (in relation with real time information provision), relations with other sectors (notably labour markets), the simultaneous existence of multiple externalities (*e.g.* congestion and pollution), congestion pricing by private road operators, and through interactions with sub-optimal capacity choice.

Second-best congestion pricing through network effects will be especially important in urban contexts, where dense networks will often prevent the application of optimally differentiated tolls on each and every link of the network. This chapter considers these issues. A first aim is to provide an overview of the basic economic ins and outs of second-best congestion pricing in networks by reviewing the classic two-route problem in Section 2 below. Most of the underlying literature conveniently, but somewhat unrealistically, treats different routes between two points ('nodes') in a network as pure substitutes, and therefore applies so-called Wardropian equilibrium principles (Wardrop, 1952) in the determination of network equilibria (with and without prices). This means that even the smallest equilibrium price difference would make a user choose the cheapest route. In reality, people may have so-called idiosyncratic preferences over different routes, meaning that they would have different preferences for competing routes when these would have equal trip prices. It is therefore of interest to study how this would affect the optimal design and welfare properties of second-best congestion pricing. Moreover, by allowing different routes to be imperfect substitutes, a framework is developed that can also be used to study

second-best congestion pricing from a multi-modal perspective. Addressing these issues is the second aim of the paper. A theoretical model is developed for the study of second-best congestion pricing in networks where parallel roads are imperfect substitute routes, and its properties are illustrated with the help of a simulation model.

2.1 SECOND-BEST CONGESTION PRICING IN TRANSPORT NETWORKS

One of the most widely studied instances of second-best congestion pricing concerns the case where not all links in a uni-modal transport network can be priced. This case has great practical relevance for actual policy making. For instance, pay-lanes as currently in operation at various places in the USA belong to this category of second-best pricing policies, since unpriced lanes are available as a direct substitute for the pay-lanes. In other (planned) road pricing schemes, be it based on marginal cost pricing principles or motivated by the desire to generate revenues, prices will often be charged on a limited number of links in a network only (*e.g.* toll roads, or toll cordons). Furthermore, the case is of interest as it can be used to illustrate some more general principles of second-best pricing that will be relevant for other types of second-best pricing too. These principles include the fact that in second-best pricing, the naïve use of Pigouvian taxes, set equal to the direct marginal external costs, will generally not be optimal, and the fact that second-best pricing itself creates distortions, which affect the optimal second-best price itself.

2.1 The classic two-route problem

The simplest version of the problem at hand concerns the two-route problem, where an untolled alternative road is available parallel to a toll road. This problem has for instance been studied by Lévy-Lambert (1968), Marchand (1968), and more recently

also by Braid (1996), Verhoef *et al* (1996), and Liu and McDonald (1999). Figure 1 illustrates the basic set-up of the problem. Two parallel, congested roads of given capacities connect an origin (*A*) and a destination (*B*). On one of these (road *T*), a congestion toll can be set, whereas the other road (*U*) remains untolled. The roads are pure substitutes, meaning that equilibrium 'generalised prices' p (the sum of monetised travel costs, c, plus a toll, τ, if levied) should be equal for both roads when both are used (this is the 'Wardropian' equilibrium principle mentioned earlier). Drivers are nearly identical (they have the same value of time, drive the same type of cars, *etc.*), with the exception of their willingness to pay for a trip, which varies over drivers whenever demand is not perfectly elastic. The two central questions are then: at what level should the second-best toll be set, and what are the welfare effects compared to the first-best option of tolling both roads?

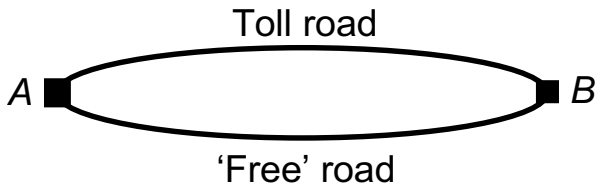


Figure 1 The classic two-route problem

An important result from the studies looking at these problems is that the second-best tax, set so as to maximise social welfare given the persistence of the second-best distortion of leaving the other road untolled, is typically below the optimal Pigouvian tax that would be set on the same road under first-best pricing, but also below the marginal external costs on the toll road in the second-best optimum. In addition, the second-best toll yields considerably smaller welfare gains than first-best pricing.

It is important to understand why the second-best toll should ideally be below the marginal external congestion costs on the tolled road. The reason is that the congestion charge on the toll road brings both good and bad news from the

perspective of efficiency. The good news is that initially excessive free-market congestion on the toll road is reduced through the toll. The bad news is that some people who are priced off this road will be diverted to the already congested untolled road, thus aggravating the congestion there. The second-best toll trades off the good news against the bad news. The good news alone would ask for a tax equal to marginal external costs for the reasons given in the introduction, but as soon as the bad news is relevant (*i.e.*, when there is congestion on the untolled road, and when a by-product of the toll is indeed to divert traffic to this untolled road), a downward adjustment of the toll is optimal. As a corollary, a policy of simply ignoring the second-best distortion, and setting the toll equal to the marginal external congestion costs on the tolled road, would lead to a social surplus below that in the second-best optimum, exactly because the policy is 'naïve' and ignores the spill-overs on the free lane. Indeed, such 'quasi first-best pricing' may even lead to a welfare loss compared to the no-toll outcome, which would certainly occur when the second-best optimal tax is negative, as in some examples in Verhoef *et al* (1996).

The second-best optimal tax rule for this classic two-route problem reflects the above mentioned trade-off between the good and the bad news from the toll. Verhoef *et al* (1996) express it as:

$$\tau = mec_T - mec_U \cdot \frac{-D'}{c_U' - D'}$$

where mec_T (mec_U) denotes the marginal external costs on the tolled (untolled) road, D' the slope of the inverse demand function, and c_U' that of the cost function on the untolled road; all evaluated at the second-best equilibrium. The tax rule clearly shows how the second-best toll is equal to the what will be called the 'direct marginal

external costs' of drivers road T (mec_T),[2] minus a fraction of those on the untolled road U (mec_U) (because $D' \leq 0$ and $c_U' \geq 0$, the weight for mec_U is always between 0 and 1).

Further intuition can be obtained by considering a few instances where the fraction mentioned would approach its possible extreme values of 0 (meaning that $\tau = mec_T$) or 1 (meaning that $\tau = mec_T - mec_U$). Intuitively, the fraction should approach zero when there is no diverting of traffic to the untolled road in the second-best optimum due to marginal changes in τ. This is the case when either $D' \to 0$ (regardless of τ, the untolled road will in equilibrium be filled with drivers up to that level N_U that equates $c_U(N_U)$ to the given value of D), or when $c_U' \to \infty$ (the cost function for road U is perfectly inelastic, and N_U is given). In these cases, the regulator can ignore spill-overs of tolling on road T upon road U, simply because these will not occur. The fraction, in contrast, approaches unity when either $D' \to -\infty$ or when $c_U' \to 0$. In these cases, the second-best toll only tries to optimise route split, by internalizing the difference between marginal external costs for the two routes, and fully ignores the effects on overall demand. This makes sense when demand is perfectly inelastic ($D' \to -\infty$), in which case overall demand would not respond to marginal changes in τ. It also makes sense if every user priced off road T immediately switches to road U, in the situation where $c_U' \to 0$ implies that private costs of using road U are independent of the route split, so that the overall demand is given and defined by the equality of D and c_U. It should be noted, however, that $c_U' \to 0$ at the

[2] The imperfection of the tax instrument causes the term 'marginal external costs' to become ambiguous, as a distinction between 'full' and 'direct' marginal external costs can be made. The full marginal external costs for drivers on the tolled road in this example are given by the tax rule. These include the second term, which is non-zero because of the imperfection of the tax instrument. Direct marginal external costs are thus defined here as all marginal external costs that are not caused by inefficient behaviour induced by the tax instrument itself.

same time implies that $mec_U = 0$, so that the tax rule in this case effectively becomes $\tau = mec_T$. Second-best pricing then becomes identical to first-best pricing because spill-overs do not result in additional welfare losses.

The relative size of the welfare losses from second-best pricing compared to first-best pricing depends on various elasticities. Verhoef *et al.* (1996) for example demonstrate that if route *U* is particularly congested, the optimal second-best toll can be negative, which typically implies relatively low efficiency gains due to pricing (even zero when the optimal second-best congestion toll happens to be zero). More generally they show that the efficiency gains depend on the relative free-flow travel times and capacities of the two routes, and increase if the tolled road becomes relatively more attractive. Furthermore, the gains depend on the price elasticity of travel demand, in combination with the cost parameters. Specifically, if demand approaches perfect inelasticity, the efficiency gains from second-best pricing may approach those from first-best pricing if the roads differ in free-flow travel times. The only factor determining efficiency in that case would be the route split, which can be optimised with one toll only. However, if the roads have identical free-flow travel times, as would often be the case for pay-lanes, the efficiency gains from second-best pricing may nevertheless fall to zero, as the free-market route split (without tolling) may then already be optimal. Equalisation of private costs then implies the optimal equalisation of marginal costs. The second-best optimal toll approaches zero and therefore induces no welfare effects whatsoever. These findings illustrate that for second-best pricing often 'anything goes', in the sense that the relative efficiency of second-best policies is often impossible to predict without prior detailed knowledge of the case at hand. Liu and McDonald (1998) constructed a model descriptive of one of the California road pricing demonstration projects (State Route 91 in Orange County),

and found that the efficiency gains from second-best congestion pricing are in the order of 10 per cent of the potential gains from first-best pricing. Bearing in mind the exposition above, this should in fact not be too disappointing, given the similarity of the pay-lanes and the free lanes, and the inelasticity of demand.

