Congestion pricing in a road and rail network with heterogeneous values of time and schedule delay**

Version of 20 March 2013

Vincent A.C. van den Berg*,#
Department of Spatial Economics
VU University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
+31-20-598 6160
v.a.c.vanden.berg@vu.nl

Erik T. Verhoef#
Department of Spatial Economics
VU University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
+31-20-598 6090
e.t.verhoef@vu.nl

Abstract

We analyse congestion pricing in a road and rail network, where the two modes are imperfect substitutes. On the road there is bottleneck congestion; in each train there is crowding congestion. We model two dimensions of preference heterogeneity; these two dimensions have opposite effects on the welfare impact of congestion pricing and lead to different distributional effects. The distributional effects also differ between road and rail. On the road, pricing is generally *more* beneficial with a higher value of time or schedule delay. In the train, pricing has no distributional effects or is *less* beneficial with a higher value.

JEL codes: D62, H23, L11, R41, R48

Keywords: Congestion Pricing; Car Travel; Train Travel; Heterogeneity; Distributional Effects

* Corresponding author.
# Affiliated to the Tinbergen Institute, Gustav Mahlerplein 117, 1082 MS Amsterdam.

** This is an Author's Accepted Manuscript of an article published in *Transportmetrica A: Transport Science*, on 11 March 2013, available online at: http://dx.doi.org/10.1080/23249935.2013.766820

1. Introduction

Road congestion is one of the most important problems in urban areas. Small and Verhoef (2007) find that congestion caused the largest marginal external cost due to car travel in the US in 2005: of this total 65% was from congestion, 6% was environmental, and 25% from accident externalities. Transport pricing is one of the main instruments that a government could use to address externalities and to affect modal choices. In studying such measures, it is important to take into account the dynamics of departure time choice, the availability of substitute modes, and preference heterogeneity. Therefore, our bi-modal model of (first- or second-best) congestion pricing takes all these issues into account, while also considering that the modes are imperfect substitutes.[1]

Although the efficiency argument for congestion pricing is firmly established, it is seldom applied in practise. Often, proposals meet fierce resistance, and it is important that analysis helps identifying where such resistance is strongest, and why. Distributional effects are an important reason why congestion pricing meets resistance. When pricing reduces travel times and increases monetary costs, an individual is better off when her value of time is higher. Since the value of time generally increases with income, this is often viewed as implying that the poor lose and the rich gain from road pricing (see, e.g., Layard, 1977). However, things might not be as simple as this. For example, Vickrey (1973) uses a bottleneck model with fixed demand, and analyses, what we will call, "proportional heterogeneity". In his setting, the values of time (α) and schedule delay (β) vary over individuals in fixed proportions (where the latter value refers to the monetary value of arriving at a different moment than most desired). Because these values depend on the inverse of marginal utility of income, proportional heterogeneity could stem from differences in incomes: a higher income usually leads to a lower marginal utility of income, and this in turn proportionally increases both values. Perhaps surprisingly, in this setting, first-best tolling brings a strict Pareto improvement (even before revenue recycling): all users gain—except those with the lowest values, who are unaffected.

Cohen (1987) and de Palma and Lindsey (2002) study heterogeneity in the value of time with fixed values of schedule delay. We call this heterogeneity "ratio heterogeneity", since it implies heterogeneity in the ratio of the value of time to value of schedule delay $\mu_i=\alpha_i/\beta$. This ratio has a clear interpretation: it reflects the willingness to accept greater schedule delays in order to reduce travel time. Heterogeneity in μ could, for example, result from differences in possibilities to use in-vehicle time productively or differences in the tightness of scheduling constraints: e.g. having small children versus not, working on an assembly line versus in an office, and travelling to a hospital appointment versus to go shopping. When only this type of heterogeneity exists, tolling

[1] Although we call the two modes rail and car, they can be any two modes that have no congestion interactions.

is harmful for users: under fixed demand, all users lose—except those with the highest value of time, who are unaffected (de Palma and Lindsey, 2002; Van den Berg and Verhoef, 2011a).

Besides these distributional effects, heterogeneity can also influence the aggregate welfare gain of a policy, as well as the relative performance of policies (see, e.g., Arnott et al. (1988), Small and Yan (2001), and Verhoef and Small (2004)).