2.2 Beyond the classic two-route problem

The classic two-route problem has received considerable attention for at least two reasons. Firstly, it is probably the simplest possible and most transparent setting for illustrating some of the main economic insights into the ins and outs of second-best pricing policies, which often also carry over to other, more complicated settings.[3] Secondly, it appears to be a reasonable approximation for many congestion pricing demonstration projects that are currently in operation. The basic model is certainly rather restrictive, and a number of extensions have been made to investigate some further properties of the problem and its solution. For example, static models, as discussed above, may underestimate the efficiency gains from second-best one-route tolling because they ignore some alternative ways in which second-best pricing can affect driver behaviour. Braid (1996) and De Palma and Lindsey (2000) allow for trip-timing adjustments by considering time-varying tolls in the Vickrey (1969) bottleneck model, applied to the same two-parallel-routes network. They find that second-best tolling yields higher absolute efficiency gains than in the static model, and a greater fraction of the first-best efficiency gains. This is because the toll not only curbs excessive total usage, but also eliminates queuing on the tolled route.

[3] Under second-best pricing, it is no longer optimal to equate taxes to the direct marginal external costs, but instead more complicated tax rules apply, which aim to partly correct for the relevant distortions, and naïve Pigouvian taxation is therefore no longer optimal. In addition to that, the relative efficiency of a given type of second-best pricing often depends rather strongly on case-specific circumstances.

The efficiency of second-best tolling may also be improved when drivers have different values of travel time. Users with a high value of time would then choose the tolled connection and benefit relatively strongly from the reduced travel time, while those with a low value of time would not suffer too much from the increased travel on the untolled connection that they prefer. Small and Yan (1999) and Verhoef and Small (1999) find that the efficiency of pay-lanes relative to the first-best optimum is indeed higher than in the equivalent model with no heterogeneity in the value of time.

Finally, optimal second-best tolls and the associated welfare gains have also been investigated for larger networks on which not all links can be tolled. Verhoef (1998) derives optimal static tolls on any subset of links (including parking spaces) in an arbitrary network. The toll formulae, which are quite complicated, include terms reflecting marginal external costs on other links, and weights that depend on various demand and cost elasticities. The solution for the classic two-route problem presented above is a special – but transparent – illustrative case of these general formulae.

2.2 THE TWO-ROUTE PROBLEM WITH IMPERFECT SUBSTITUTES

In this section the network of Figure 1 is further discussed. As announced, this network will be reconsidered for the case where the two routes are imperfect substitutes. There are, as before, two nodes, A and B, connected by two links, which are now denoted 1 and 2. Besides tolls, cases where capacities on these links can vary will also be considered, and second-best issues arise if at least one of these four policy variables is restricted.

The links may refer to roads, as in the classic two-route problem, and second-best problems emerge, for instance, if only one link is unpriced or if it is priced sub-optimally. The links may also represent different modes, which is what will be

considered in Section 4. For competing modes it is common to regard these as imperfect substitutes. In addition, the type of cost functions for users, as well as the cost of providing capacity for operators, may then be different between the two links. Arnott and Yan (2000) have recently provided a review of the problems that arise if travel with one mode is subsidised, while the policy maker can choose the price of travelling on the other mode and capacity on both links. This 'two-mode problem' appears to be hard to solve in general terms.

If two roads are perfect substitutes, any price difference implies that one route will not be used. However, if they are imperfect substitutes, price differences do not necessarily imply that demand for the most expensive route falls to 0. In reality, two roads that connect the same pair of towns need not necessarily be considered as perfect substitutes. For instance, consumers may prefer one route to the other because their residential location in A is closer to the point where route 1 starts, because their destination in B is closer to where route 2 ends, or because they have a preference over highways (easier driving) or secondary roads (nicer scenery) *per se*. The random term in the logit model presented below may be interpreted as reflecting such idiosyncratic differences.

3.1 The general model

In what follows a model that can cover all the aforementioned cases is constructed. In the general model the social planner attempts to maximise the social surplus, which is defined as consumer surplus plus toll revenues minus cost of infrastructure capacity:

$$SS = CS + TR - K \qquad (1)$$

The Marshallian notion of consumer surplus is used. For a single good, it is the area under the inverse demand curve and above the prevailing price, between the origin and the equilibrium quantity. If two (or more) goods are involved, the total consumer

surplus is not usually the sum of the consumer surpluses for the individual goods as they could be calculated on the basis of the two (or more) demand functions. The reason is that the demand curves are interrelated, and when the price of one good is increased, the demand curves for the other good(s) and the associated consumer surplus(es) change(s). However, total surplus can still be defined unambiguously if the demand functions are independent of income, and that is what will be assumed here. The situation with two goods, indicated by suffixes 1 and 2, is considered. The quantities q demanded can be derived from the consumer surplus by taking the first derivatives and putting a minus sign in front:

$$q_i = -\frac{\partial CS}{\partial p_i}, \quad i = 1,2. \tag{2}$$

Tolls are denoted τ, and have to be multiplied by the number of trips in order to find toll revenues:

$$\begin{aligned} TR &= TR_1 + TR_2 \\ &= \tau_1 q_1 + \tau_2 q_2 \end{aligned} \tag{3}$$

Cost depends on capacity:

$$K = K_1(cap_1) + K_2(cap_2) \tag{4}$$

Constant returns to scale are not imposed. Capacity is allowed to be vector valued (e.g. the frequency of trains and the number of seats per train may be relevant dimensions of capacity), but it will usually be assumed to be a scalar (e.g. road width).

Only cases in which a user equilibrium is obtained will be considered. User equilibrium is found where the inverse demand, as represented by the full price p for both routes, is equal to the sum of private travel cost c, which depends on capacity and number of trips, and toll τ:

$$p_i = c_i(cap_i, q_i) + \tau_i, \quad i = 1,2. \tag{5}$$

The first-best problem is the maximisation of social surplus by choosing the two tolls and capacities under the side condition that a user equilibrium should be obtained. The Lagrangian is:

$$\begin{aligned} L &= SS + \lambda_1(p_1 - c_1 - \tau_1) + \lambda_2(p_2 - c_2 - \tau_2) \\ &= \sum_i \int_{p_i}^{\infty} q_i(x)\mathrm{d}x + \tau_i q_i(p_i) - K_i(cap_i) + \lambda_i(p_i - c_i(q_i(p_i), cap_i) - \tau_i) \end{aligned} \tag{6}$$

The first order conditions with respect to p_1, p_2, τ_1, τ_2 and cap_1 and cap_2 are:

$$\begin{aligned} &-q_1 + \tau_1 \frac{\partial q_1}{\partial p_1} + \tau_2 \frac{\partial q_2}{\partial p_1} + \lambda_1\left(1 - \frac{\partial c_1}{\partial q_1}\frac{\partial q_1}{\partial p_1}\right) + \lambda_2\left(-\frac{\partial c_2}{\partial q_2}\frac{\partial q_2}{\partial p_1}\right) = 0 \\ &-q_2 + \tau_1 \frac{\partial q_1}{\partial p_2} + \tau_2 \frac{\partial q_2}{\partial p_2} + \lambda_1\left(-\frac{\partial c_1}{\partial q_1}\frac{\partial q_1}{\partial p_2}\right) + \lambda_2\left(1 - \frac{\partial c_2}{\partial q_2}\frac{\partial q_2}{\partial p_2}\right) = 0 \end{aligned} \tag{7}$$

$$q_i = \lambda_i, \quad i = 1,2. \tag{8}$$

$$\frac{\partial K_{i1}}{\partial cap_i} = -\lambda_i \frac{\partial c_i}{\partial cap_i}, \quad i = 1,2. \tag{9}$$

In the first-best optimum, all conditions (5) and (7)-(9) are satisfied. In second-best situations one or more of the four equations (8) and (9) are not satisfied. Equations (5) and (7) are always valid, since they are the consequence of the basic requirement that there should be a user equilibrium.

Equations (8) are the requirements for optimal tolling. They state that under optimal tolling the Lagrange multipliers will be equal to the associated numbers of trips.[4] If equations (8) are valid, the Lagrange multipliers can easily be eliminated

[4] In transportation economics often an alternative (but equivalent) formulation of consumer surplus is used, based on the inverse demand function. In that case, partial derivatives of the Lagrange function with respect to the volumes of trips have to be used, and the Lagrange multipliers are then equal to zero under optimal tolling. In the present context, with possibly imperfect substitutability between the two routes/modes, the formulation based on the 'ordinary' demand function is more convenient.

from the system of equations. This implies a substantial simplification. Equations (7) can then be written as:

$$\begin{aligned}\left(\tau_1 - q_1 \frac{\partial c_1}{\partial q_1}\right)\frac{\partial q_1}{\partial p_1} + \left(\tau_2 - q_2 \frac{\partial c_2}{\partial q_1}\right)\frac{\partial q_2}{\partial p_1} = 0 \\ \left(\tau_1 - q_1 \frac{\partial c_1}{\partial q_1}\right)\frac{\partial q_1}{\partial p_2} + \left(\tau_2 - q_2 \frac{\partial c_2}{\partial q_1}\right)\frac{\partial q_2}{\partial p_2} = 0\end{aligned} \tag{10}$$

This implies that both tolls should be equal to the external congestion effect.[5]

Equations (9) become:

$$\begin{aligned}\frac{\partial K_1}{\partial cap_1} = -q_1 \frac{\partial c_1}{\partial cap_1} \\ \frac{\partial K_2}{\partial cap_2} = -q_2 \frac{\partial c_2}{\partial cap_2}\end{aligned} \tag{11}$$

Equation (11) states that the marginal cost of capacity should be equal to the marginal benefit in the form of reduced travel costs.

It is standard to assume that private travel costs are homogeneous of degree zero in capacity and number of users. This implies:

$$q_i \frac{\partial c_i}{\partial q_i} + cap_i \frac{\partial c_i}{\partial cap_i} = 0, \quad i = 1,2. \tag{12}$$

Using Equation (12) the first order conditions with respect to capacity can be rewritten as:

$$cap_i \frac{\partial K_i}{\partial cap_i} = \lambda_i q_i \frac{\partial c_i}{\partial q_i}, \quad i = 1,2. \tag{13}$$

If there are constant returns to scale, $\partial K_i / \partial cap_i$ is a constant, k_i, and the left hand side of Equation (10) is equal to the total cost of capacity of link *i*. The right hand side is equal to λ_i times the external (congestion) effect of traffic on link *i*. In the first-best

[5] If the matrix of partial derivatives $\partial q_i/\partial p_j$ has an inverse this is the unique solution. The matrix has no inverse if total demand is fixed or if the shares of the two routes/modes are fixed.

equilibrium the Lagrange multiplier equals the number of users and the cost of capacity is therefore equal to optimal toll revenues. This self-financing result was first obtained by Mohring and Harwitz (1962) for a single link. Yang and Meng (2002) recently showed that it also holds in a general network. Homogeneity of degree 0 of the user cost function will not always be imposed: the function used for public transport in the two mode model does not satisfy this requirement.

3.2 Trip generation and trip distribution

A distinction between trip generation and trip distribution is often made in transportation planning models. It is convenient because it allows for a decomposition of the total planning problem in parts that can be treated relatively independent of each other and it is therefore useful to introduce it in economic models as well. In order to introduce this distinction in a way that is consistent with economic theory, assume that consumer's surplus can be written as a function of a composite price P of the prices of the two transport services:

$$CS(p_1, p_2) = CS^*(P) \tag{14}$$

with

$$P = P(p_1, p_2) \tag{15}$$

P can be interpreted as a composite price for transport services and equation (14) states that the consumer surplus derived from both transport services can be written as the consumer's surplus of a single transport service.

It can then be easily verified, using Equation (2), that for the sum of the demand for both links, $Q = q_1 + q_2$:

$$Q = -\frac{\partial CS^*}{\partial P}\left(\frac{\partial P}{\partial p_1} + \frac{\partial P}{\partial p_2}\right) \tag{16}$$

whereas the distribution shares *s* of the trips over the two modes are:

$$s_i = \frac{q_i}{Q} = \frac{\partial P}{\partial p_i} \Bigg/ \left(\frac{\partial P}{\partial p_1} + \frac{\partial P}{\partial p_2} \right), \quad i = 1,2. \tag{17}$$

Considerable simplification of the latter two equations is possible if it is assumed that the partial derivatives of *P* add up to 1:

$$\frac{\partial P}{\partial p_1} + \frac{\partial P}{\partial p_2} = 1 \tag{18}$$

According to Equation (18) a small change *dp* in the prices of trips on the two roads leads to an identical change in the composite price of transport on both routes, which is intuitive.

Equation (16) refers to *trip generation*, and Equation (17), to *modal split* or *route choice*. These equations show that, when both assumptions are made, trip generation does not depend on the specification of *P*, whereas trip distribution does not depend on the specification of CS^*. Specification (14) with *P* satisfying (18) is therefore consistent with the use of different sub-models for these two aspects of travel demand, which are relatively independent of each other.

The effect of an increase in the price of one route or mode can be decomposed into a trip generation effect and a trip distribution effect as follows:

$$\begin{aligned} \frac{\partial q_i}{\partial p_j} \left(= \frac{\partial (Q\, s_i)}{\partial p_j} \right) &= s_i \frac{\partial Q}{\partial p_j} + Q \frac{\partial s_i}{\partial p_j} \\ &= s_i s_j \frac{\partial Q}{\partial P} + Q \frac{\partial s_i}{\partial p_j} \end{aligned} \tag{19}$$

This equation holds for *i,j*=1,2.

The logit model, which is a standard specification for mode choice, results if the following P is chosen:[6]

$$P = \ln\left(\frac{\alpha_1 e^{\beta p_1} + \alpha_2 e^{\beta p_2}}{\alpha_1 + \alpha_2}\right)^{\frac{1}{\beta}} \tag{20}$$

with $\beta<0$ and α_1, $\alpha_2>0$. The shares can be determined using Equation (17) as:

$$s_i = \frac{\alpha_i e^{\beta p_i}}{\alpha_1 e^{\beta p_1} + \alpha_2 e^{\beta p_2}} \tag{21}$$

α_i can be rewritten as $\exp(\alpha_i^*)$ with $\alpha_i^*=\ln(\alpha_i)$ and it is possible that α_i^* depends on mode or route characteristics, as is usual in empirical logit models used to study trip distribution.[7]

In the limiting case in which $\beta=0$, link choice is independent of the prices. It can be verified from Equation (20) that if $p_1=p_2=p$, P is also equal to p, that P is increasing in both prices, and that P always takes on a value in-between that of the two prices. This motivates the interpretation of P as a composite price for transport on the two modes or routes. P is homogeneous of degree 1 in the prices of all goods if p_1 and p_2 are interpreted as relative prices.

The two links of the network are perfect substitutes if the users always choose the one with the lowest cost. In this situation the share of link 1 should be equal to zero if $p_1>p_2$, and equal to 1 in the opposite case, whereas the shares are indeterminate when both prices are equal. If the logit specification for the shares is used, this situation can be approximated arbitrarily closely by choosing taking the limit for $\beta \to -\infty$ of (21). This is easily verified by rewriting Equation (21) for i=1 as:

[6] Equation (21) is not the only possible specification of P that satisfies (18). Any function $P=\ln[g(\exp(p_1),\exp(p_2))]$ with g homogeneous of degree 1 in its two arguments does.
[7] The function P is equal to the sum of a constant ($-\ln(\alpha_1+\alpha_2)$) and the logsum measure that can be used for welfare economic analyses (Small and Rosen, 1981).

$$s_1 = \frac{\alpha_1 e^{\beta(p_1 - p_2)}}{\alpha_1 e^{\beta(p_1 - p_2)} + \alpha_2} \tag{22}$$

and taking the limit for $\beta \to -\infty$. When β equals 0 there is no price sensitivity at all. The model developed in the present subsection seems attractive as the basis of a simulation model. In particular, it offers the possibility to compare how second-best problems change if the two links become closer substitutes.

2.3.3 A simulation model with two routes

The previous two subsections provide the main ingredients for a simulation model for the two route and two mode situations that will be used in the remainder of this section. The two-route case is discussed here and the model is completed by introducing specific forms for the (total) demand and travel time functions.

A quadratic specification for consumer's surplus is adopted as a function of the transport price index *P*:

$$CS^* = -a\,P - .5\,b\,P^2. \tag{23}$$

with *a>0* and *b<0*. Total demand for trips is therefore a linear function of the price of transport:

$$Q = a + b\,P \tag{24}$$

The logit formulation (20) is used for *P*. User cost is given by the commonly used Bureau of Public Roads formulation:[8]

$$c_i = vot\ fftt \left[1 + c_{i1} \left(\frac{q_i}{cap_i} \right)^{c_{ie}} \right] \tag{25}$$

where *vot* denotes the value of time and *fftt* free-flow travel time. This function is homogeneous of degree zero in the number of trips and capacity.