Van den Berg and Verhoef (2011b) show that price sensitivity of demand changes the effects of heterogeneity in the bottleneck model. With proportional heterogeneity, first-best tolling lowers the generalised price for users with high values and, consequently, they demand more travel, which increases congestion and makes tolling more harmful for drivers with low values. This paper finds that also the amount of substitutability of the two modes changes the effects of pricing, and how heterogeneity influences these effects. Hence, our analysis confirms that it is not only important to control for heterogeneity and price sensitivity of demand, but also for the degree of substitutability between different modes.

We study both first-best pricing of both modes and second-best pricing of one mode only. The latter bears close resemblance with second-best pricing of one of two parallel roads. Such roads are usually closer substitutes than two modes. Braid (1996) was one of the first to study this second-best policy in the dynamic bottleneck model: the welfare-maximising toll has a time-variant term that equals the congestion externality, and a time-invariant term that is negative. This negative term attracts users away from the unpriced link, where marginal social cost is above private cost, and thereby increases welfare.

Besides being an imperfect substitute to car travel, public transport also needs a different specification of user costs than road transport. To accommodate departure time choice, we need a dynamic model of public transport. Various such models have been proposed. Kraus and Yoshida (2002) develop a model with queuing at the platform. Their model is similar to the bottleneck model: also here queuing is a pure loss and dynamic pricing completely eliminates it. Tabuchi (1993), Huang (2000), and Rouwendal and Verhoef (2004) use a parallel road and rail track. Tabuchi (1993) has bottleneck congestion on the road and no congestion in the single train service, which arrives on the preferred arrival time, t^*. The train operator's costs consist of fixed and per-user cost. If the train fare equals average operating cost, the second-best road toll is the standard time-variant toll plus a *positive* term that pushes users to the train. This difference from Braid's (1996) result stems from the fact that Tabuchi's train has economies of scale, and thus increasing the number of users lowers average cost. This shows that seemingly similar policies in road transport and public transport may work out rather differently, making it important to take into account carefully whether there are different modes. Huang (2000) considers two user groups with different values of time, schedule delay, and crowding. The road has dynamic bottleneck congestion. The single train service arrives at t^* and has crowding congestion: that is

costs associated with being in a crowded train. Rouwendal and Verhoef (2004) also use crowding congestion. However, they use a static road model, and car and train are imperfect substitutes.

We analyse a parallel road and rail line using a dynamic model with continues time. On the road there is bottleneck congestion. In each train service there is crowding congestion. Our train model is dynamic, which differs from Tabuchi (1993) and Huang (2000). We consider profit maximisation on only road or rail, which appears to have been ignored for rail-road networks. In reality, rail companies are often private or at least partly privatised; therefore, profit maximisation by a rail operator seems important. This relevance increases when rail franchises are auctioned off to the highest bidder, since this drives the winning firm towards profit maximisation. Private provision of roads is also an increasingly popular policy option.

We find that heterogeneity affects road and rail differently. Without congestion pricing, the mean price of car travel decreases with the degree of ratio heterogeneity; while proportional heterogeneity has no effect. Conversely, the average price of train travel is unaffected by ratio heterogeneity, but decreases with the degree of proportional heterogeneity. For second-best pricing, we find that the relative efficiency of "only pricing the road" decreases with proportional heterogeneity (relative efficiency equals the welfare gain of a policy relative to the first-best gain). This contrasts with the result with two parallel bottlenecks in Van den Berg and Verhoef (2011b), who have the same two dimensions of heterogeneity as here, and find that the relative efficiency of single-link pricing increases with proportional heterogeneity. As we will explain, this difference is due to the assumption that two roads are perfect substitutes, whereas here road and rail are imperfect substitutes.[2] The current paper further extends our previous work by looking at a bi-model system where the public transport has a different type of congestion than the road, and by also looking at private pricing. Van den Berg (2011a) use two parallel bottlenecks under only ratio heterogeneity, and investigate first-best and second-best pricing, private pricing and a uniform toll that is constrained to be constant over time.

Our use of two separate dimensions of heterogeneity—ratio and proportional—is motivated by the observation that the two types together would allow for an unrestricted bivariate distribution. The interpretation of results then becomes tedious, as these are the combined result of the effects of the two distributions. Considering them separately is therefore primarily motivated by reasons of ease of exposition. Note that Van den Berg and Verhoef (2011b) find that the qualitative effects of the one type of heterogeneity are not affected by the degree of the other type.