[8] See e.g. Small (1992), p. 70.

For the cost of capacity in the two-route problem the following specification is used:

$$K_i(cap_i) = k_i cap_i \qquad (26)$$

This cost function has constant returns to scale.

Table 1 Parameter values for the two-route problem

Parameter	Route/mode 1	Route/mode 2
Demand		
a	7500	
b	-100	
β	-0.1	
α	1	1
User cost		
vot	7.5	
fftt	0.75	0.5
c_1	0.15	0.15
c_e	4	4
Cost of capacity		
k	6	6

The parameter values used for the two-route problem (see Table 1) refer to a situation in which two cities are connected by roads that differ in capacity and free-flow travel time. The route with the highest free-flow travel time has the lowest cost of capacity per road kilometre. Preferences with respect to the use of the two routes are symmetric, which means that consumers do not systematically prefer one road over the other.

Link 1 has a free-flow travel time of half an hour and link 2 has a free-flow travel time of three-quarters of an hour, and the value of time is set at the Dutch average of €7.5. The BPR parameters c_1 and c_e have their conventional values.

The unit price of capacity was determined as follows. A value of *cap* equal to 1,750 for the BPR cost function implies a doubling of travel times at a flow of around

2,800 vehicles per hour. This is roughly the flow at which, empirically, travel times double for a single highway lane. The hourly unit price of capacity of €6 for link 2, the fast connection, was determined by dividing the estimated average annual capital cost of one highway lane kilometre in The Netherlands (€0.2 million) by 1,100 (220 working days multiplied by 5 peak hours per working day, assuming two peaks) and then by 1,750 (the number of units of capacity of a standard highway lane), and finally by multiplying that result by 60 (the number of kilometres that can be travelled at free-flow speed in half an hour). The price of capacity at link 1 was set at the same value in order to facilitate comparability. Although the road is longer, which might call for a higher *k*, construction may be cheaper, which may call for a lower *k*.

2.3.4 Base case

The base case refers to a user equilibrium without tolls and with capacities as shown in Table 2, with a higher capacity for link 2. The main characteristics of the base case are given in Table 2.

Table 2 The base case of the two-route problem

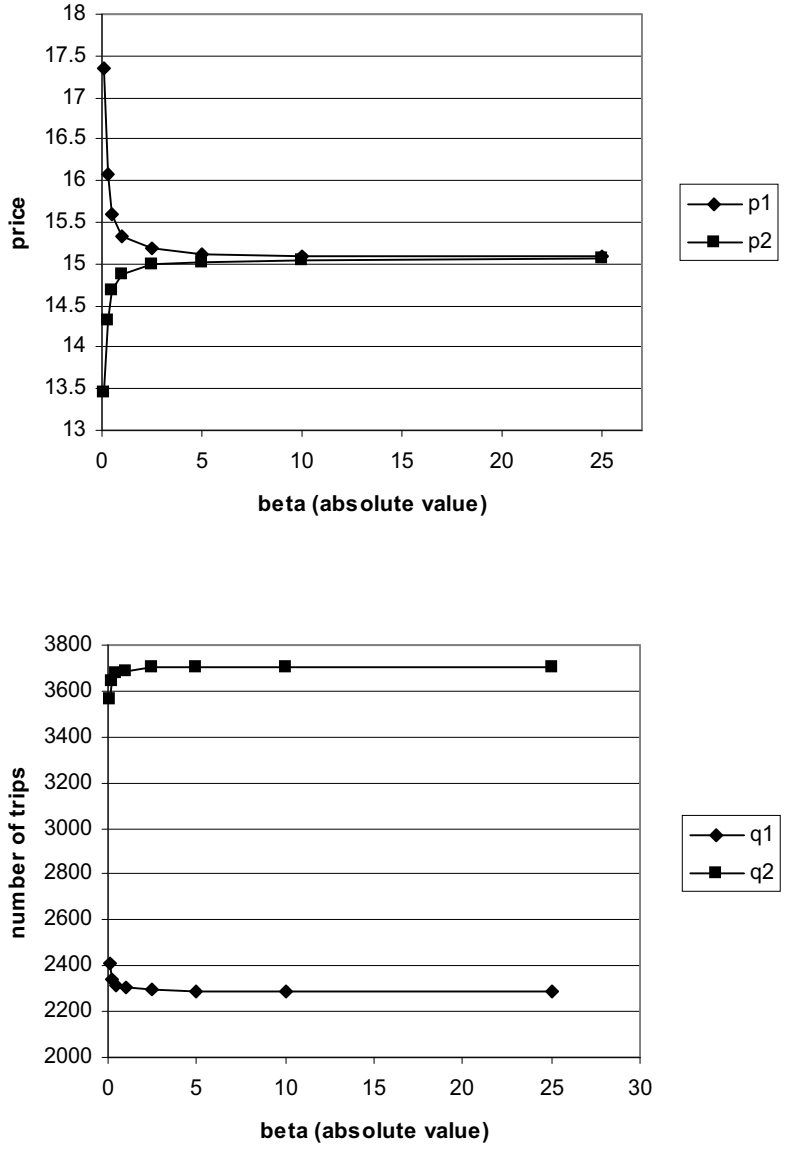
Variable	Route/mode 1		Route/mode 2
cap	1,250		1,750
q	2,413		3,566
Q		5,925	
c (= *p*)	17.35		13.45
P		15.21	
Elasticity of *q* with respect to p_1	-1.15		0.67
Elasticity of *q* with respect to p_2	0.58		-0.68
Elasticity of *Q* with respect to *P*		-0.25	
CS		178,752	
K	7,500		10,500
SS		160,752	

Note: *cap*, *q* and *Q* are measured as numbers of trips per hour, *c*, *p, P, CS, K* and *SS* in euros. Source: own calculations.

It is interesting to see what happens when the two routes become closer substitutes or in other words, when the absolute value of β is increased. Figure 2 shows the equilibrium values of prices and demand for travel on both links for successively higher values of β. The prices (user costs) for the two links in Figure 2 approach each other closely. Since capacity is fixed, this implies that the link that had the lower user cost in the base case is used more intensively. Its lower price attracts more users when price sensitivity increases, and for this reason user cost increases and the price difference disappears. Convergence to the case of perfect substitutes for $\beta \rightarrow \infty$ has been confirmed by means of a separate model that uses the Wardrop equilibrium conditions.[9]

[9] This check has also been carried out in other uni-modal cases considered below.

Figure 2 Effect of increasing substitutability on user cost (upper panel) and route choice (lower panel) in the base case of the two-route model



2.3.5 First-best

Finding the first-best solution is facilitated by the assumptions of constant returns to scale of the capacity cost function and homogeneity of degree 0 of the travel cost function. Equations (9) imply:

$$k_i = vot\ fftt\ c_{i1} c_{ie} \left(\frac{q_i}{cap_i} \right)^{c_{ie}+1} \tag{27}$$

and this determines the ratio between number of users and capacity. The optimal toll is a function of this ratio only:

$$\tau_i = vot\ fftt\ c_{i1} c_{ie} \left(\frac{q_i}{cap_i} \right)^{c_{ie}} \tag{28}$$

as is the user cost of travel given in Equation (25). Together they determine the values of the prices p_i by Equation (5). The prices p_i in turn allow total demand and route choice to be computed. Consumer surplus, toll revenues, capacities and associated costs are presented in Table 3.

Table 3 First-best situation of the two-route problem

Variable	Route/mode 1	Route/mode 2
cap	2,511	2,942
q	2,817	3,579
Q	6,396	
c	6.96	4.98
τ	5.35	4.93
p	12.31	9.91
P	11.04	
Elasticity of *q* w.r.t. p_1	-0.77	0.47
Elasticity of *q* w.r.t. p_2	0.46	-0.52
Elasticity of *Q* w.r.t. *P*	-0.17	
CS	204,542	
TR	15,063	17,650
K	15,063	17,650
SS	204,542	

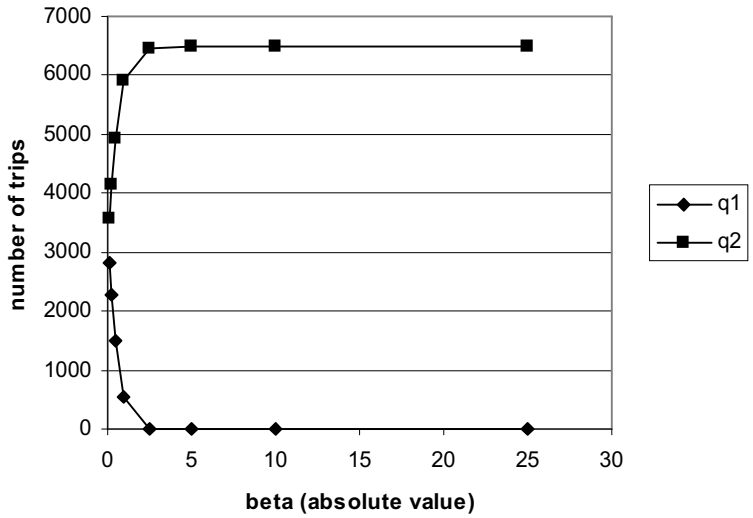
Source: own calculations.

Note: *cap*, *q* and *Q* are measured as numbers of trips per hour, *c*, τ, *p, P, CS, K* and *SS* in euros.

For the first-best case, it is also interesting to see what happens if the price sensitivity of route choice increases. The first thing to observe is that in this case the

prices (user costs) on the two routes cannot become equal. As explained above, these prices are determined by the ratio between number of users and capacity, and this ratio is fixed by the first order condition for capacity choice. The constant price difference between the two routes will have a larger effect on demand when price sensitivity increases. Since capacity is adjusted to the number of users because of the constant ratio between the two, this implies that the most expensive road will gradually disappear with enough substitutability. Figure 3 shows that demand shifts towards the first route when the absolute value of β increases. For values larger than 2.5 demand for trips on route 1 is virtually equal to 0.

Figure 3 Effect of increasing substitutability on route choice in the first-best situation of the two-route model



2.3.6 Second-best: the two route problem

In order to solve second-best problems an iterative procedure was used. As a starting point a user equilibrium for given values of the policy variables was taken. Then a linearised version of the first order conditions was solved for the second-best optimum. In the following step a weighted average of the original values and those

suggested by the linearised first order conditions for the policy variables that can be freely chosen was used. This procedure was repeated until convergence occurred.
A sequence of second best situations, where in each step an additional policy instrument is added, will be discussed. In the base case no instruments are used, a toll on road two is then introduced and the capacities of roads 2 and 1 are subsequenlty added. If, finally, a toll on road 1 comes available as a policy instrument, the first-best situation results.

One toll, fixed capacities

First, the situation in which the toll on route 1 is equal to zero, capacities on both routes are fixed and the policy maker's single instrument is the toll on route 2 is analysed. This implies that of the first order conditions (7)-(9), only (7) and (8) for i=2 are satisfied. In that situation only a modest welfare gain in comparison to the base case can be achieved: social surplus rises to € 166,654. The index of relative welfare improvement, ω,[10] is equal to 0.13. The optimal toll is equal to 8.53, and severe congestion remains on both routes. When price sensitivity increases, user costs on both routes converge, whereas the numbers of trips diverge somewhat, just as in the base case.

These results are surprisingly close to the results from the model with perfect substitutability (τ_2=8.39 and ω=0.15, with SS=€161,579 in the base case, SS=€169,029 in the second-best case, and SS=€211,809 in the first-best case, where road 1 is in fact eliminated). This suggests that, although the assumption of imperfect substitutability between routes makes the model richer and allows for mode choice, it

[10] The index of relative welfare improvement is the increase in social surplus compared to the base equilibrium as a fraction of the increase in the first-best optimum. Note that the value of this index is dependent on the parameter values chosen in the base case. For instance, the

does not strongly affect the policy conclusions on the design and desirability of second-best pricing.

One toll, one capacity

The second phase concentrated on the analysis of the situation in which the toll and capacity of route 2 can be determined by the policy maker. First-order conditions (7) are both satisfied, but conditions (8) and (9) are only valid for i=2. In the second-best optimum, the capacity of route 2 is extended substantially to 4,048 trips per hour, whereas the toll is negative: € –2.79. Congestion on route 1 remains significant, with the number of trips at 2,003. The welfare gain is substantial: social surplus is now equal to €194,510, with ω=0.77. When price sensitivity increases, the optimal toll increases. It becomes positive for values of β higher than 0.25. At that level ω is still positive (and equal to .91), which reflects the welfare gain from optimizing route 2's capacity, even when the second best optimal toll is zero. When β increases, the optimal capacity of route 2 increases, but congestion on route 2 also grows. In the limiting case of perfect substitutability, the second-best toll is equal to €1.71, the capacity, to 4,525.25 trips per hour, and ω increases to 0.92. These results suggest that when capacity choice is a policy variable, the impact of allowing imperfect substitutability upon policy design and evaluation increases.

One toll, two capacities

Finally, the case in which route 1 cannot be tolled, whereas the toll on route 2 and the capacities on both routes can be freely chosen was considered. In this situation all

value of this index for the present case "one toll, fixed capacities" would have been larger if the capacities in the base case had been closer to their first best optimal levels.

first-order conditions, except (8) for i=1 are satisfied. The optimal values of these policy variables are presented in Table 4.

Table 4 Second-best: the two route problem with one toll and two capacities

Variable	Route/mode 1	Route/mode 2
cap	2,806	2,955
q	3,244	3,596
Q	6840	
c	7.13	4.98
τ	0	1.11
p	7.13	6.10
P	6.60	
Elasticity of q w.r.t. p_1	-0.42	0.27
Elasticity of q w.r.t. p_2	0.29	-0.34
Elasticity of Q w.r.t. P	-0.10	
CS	233,916	
TR	0	4,016
K	16,841	17,732
SS	203,359	

Source: own calculations.

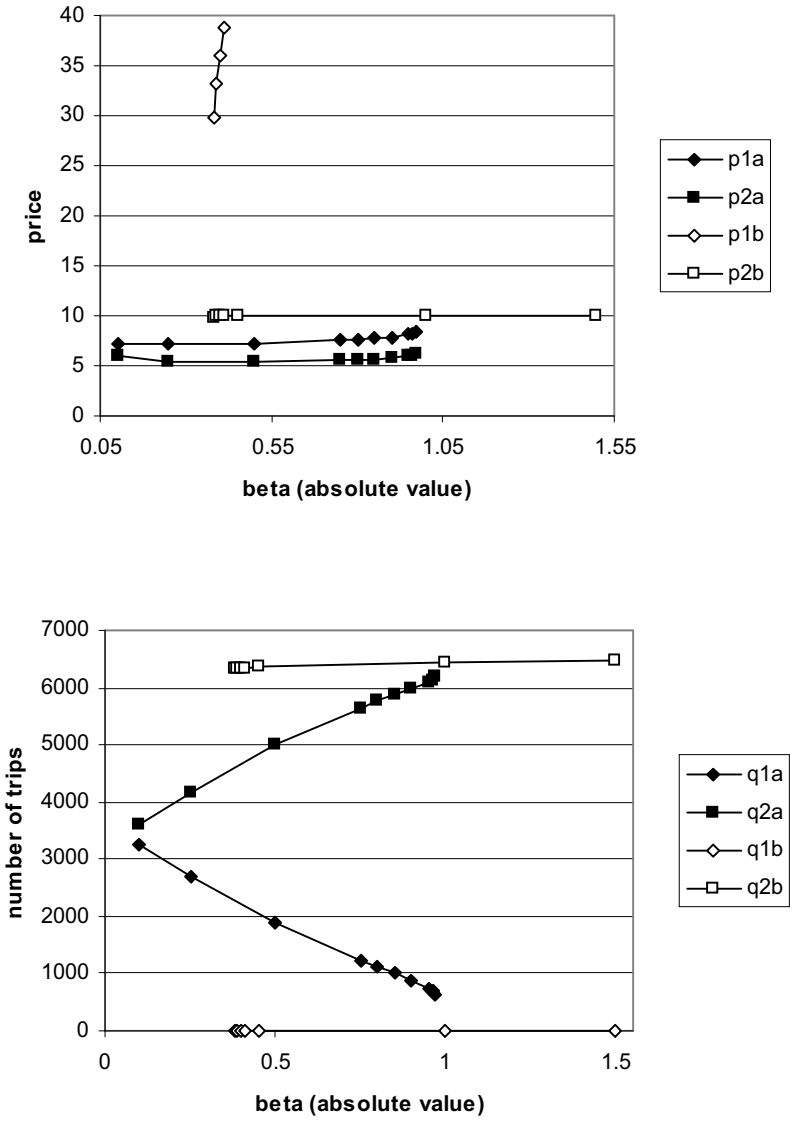
Note: *cap*, *q* and *Q* are measured as numbers of trips per hour, c,τ,p,P, *CS*,*K* and *SS* in euros.

In comparison to the first-best situation, the capacity of route 1 is increased, whereas the toll on route 2 is decreased. The price of transport is lower, which increases consumer surplus. The costs of capacity are no longer covered by toll revenues, and the deficit decreases social surplus. These two changes result in a social surplus that is almost as high as in the first-best situation (ω=0.97).