[2] There are other types of preference heterogeneity than those studies here. For instance, heterogeneity in the preferred arrival time is certainly present in reality; but it seems to have limited effects, if the morning queue is single peaked. Arnott et al. (1988, 1994) find that then total schedule delay cost are lower than with homogeneity, but by the same amount with and without tolling, and hence tolling has no distributional effects. Vickrey (1969, 1973), Hendrickson and Kocur (1981), and Cohen (1987) also include this heterogeneity; and Gonzales and Daganzo (2012) use it in a model with two modes: 1) car using a bottleneck and 2) uncongestible public transport which uses part of the bottleneck capacity when operating, where the two modes are perfect substitutes. Another type of heterogeneity is between the values of schedule delay early and late. Arnott et al. (1988, 1994) also study this case; they find that, with two user groups, it does affect the aggregate gain of tolling, but does not lead to distributional effects.

The paper is arranged as follows. Section 2 discusses the general set-up of the model, Section 3 derives the analytical model for car travel, and Section 4 develops the model for rail travel. Section 5 describes the numerical set-up, which Sections 6 and 7 use in their analyses of the numerical models for respectively proportional and ratio heterogeneity. These sections also give sensitivity analyses on the degree of heterogeneity. Section 8 presents further sensitivity analyses. Section 9 concludes.

2. The model set-up

Our model considers one road and one rail track that connect a single origin and destination. To simplify the analysis, we analyse the two types of heterogeneity separately: we first study ratio heterogeneity and then proportional heterogeneity.

We use a deterministic model, without uncertainty of travel times, but with continuous heterogeneity in preferences. We only look at the short-term policy of price setting, and ignore long-term policies such as train schedule or capacity. Furthermore, it is assumed that rail-operating costs are a constant amount per user, and that there are no fixed costs. We ignore heterogeneity in the value of crowding congestion. To simplify the rail model, this paper uses the common assumption that users cannot arrive after the preferred arrival time, t^*,[3] which is the same for all users. Arrivals before t^* have a schedule delay costs, which has an per hour value of β. The monetary value of an hour of travel time is α.

A "user type" comprises of all the users with a certain combination of values of schedule delay and time. The inverse demand (D_i^j) for user type i of mode j gives the marginal willingness to pay to travel of the $n_i^j th$ user:

$$D_i^j = A_i^j + B_i^j n_i^j + E_i^j n_i^k .$$

This willingness to pay is in terms of the generalised price (P_i^j), which is the sum of the user cost and monetary transfers (i.e. the road toll and the train-ticket fare), and is henceforth referred to as "price" for brevity. The n_i^j is the density of type i users on mode l, and A_i^j is the intercept or the maximum willingness to pay. Both coefficients B_i^j and E_i^j must be negative; they respectively measure how much the inverse demand decreases when there are more users of type i on mode j

[3] With late arrivals, the effect on prices of the number of early and late train services is asymmetric and the exact form of the price equation depends on the ratio of the values of schedule delay early and late. Arnott and Kraus (1993; 1995), Kraus and Yoshida (2002), and Kraus (2003) also use this simplification. Although the restriction of no late arrival is unrealistic; it does not affect the general results of the bottleneck (see Arnott and Kraus, 1993) or the crowding model.

and k. If E_i^j were positive, the modes would be complements, and the users of mode j would accept a higher generalised price when there are more users of k.[4]

In user equilibrium, for all user types, the price of each mode equals the value of its inverse demand function (superscript c indicates the car, and r the rail link), giving the following equilibrium conditions:

$$\begin{aligned} D_i^c &= P_i^c \\ D_i^r &= P_i^r \end{aligned} \qquad \forall i. \tag{1}$$

Hence, different from with perfect substitutes, prices now generally differ between the two modes. There is typically no closed-form solution to this continuum of equilibrium conditions,[5] since the price of a mode for a type not only depends on its own number of users, but also on those of all other types. Section 5 will describe how the numerical model finds the user-equilibrium.

Gross consumer benefit (G_i) for type i users is the line integral of the two inverse demands (see also footnote 4). Consumer surplus for i equals benefit minus what we could call the total price: $P_i^r \cdot n_i^r + P_i^c \cdot n_i^c$. Since we focus on continuous heterogeneity, total consumer surplus is the integral of these surpluses over the user types; with discrete heterogeneity, it would be the sum. Welfare (W) is the integral of gross consumer benefit over all types minus total cost for all users and operators (TC):

$$W = \int G_i \mathrm{d}i - TC.$$

3. Introducing heterogeneity: analytical exposition for a road with a bottleneck

Before we turn to the full model, we will present the sub-models for road and public transport separately. We start with the former, and use the exposition to summarise the insights from the literature on the impacts of heterogeneity.