When looking at the sensitivity of these results to changes in the price sensitivity of mode choice it was found during the simulations that the number of trips on route 1 in the second-best optimum suddenly falls to (a value very close to) zero when the

absolute value of β increases from 0.97 to 0.98.[11] Further investigation revealed that for the interval [-0.97,-0.37] there exist two equilibria: (a) one in which both routes have a substantial number of users and another (b) in which road 1 is virtually unused.

Figure 4 Effect of increasing substitutability on user cost (upper panel) and route choice (lower panel) in the two-route problem with one toll and two capacities



[11] The number of trips on route 1 falls from 635 to 0, the number of trips on route 2 increases from 6,192 to 6,438.

For values of β smaller than 0.37 (in absolute value) the equilibrium in which route 1 is virtually unused disappears. Figure 4 shows the effects on user costs and quantities. The prices and quantities ending with 'a' refer to the equilibrium in which both routes are used, and those ending with 'b', to the equilibrium in which only route 2 is used.

The price for trips on route 1 in equilibrium 'b' is only indicated for values of β lower than 0.42, since it increases rapidly to values higher than 100. Use of route 1 is always close to 0. Only for values of β smaller than 0.42 is the number of trips on road 1 in equilibrium 'b' larger than 0.01. The equilibrium user costs on the two routes in equilibrium 'a' do not converge and in fact the difference between them increases. As a result, the numbers of users on the two routes diverge. Eventually, it becomes inefficient to maintain route 1. With perfect substitutability, the ω for this policy is equal to 1. This implies that eliminating a road in the first-best optimum is not harmful to efficiency if its toll is restricted to be equal to zero. When the two equilibria exist simultaneously, the more balance equilibrium 'a' always has the highest social surplus. When β increases, the difference between social surpluses decreases.

The findings just reported show that there may be two solutions of the relevant subset of first-order conditions in the two-mode problem with one toll and two capacities as policy instruments. The two solutions correspond with different values of the policy instruments. When there are two equilibria, both are stable in two senses. First, the two solutions correspond with different values of the policy instruments, but

for both configurations the user equilibrium in the network is table. Second, both second best equilibria represent local welfare maxima, not minima.[12]

What has been found is that a policy maker who starts from an arbitrary use equilibrium and adjusts the values of the instruments so as to reach a second best optimum may end up in two different situations if the parameter β lies in a particular interval. The initial values of the policy instruments determine in which equilibrium he ends up. Although no extensive mathematical or numerical analysis was undertaken, previous experience indicates that when there exist two equilibria, both can be reached with substantially different initial values of the policy instruments.[13]

2.3 A SIMULATION MODEL FOR THE TWO-MODE PROBLEM

2.4.1 Set-up of the two-mode problem

There are two main differences between the two-route and the two-mode problems. The first is that studies of the two route problem usually assume perfect substitutability. This was an important motivation for considering the alternative case of imperfect substitutes in the previous section. The second difference involves the transport technology. In the two route problem, constant returns to scale in capacity cost and homogeneity of degree zero in capacity and demand are usually assumed, whereas public transport is often produced under increasing returns to scale and with a technology that is not homogeneous. In the present section a simulation model for

[12] This is confirmed, for instance, if social surplus is maximised by choosing capacity and toll for road 2 with various exogenously determined values of the capacity of road 1. For β=0.75 there exists a local interior maximum when the capacity of road 1 is chosen at the level of equilibrium (a) and another one at the boundary, when the capacity of road 1 equals zero. When β is close to (but larger than) .37 the second maximum also becomes an interior maximum.

the two-mode problem that incorporates a different transportation technology for the second mode is developed. It is otherwise comparable to the two-route problem of the previous section.

For this two-mode model, demand, user cost of roads (mode 1) and cost of road capacity are all specified as in the two-route model. For rail (mode 2), two aspects of capacity are distinguished: frequency and number of seats per train. A peak period of one hour is considered, during which the arrival flow of commuters at the train platform is assumed to be constant. The average waiting time is therefore equal to 0.5/*N*, where *N* denotes the frequency of train departures. Passengers dislike waiting and the cost of waiting time is equal to (1+*w*) *vot*, where *vot* denotes the value of time and *w* gives the extra weight given to the value of waiting time (Mohring, 1972).

The discomfort associated with travelling by train is assumed to be (a) proportional to travel time (*tt*) and (b) increasing in the ratio between the number of passengers and the number of seats. The total number of passengers is q_2 and the total number of seats available is *Ns*, where *s* denotes the number of seats per train. The cost function for travelling by train is defined as:

$$c_2 = vot\left[.5(1+w)/N + c_{21}\, tt\left(\frac{q_2}{Ns}\right)^{c_{2e}}\right] \qquad (29)$$

This function is homogeneous of degree 0 in the number of trips and the number of seats. It is *not* homogeneous of degree 0 in the number of trips and trip frequency.

The cost of providing rail transport is determined by the number of train departures and by the product of the number of departures and the number of seats per train:

[13] The case in which the two capacities are the only policy instruments (both tolls are equal to zero) has also been considered and found no indications of multiple policy equilibria. When β gets large (in absolute value) route 1 gradually disappears.

$$K_2 = k_{21}N + k_{22}Ns \qquad (30)$$

The first order conditions for the optimum are analogous to conditions (7) - (9). The only difference is that there are now two first order conditions for the capacity of the second mode.

Table 5 Parameter values for the two-mode problem

Parameter	Route/mode 1	Route/mode 2
Demand		
a		7,500
b		-100
β		-0.1
α	1	1
User cost		
vot		7.5
fftt /tt	0.5	0.5
w		2
c_1	0.15	1.0
c_e	4	4
Cost of capacity		
k	6	$k_{21} = 365$
		$k_{22} = 4.5$

The parameter values for the simulation model are listed in Table 5. The demand parameters are left unchanged. Travel time by rail is assumed to be equal to free-flow travel time by car. The additional cost of waiting is assumed to be 100 per cent of the reference value of time, as is conventional. The exponent c_{2e} has a value of 4. The cost parameters are based on MuConsult (1999), where the total costs of passenger transport of Dutch Railways are approximated with a linear function of seat kilometres, train kilometres and passenger kilometres.[15] The passenger kilometres component is relatively small and is ignored here. The marginal cost of train

[14] A thorough discussion on the value of time, including the value of waiting time for commuters, is presented in Chapter 2 of this volume.

kilometres was found to be much higher than that of seat kilometres. The distance between A and B is assumed to be 60 kilometres.

Table 6 The base case of the two-mode problem

Variable	Route/mode 1	Route/mode 2
cap	1,250	*s*: 300 *N*: 6
q	2,887	2,582
Q	5,469	
c	1,976	17.13
τ	0	3.75
p	19.76	20.88
P	20.31	
Elasticity of *q* w.r.t. p_1	-1.12	0.81
Elasticity of *q* w.r.t. p_2	0.85	-1.28
Elasticity of *Q* w.r.t. *P*	-0.37	
CS	149,574	
TR	0	9,683
K	7,500	9,390
SS	142,367	

Source: own calculations.

Note: *cap* for the road , *N*, *q* and *Q* are measured as numbers of trips per hour, *c*, τ, *p, P, CS, K* and *SS* in euros.

In the base case it is assumed that trains with a capacity of 300 seats depart every 10 minutes. In the Netherlands the fare for a return trip of 60 kilometres is around €15. Given that for peak hours most of the trips are return trips, it seems appropriate to use half the value of a return ticket. Many of these trips, however, are made by season ticket holders, who pay a lower price per trip. For this reason 25 per cent (instead of 50 per cent) was used as the appropriate price for a one-way trip of 60 km.

[15] These are only operating costs, not capital costs. Moreover, fixed cost and costs directly related to the number of passengers, which turned out to be relatively small, are ignored here.

Results for this base case are presented in Table 6. The price of transport is higher than in the case of the two-route problem. The main reason for this difference is that the cost function for passenger transport is more sensitive to congestion than that for road transport. For the latter, the ratio between the number of users and capacity can be made equal to two. For passenger transport, at least in the Netherlands, it is virtually impossible to carry twice as many passengers as there are seats. This difference is reflected in the specifications of the cost functions for rail and road transport as discussed above. In the base case the ratio between use and capacity is 1.4 for trains. This figure seems reasonable for commuter trains in the Randstad area during peak hours. Revenues from train tickets are sufficient to cover all costs of rail passenger transport, as seems realistic for peak traffic in the Randstad area.

In order to study the effect of changes in the substitutability between the modes, the same range of values of the coefficient β that was used for road transport in the two-mode version simulation model were simulated. Only minor differences with the two-route situation were found. Again, the prices for the two modes, which are already close to each other in the base case, become virtually identical. The price elasticities of demand for the separate modes become very large (in absolute values) but the overall price of transport and its price elasticity remain almost identical. It must therefore be concluded that, at least in this base-case version of the two-mode model, the differences between the situations of imperfect and perfect substitutes are minor and comparable to the differences in the two-route case.

4.2 The first-best situation and the two-mode problem

Determining the first-best equilibrium is more complicated than in the two-route case. The ratio q_2/Ns is determined by the first order condition for s, but this does not determine N. This means that the cost of railway transport cannot be computed from

the first order condition with respect to capacity alone, and an iterative procedure is therefore needed in order to compute the first-best equilibrium. The equilibrium was first computed conditional upon a given value of N, the value of N suggested by the relevant first order condition in this equilibrium was then computed, and N adjusted towards this optimal value. This procedure was repeated until convergence occurred.

Table 7 First-best situation of the two-mode model

Variable	Route/mode 1	Route/mode 2	
cap	2385	*s* :	556
		N :	8.77
q	2,901	3,743	
Q	6,644		
c	4.98	2.16	
τ	4.93	5.21	
p	9.91	7.37	
P	8.56		
Elasticity of q w.r.t. p_1	-0.62	0.35	
Elasticity of q w.r.t. p_2	0.37	-0.38	
Elasticity of Q w.r.t. P	-0.13		
CS	220,712		
TR	14,307	19,500	
K	14,307	22,701	
SS	217,511		

Source: own calculations.

Note: *cap* for the road, N, q and Q are measured as numbers of trips per hour, c, τ, *p, P, CS, K* and *SS* in euros.

The simulation results are presented in Table 7. In comparison with the base case, the number of rail passengers increases substantially. The quality of rail transport is improved substantially: the number of departures per hour increases from 6 to 8.77, and the number of seats per train almost doubles, and becomes equal to 556. The ratio between passengers and seats drops to 0.76. These improvements result in a much

lower user cost of travel time for rail transport (€1.96 instead of €16.99). Even though the railway fare increases considerably to €5.21 per (one way) trip, the generalized price of train transport falls considerably because of the lower travel time, and is now lower than that of car travel.

Road capacity is also expanded substantially and the price of road transport is halved in comparison with the base case. Toll revenues are exactly equal to the cost of road capacity. Fare revenues from rail transport are now insufficient to cover all costs. These revenues exactly cover the cost of seat kilometres (i.e. they are equal to $k_{22}Ns$ in Equation (2.30)). There are economies of scale associated with increasing the number of seats per train and marginal cost pricing is therefore unable to cover total exploitation costs.

When train and road become closer substitutes, demand shifts towards the cheapest mode. With the assumed parameter values, the cheapest mode is the train. With economies of scale in the railways, this demand shift causes a further drop in the cost of rail transport and road as a mode of transport eventually disappears.

Railway fare, fixed capacities

When the capacities of both modes are fixed and the only policy instrument is the train fare, possible welfare gains are small when compared to the base case (SS=€151,737, ω=0.11). The optimal railway fare turns out to be considerably *higher* than in the first-best situation (€16,15). There is substantial congestion on the road (q_1=2,980). User costs are equal to €21.92 for the road and €25.03 for the railway. When substitutability between the two modes is increased, user costs converge to €22.68. The optimal toll decreases somewhat.

Railway fare, frequency and seat capacity

When the railway fare, number of seats and frequency of departure can all be chosen by the policy maker, welfare gains are considerably higher than when only capacity can be used as a policy instrument. Social surplus is equal to €208,063 (ω=0.88), which is close to the first-best value. The optimal railway fare now turns out to be *negative* (-€2.09). User cost is equal to €8.45 for road and €1.87 for railway transport, implying that the price of the latter mode is also negative (-€0.22). When β is increased the railway fare becomes positive, while user costs of the two modes converge gradually.

Table 8 Second-best: the two mode problem

Variable	Route/mode 1	Route/mode 2	
cap	2,639	*s* :	556
		N :	8.79
q	3,298	3,757	
Q	7,055		
c	5.12	2.16	
τ	0	1.66	
p	5.12	3.82	
P	4.45		
Elasticity of *q* w.r.t. p_1	-0.31	0.17	
Elasticity of *q* w.r.t. p_2	0.21	-0.21	
Elasticity of *Q* w.r.t. to *P*	-0.06		
CS	248,870		
TR	0	6,247	
K	15,833	22,783	
SS	216,501		

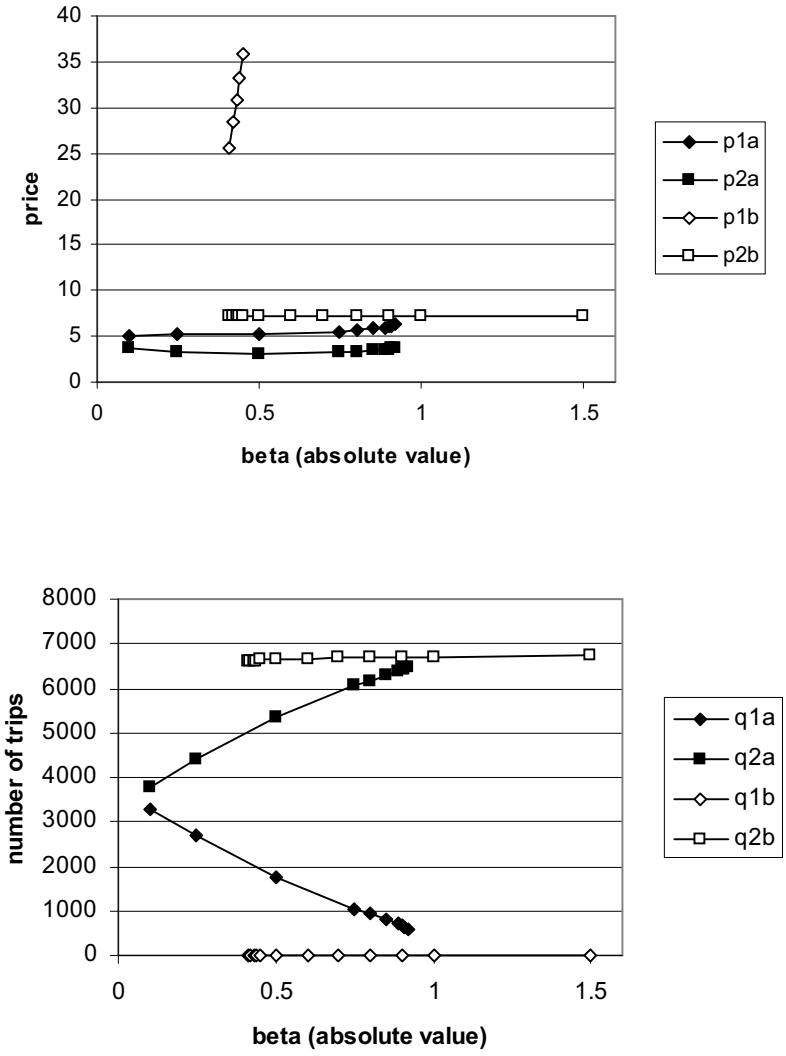
Railway fare, frequency and seat capacity, and road capacity

The results for the two-mode problem, where capacity of the road can also be chosen optimally, are presented in Table 8. As can be seen in the table, the share of potential welfare gains is even larger than in the case where only the capacity of rail could be

used as a policy instrument. The social surplus equals €215,889 (ω=0.98). Both road and railway capacities are considerably higher than in the first-best situation. The revenues from railway fares are less than one third of the costs of railway transport.

When substitutability between the two modes increases, results similar to those from the two-route problem are obtained. Figure 5 illustrates this.

Figure 5 Effect of increasing substitutability on user cost (upper panel) and route choice (lower panel) in the two-mode problem



For the interval [-0.92,-0.41] there are again two equilibria. In one of these the road remains virtually unused. For absolute values of β higher than 0.92 there is only one equilibrium, in which there is no road. For absolute values of β lower than 0.41 there is only one equilibrium in which both road and railway attract a substantial number of trips. For small values of β the equilibrium in which two modes are used has the highest social surplus, for larger values the equilibrium in which only one mode is used has the highest social surplus.