3.1. Ratio heterogeneity and pricing of a single bottleneck

The first type of heterogeneity we consider is "ratio heterogeneity". This refers to heterogeneity in the value of time (α) with a fixed value of schedule delay (β): the ratio follows $\mu_i \equiv \alpha_i/\beta$.[6] This

[4] Following Kraus (2003), we impose that the cross-substitution effects are the same for both modes for all types of users (i.e. $E_i^r = E_i^c = E_i$, superscript c indicates the car and r the rail), and there are no income effects. Users are utility maximisers, and consequently $B_i^r \cdot B_i^c > E_i \cdot E_i$ must hold. Under these assumptions, gross consumer benefit (G_i) for type i of users is the line integral of the two inverse demands and it is independent of the path used for integration: $G_i = \int_{(0\ ,\ 0)}^{(n_i^c\ ,\ n_i^r)} \left(D_i^c[x_i^c, x_i^r]\mathrm{d}x_i^c + D_i^r[x_i^c, x_i^r]\mathrm{d}x_i^r \right) = n_i^r \left(A_i^r + \frac{B_i^r \cdot n_i^r}{2} \right) + n_i^c \left(A_i^c + \frac{B_i^c \cdot n_i^c}{2} + n_i^r E_i \right)$.

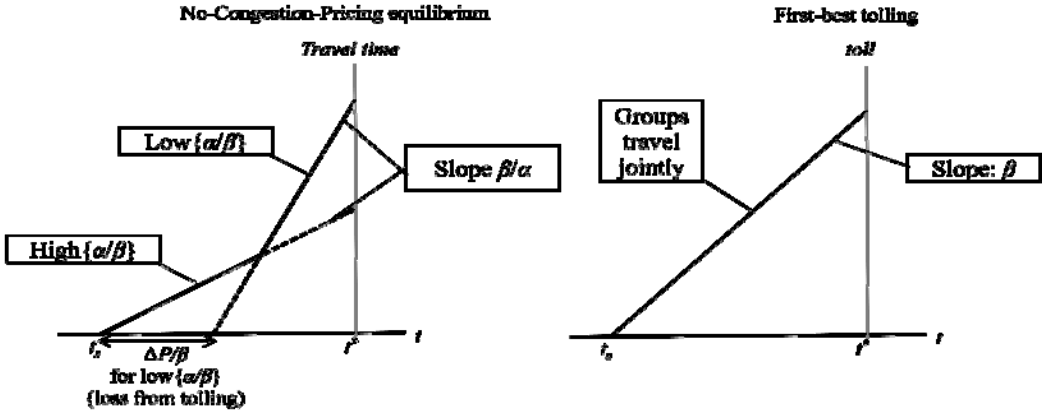
[5] The types are continuously distributed, and the equilibrium condition has to hold for all i in this continuum.

[6] Here, as is conventional in the literature, we assume $\alpha_i > \beta$ for all i, because otherwise the standard no-toll equilibrium would not hold.

heterogeneity should not be viewed as stemming from income differences, as income differences should also lead to heterogeneity in β. Hence, P_i is the price for users with a value of time of α_i. This price is the sum of the possible toll and the generalised cost. In the analytical road models, the generalised cost is the sum of travel delays cost and schedule delay cost. The numerical models will add free-flow travel time and fuel costs to this.

While our full model uses continuous distributions, it is insightful to first consider a simplified graphical example with only two groups of drivers and fixed demand. Figure 1 illustrates this situation. The left panel gives the development of travel time in the No-Congestion-Pricing (NCP) equilibrium; the right panel shows the first-best (FB) toll. The lines can be interpreted as iso-price lines (i.e. lines along which the generalised price is constant). A higher line naturally implies a higher price. The line for group i is derived from i's user-equilibrium condition that the price must be constant over time during the entire period when type i users travel, and not lower for other arrival times. In the NCP equilibrium, this condition requires travel times by arrival time to grow at a rate $1/\mu_i \equiv \beta/\alpha_i$ when group i travels. The High group has a higher ratio $\mu_H \equiv \alpha_H/\beta$. In the NCP case, the groups travel temporally separated. The high μ_H of the High group means that they choose to arrive early to avoid long travel times, and reversely for the Low group. First-best (FB) tolling removes all queuing (see, e.g., Arnott et al., 1988). This requires the toll to increase with β, since this ensures that travel delays that are always zero constitutes the dynamic equilibrium. Now the groups travel jointly, since there are only scheduling costs and tolls, which (by assumption) they value equally. See the right panel of Figure 1 for the FB toll in the example.