2.4.3 Summary

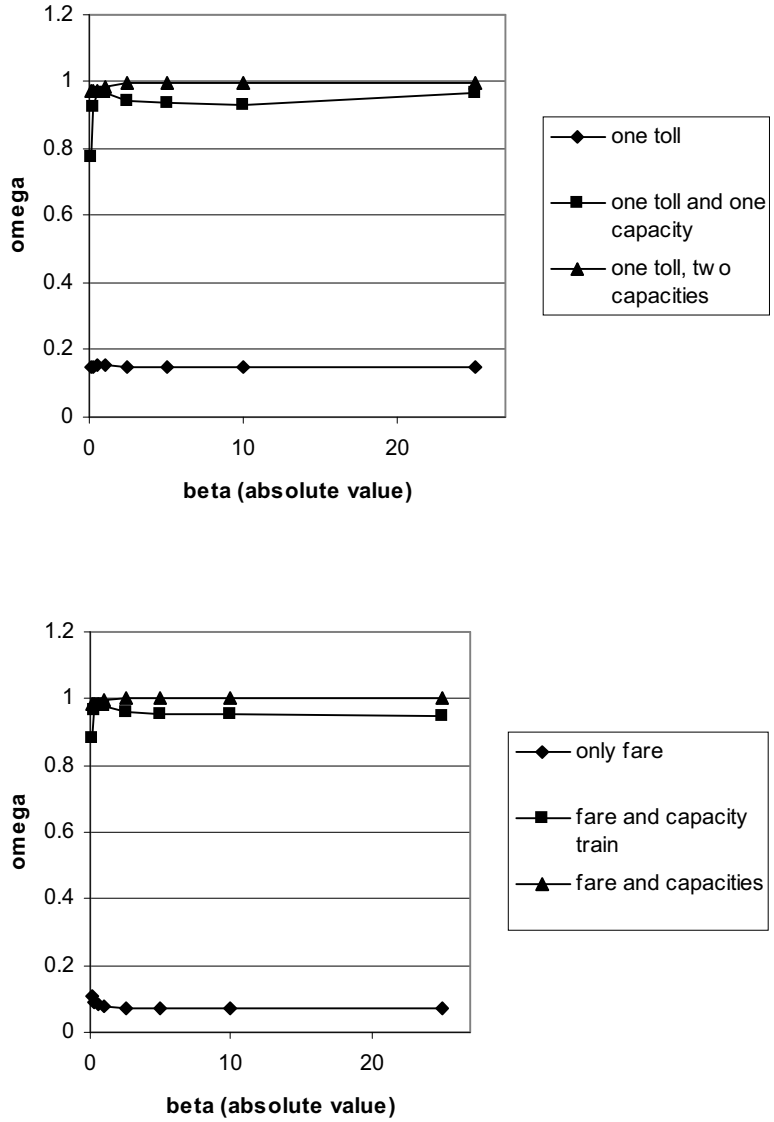
The main purpose of this paragraph was to provide a comparison of the results of two-mode situations with those of the two-route situations discussed in the previous section. The most important conclusion is that these results appear to be so similar. Introduction of imperfect substitutability into the two-road model makes it behave much like the analogous two-mode model, despite the substantial difference in the cost structure.

The similarity between the two models is nicely illustrated by the development of the index of relative welfare improvement as a function of the price substitutability parameter β, which is shown in Figure 6.[16] In both cases only modest improvements in welfare are possible when one road or mode has to remain untolled and capacity has to be taken as given. If capacity of the road or mode that can be tolled can also be used as a policy instrument, a much larger share (typically more than three quarters) of the potential (first best) welfare gains can be realized. If the capacities of both roads or modes can be used as policy instruments, 100 per cent of the potential welfare gains is realised for large values of β. The reason is that in these

[16] When there are two equilibria, the one yielding the highest social surplus was used in Figure 6.

circumstances the unpriced road or mode disappear, which implies that its price becomes irrelevant.

Figure 6 Effect of increasing substitutability on user cost and route/mode choice in the two-route (upper panel) and two-mode (lower panel) models



The simulation results clearly suggest that there is a qualitative difference between situations in which substitution between roads or modes is difficult (with the absolute value of β less than 1) and situations in which it is easier, but that there is not a

qualitative difference between perfect substitutability and imperfect but relatively easy substitutability (a large absolute value of β). Two important reasons or this conclusion are that one road or mode disappears completely in the first best case for finite values of β and that in the case with two capacities and one toll the transition from an equilibrium in which both roads/modes are used to one in which only one is used takes place for relatively small values of β.

2.4 CONCLUSIONS

This paper considers second-best pricing as it arises through incomplete coverage of full networks. The main principles were first reviewed by considering the classic two-route problem and some extensions that have been studied more recently. In most of these studies the competing routes are assumed to be perfect substitutes, which is probably not the case for most parallel roads in reality, and even less likely for the case where competing connections represent different transport modes.

A modelling framework in which the alternatives are imperfect substitutes was developed and numerical results for two roads and two modes were presented. In the model, trip generation and trip distribution are distinguished in a way that is consistent with economic theory. A linear demand equation was used for trip generation, and the logit model for trip distribution. The model offers the possibility to study the effect of changes in the substitutability between the two routes or modes. Perfect substitutability is a limiting case, in which the absolute value of one parameter becomes infinitely large.

The model was used to consider situations in which one route or mode cannot be tolled. Simulation results show that, for the chosen parameter values, there is a substantial difference between the effectiveness of policies in which the capacities

have to be taken as given, and those in which capacity of at least one mode can be changed. If only a toll on route or mode two can be used typically less than a quarter of the total possible welfare gains is realised. When the capacity of at least one route or mode can be determined by the policy maker, typically more than three quarters of the maximum possible welfare gains are realised. These figures are not very dependent on the substitutability between the two routes or modes.

A striking feature of the policy in which the capacities of both modes and the railway fare can be used as policy instruments is the existence of two equilibria for a range of values of β. In one equilibrium there are substantial numbers of users of both modes, whereas in the other use of one mode is negligible. In this case there is a regime shift that is related to the possibility to substitute use of one mode for another, but it does not coincide with the difference between perfect and imperfect substitutability, interpreted as a finite and an infinite value of β respectively.

For the first-best case there is no regime shift that coincides with the difference between perfect and imperfect substitutability either. The situation in which only one mode is used is now approached in a continuous way, and it is reached for a finite and relatively small value of β.

The results for the classic two-route problem with imperfect substitutes, in which only a toll can be charged on one of two parallel roads, are surprisingly close to the results for the model with perfect substitutes. This suggests that although the assumption of imperfect substitutability between routes makes the model richer and allows for mode choice, it does not strongly affect the policy conclusions on the design and desirability of second-best pricing. This makes the results of prior studies appear robust to the assumption of perfect substitutability.

This picture changes when capacity choice is also included as a policy instrument. When only the capacity of the priced road can be chosen, the choice may make the difference between a negative second-best toll (for imperfect substitutes) and a positive one (for closer or perfect substitutes), as illustrated by the results. When both capacities can be chosen, the choice may make the difference between keeping a road (for imperfect substitutes) and eliminating it in the long run (for closer or perfect substitutes).

REFERENCES

Arnott, R. and Yan, A. (2000). The Two-Mode Problem: Second-Best Pricing and Capacity. *Review of Urban and Regional Development Studies 12,* 170-199.

Braid, R.M. (1996). Peak-Load Pricing of a Transportation Route with an Unpriced Substitute. *Journal of Urban Economics 40*, 179-197.

De Palma, A. and Lindsey, R. (2000). Private Toll Roads: Competition under Various Ownership Regimes. *Annals of Regional Science 34*, 13-35.

Knight, F. (1924). Some Fallacies in the Interpretation of Social Costs. *Quarterly Journal of Economics 38*, 582-606.

Lévy-Lambert, H. (1968). Tarification des Services à Qualité Variable: Application aux Péages de Circulation. *Econometrica 36*, 564-574.

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

Liu, L.N. and McDonald, J.F. (1998). Efficient Congestion Tolls in the Presence of Unpriced Congestion: a Peak and Off-Peak Simulation Model. *Journal of Urban Economics 44*, 352-366.

Marchand, M. (1968). A Note on Optimal Tolls in an Imperfect Environment. *Econometrica 36*(3-4), 575-581.

Mohring, H. (1972). Optimization and Scale Economies in Urban Bus Transportation. *American Economic Review 62*, 591-604.

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

MuConsult (1999). Voorbereiding Prestatiecontract NS (Preparation Achievements Contract Dutch Railways). Report for the Dutch Ministry of Transport (in Dutch).

Pigou, A.C. (1920). *Wealth and Welfare*. Macmillan, London.

Small, K.A. (1992) *Urban Transportation Economics.* Harwood, Chur.

Small, K.A. and Rosen, H.S. (1981). Applied Welfare Economics with Discrete Choice Models. *Econometrica 49,* 105-130.

Verhoef, E.T., Nijkamp, P. and Rietveld, P. (1996). Second-Best Congestion Pricing: the Case of an Untolled Alternative. *Journal of Urban Economics 40,* 279-302.

Vickrey, W.S. (1963). Pricing in Urban and Suburban Transport. *American Economic Review* 53, 452-465.

Vickrey, W.S. (1969). Congestion Theory and Transport Investment. *American Economic Review* (*Papers and Proceedings*) 59, 251-260.

Walters, A.A. (1961). The theory and measurement of private and social cost of highway congestion. *Econometrica 29,* 676-697.

Wardrop, J. (1952). Some Theoretical Aspects of Road Traffic Research. *Proceedings of the Institute of Civil Engineers, 1(II),* 325-378.

Yang, H. and Meng, Q. (2002) A Note on Highway Pricing and Capacity Choice under a Build-Operate Transfer Scheme. *Transportation Research 36A,* 659-663.