Figure 1: Example of ratio heterogeneity with two discrete groups



The price, when arriving at t, includes travel-time costs ($C_T[i,t]$), scheduling costs ($C_{SD}[t]$), and possibly a toll ($\tau^c[t]$). Observing that the price for a user equals the schedule delay when its group iso-price line intersects the horizontal axis, we can determine the following prices (where subscript L indicates the Low group, and H the High group):

$$P_L^{NCP} = C_T[i,t] + C_{SD}[t] = \beta \cdot \frac{n_L^c + n_H^c \cdot \mu_L / \mu_H}{s} = \beta \cdot \frac{n_L^c + n_H^c \cdot \alpha_L / \alpha_H}{s}, \quad (2a)$$

$$P_H^{NCP} = C_T[i,t] + C_{SD}[t] = \beta \cdot \frac{n_L^c + n_H^c}{s}, \quad (2b)$$

$$P_L^{FB} = C_{SD}[t] + \tau^c[t] = \beta \cdot \frac{n_L^c + n_H^c}{s}, \quad (2c)$$

$$P_H^{FB} = C_{SD}[t] + \tau^c[t] = \beta \cdot \frac{n_L^c + n_H^c}{s}; \quad (2d)$$

where n_i^c is the number of drivers of group *i*. The total number of drivers is $N^c = n_L^c + n_H^c$ The capacity of the bottleneck is *s*.

Without heterogeneity, the price would be $\beta \cdot N^c/s$ for all drivers in both regimes. Hence, ratio heterogeneity does not affect the price of the High group, but does lower the NCP price for the Low group. The high μ_H means that High users build up the queue more slowly than Low users would, lowering the price for Low users compared with the situation with N^c homogeneous users. This also means that High users impose lower externalities than Low users: the ratio of marginal effects being $(\partial P_L^{NCP} / \partial n_H^c) / (\partial P_L^{NCP} / \partial n_L^c) = \alpha_L / \alpha_H$ (see also Lindsey, 2004). A further consequence is that the weighted average of the marginal external costs is lower with ratio heterogeneity than with homogeneity.

With first-best pricing, the price is $\beta \cdot N^c/s$ for both groups. Hence, tolling does not affect the price of the High group, but raises it for the Low group by $\Delta P = \left(1 - \alpha_L / \alpha_H\right) \beta \cdot n_H^c / s$. This implies that tolling is less welfare increasing with ratio heterogeneity than with homogeneity, where prices are unaffected by tolling. With two groups, more heterogeneity means that the ratio α_L/α_H is higher (while keeping the average values fixed). With more heterogeneity, congestion externalities are lower, and tolling raises the price for the Low group more.

Arnott et al. (1988, 1994) and de Palma and Lindsey (2002) consider cases with multiple groups and fixed demand. Their results are consistent with our discussion. Without tolling, users arrive in order of decreasing ratio $\mu_i \equiv \alpha_i/\beta$. The price for a group increases with μ_i. All users lose due to first-best pricing, except the users with the highest values, who are unaffected.

Van den Berg and Verhoef (2011a) show that price sensitivity of demand changes this conclusion. Users with low ratios μ_i decrease their demand when pricing is introduced, since it increases their price. This in turn lowers the congestion, making tolling beneficial for users with high ratios. Under continuous heterogeneity, more ratio heterogeneity again has the same effects

as in the discrete case.[7] Moreover, they also find that the relative efficiency of tolling one of two parallel bottlenecks decreases with the degree of ratio heterogeneity.

These results differ from what Small and Yan (2001) and Verhoef and Small (2004) find. They use static flow congestion and two parallel links, and find that the gain of first-best pricing and the relative efficiency of single-link tolling increase with the amount of heterogeneity in the value of time. Their result is due to self-sorting by link. Drivers with high values use the link with the shorter travel time but higher monetary cost, making the shorter travel time of this link more valuable. Drivers with low values use the link with the longer travel time, making its extra travel time less costly. Consequently, the type of congestion also influences how heterogeneity affects tolling.

Following Van den Berg and Verhoef (2011a), we can generalise the equilibrium prices from the discrete group case to the case where the value of time is continuously distributed:

$$P_i^{NCP} = C_{SD}[t] + C_T[i,t] = \frac{\beta}{s} N^c \left(F^c[\alpha_i] + \alpha_i \int_{\alpha_i}^{\bar{\alpha}} \frac{f^c[\alpha_j]}{\alpha_j} d\alpha_j \right). \tag{3}$$

The distribution of car users follows the probability density function $f^c[\alpha_i]$; the density of users is $n_i^c = f^c[\alpha_i] \cdot N$. The cumulative distribution function is $F^c[\alpha_i]$, and the maximum value of time is denoted $\bar{\alpha}$. Scheduling costs for drivers of type i equal the left term between round brackets, $F^c[\alpha_i]$, multiplied by the term outside them: $C_{SD}[t] = \beta \cdot N^c \cdot F^c[\alpha_i]/s^c$. Queuing costs are the right term in round brackets multiplied by the term outside then. Again, users with a certain μ_i gain when some users with a higher $\mu_j \equiv \alpha_j/\beta$ are replaced by users with an even higher μ_k, but do not suffer if some users with a lower μ_j are replaced by users who have an even lower μ_k. It is this asymmetry that causes total costs to decreases with the degree of ratio heterogeneity.

With first-best (FB) tolling, users arrive in order of increasing value of schedule delay. However, since this value is the same for all, the order is undetermined, and the FB price in (4) is the same for all. Here, N^c is the endogenous number of drivers. Since tolling affects users with different values of time differently, it also alters the equilibrium distribution of users as demand is price sensitive.

$$P_i^{FB} = C_{SD}[t] + \tau^c[t] = \beta \cdot N^c / s \tag{4}$$

If crowding externalities in the train cannot be priced we are in a second-best world. Road pricing then uses the same formula for the time-variant part of the toll as the FB toll, but adds a

[7] Under continuous heterogeneity, we define an increase in the heterogeneity as an increasing variance of the distribution while the mean remains the same. More heterogeneity could result from more users in the tails of the distribution or the highest and lowest values becoming more extreme.

time-invariant component. With welfare maximisation, the time-invariant part is negative to attract users away from the train (similar to what Braid (1996) finds for two parallel roads). With profit maximisation, the time-invariant term is positive, and maximises toll revenue.

3.2. Proportional heterogeneity and pricing of a single bottleneck

The second type of heterogeneity that we consider concerns "proportional heterogeneity". Here, the ratio of values of time and schedule delay (α_i/β_i) is constant, but all values vary in fixed proportions following the scalar k_i: $\alpha_i \equiv a \cdot k_i$ and $\beta_i \equiv b \cdot k_i$. Our road model is adapted from Van den Berg and Verhoef (2011b). This type of heterogeneity was introduced by Vickrey (1973), and it might be interpreted as stemming from income differences.

Figure 2: Example of proportional heterogeneity with two discrete groups

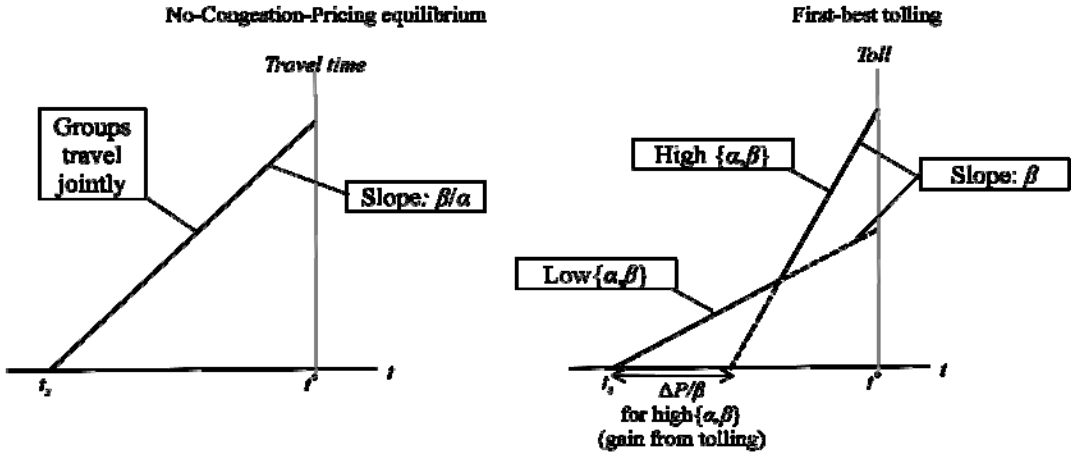


Figure 2 gives examples of the NCP and FB equilibria for two groups of users under fixed demand. Now the groups travel jointly in the NCP case, since the slope of the iso-price line is β/α, and this ratio is the same for all. First-best pricing again removes all queuing. Now, High users—here defined as users with the high α and β—arrive closest to t^*, because they are most willing to pay a toll in order to reduce schedule delays. NCP and FB equilibria prices can again be written as the schedule delay cost at the moment that the relevant iso-price line intersects the horizontal axis:

$$P_L^{NCP} = C_{SD}[i,t] + C_T[i,t] = \beta_L \cdot \frac{n_L^C + n_H^C}{s}, \tag{5a}$$

$$P_H^{NCP} = C_{SD}[i,t] + C_T[i,t] = \beta_H \cdot \frac{n_L^C + n_H^C}{s}, \tag{5b}$$

$$P_L^{FB} = C_{SD}[i,t] + \tau^c[t] = \beta_L \cdot \frac{n_L^C + n_H^C}{s}, \tag{5c}$$

$$P_H^{FB} = C_{SD}[i,t] + \tau^c[t] = \frac{\beta_L \cdot n_L^C + \beta_H \cdot n_H^C}{s}. \tag{5d}$$

The High group gains from tolling, because it shifts them to a lower iso-price curve: the intersection of its iso-price curve with the horizontal axis is closer to t^* in the right panel than in the left panel. For the Low group, the price remains the same. The High group gains because users with low values require a less steep toll schedule to prevent queuing—the slope of the toll schedule being β—and this lowers the tolls that users with high values will have to pay.

Vickrey (1973) shows that with continuous heterogeneity the same mechanisms apply. In the NCP equilibrium, the order of users is undefined. With FB pricing, users with the highest values arrive closest to t^*. Under fixed demand, users with the lowest values are unaffected by tolling, while all other users gain—and more so, the higher their values are. Tolling makes the arrival order more efficient: total scheduling costs decrease, because drivers with high values of schedule delay now arrive closest to t^*. Van den Berg and Verhoef (2011b) show that the gain from tolling increases with the proportional heterogeneity, since the gain from the more efficient arrival ordering increases. Furthermore, the relative efficiency of second-best tolling of only one of two parallel bottlenecks also increases: with more heterogeneity, pay-lane users with high values have higher mean values of time and schedule delay, making the travel time and schedule delay savings that the pay-lane offers more valuable.

With continuous heterogeneity and no tolling, the NCP prices generalise to

$$P_i^{NCP} = C_{SD}[i,t] + C_T[i,t] = \beta_i \cdot N^c / s. \tag{6}$$

The FB prices are

$$P_i^{FB} = C_{SD}[i,t] + \tau^c[t] = \frac{N^c}{s}\left(\beta_i(1 - F^c[\beta_i]) + \int_{\underline{\beta}}^{\beta_i} \beta_j f^c[\beta_j]\, d\beta_i\right). \tag{7}$$

The $\underline{\beta}$ is the minimum value of schedule delay in the distribution. The distribution of users in the FB equilibrium follows the density $f^c[\beta_i]$, the distribution function is $F^c[\beta_i]$. The distribution of users is endogenous: it depends on the tolling policy and the price sensitivity of each type of user. Although the equations might look tedious, they are entirely consistent with (5).[8]

[8] Section 3 of Arnott et al. (1988, 1994) has independent heterogeneity in values of time and schedule delay. In our terminology, this means that they consider a mix of proportional and ratio heterogeneity. This explains their finding that the welfare gain of tolling can be higher or lower with heterogeneity than with homogeneity: it depends on the relative degrees of proportional heterogeneity—which raises the gain—and ratio heterogeneity, which lowers the gain.

4. Analytical rail model

With the road model described, we now turn to the train. Congestion in public transport can take many forms, including crowding in the train, longer times to board and leave the train, and queuing on the platform to enter a train for several headways between services. As discussed below, crowding congestion seems the most relevant to consider in a model for inter-city travel. Boarding and leaving times are short relative to total travel time, and thus their costs seem minor. Queuing at the train platform for several headways between services seems rare. Conversely, crowded trains are indeed an often mentioned discomfort.

In the crowding model of Kraus (1991), the value of travel time is higher for standing than for seated passengers. Yet, crowding cost may also depend directly on the number of users: it is more unpleasant to sit or stand in a fully packed train than in a half-full train. The meta study for the UK of Wardman and Whelan (2011) confirms this. When all seats are taken, the most supported function is a value of time that increases linearly with the number of users. When there are empty seats, the value of time seems constant when most seats are empty; when many seats are taken, some studies find that the value of time increases with the number of users. Li and Hensher (2011) review international empirical studies, and find three types of crowding cost functions in the literature: (1) crowding increases the value of time (so this set-up follows Wardman and Whelan (2011)), (2) a separate per minute monetary value of crowding, and (3) a separate per trip monetary value of crowding. Consistent with case (1), Rouwendal and Verhoef (2004) and Wu and Huang (2010) use a value of time that increases with the number of users due to crowding. Instead, Huang (2000) follows cases (2) and (3), which are equivalent in his case, because the travel time is constant. He has separate crowding and travel time costs, crowding costs are linear in the number users, and the distinction between standing and seated passengers is ignored. This paper follows Huang's set-up.

The analytical models ignore travel time and the train fare that covers the marginal operating costs, since they are assumed to be constant over time. The numerical models will include these.

4.1. Ratio heterogeneity and rail pricing

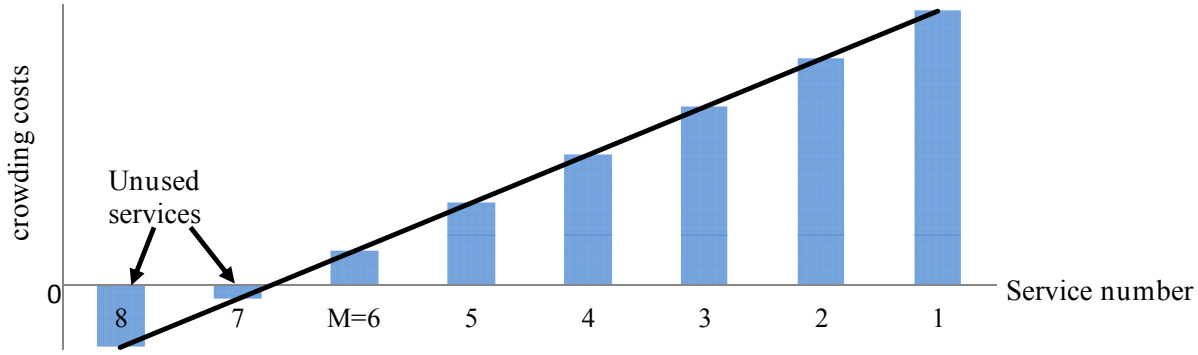
Since β is generic with ratio heterogeneity, there is no product differentiation or self-selection over time (recall that that train travel times are constant). The price for a user must be the same in all used train services. We assume that it is impossible to arrive after t^*, and denote the train that arrives at t^* as service 1; earlier services have a higher index. Hence, a service is a particular scheduled train. The headway between train services is given, and equals h. The (generalised) price of using service m is the sum of crowding cost ($C_{CR}[m]$), scheduling costs ($C_{SD}[m]$), and a possible crowding levy. The price is independent of the value of time. Crowding costs in service

m are $g \cdot N_m^t$; scheduling costs are $h(m-1)\beta$. The *g* is the crowding coefficient, N_m^r is the number of users in service *m*.

Figure 3 shows how crowding costs should vary over services to make all services equally attractive to use, as required for equilibrium. Because services 7 and 8 in the example would need negative crowding costs, they will not be used. Service 6 is the earliest service to be used. It has the highest scheduling costs and the lowest number of users and crowding cost. For the price of between two successive services *l* and *l+1* to be the same, the difference in crowding cost, $g \cdot \left(N_l^r - N_{l-1}^r\right)$, should balance the difference in scheduling cost, $\beta \cdot h$; and, hence, the difference in numbers of users is $N_l^r - N_{l-1}^r = \beta \cdot h / g$. Accordingly, crowding costs increase linearly over time.

The earliest service to be used is indicated by *M*. Note that *M* is not a policy instrument; it is determined by a user equilibrium condition. We can ignore earlier services that are unused—such as services 7 and 8 in Figure 3—since, by assumption, the provision of services is costless, and thus these empty services do not harm welfare or profits.

Figure 3: Crowding costs per service that would make each service equally attractive to use



Knowing how N_m^r varies between services, we can write the total number of users as[9]

$$N^r = \sum_1^M N_m^r = N_1^r + \left(N_1^r - \beta \cdot h / g\right) + \left(N_1^r - 2\beta \cdot h / g\right) + \ldots + \left(N_1^r - (M-1)\beta \cdot h / g\right),$$
$$= N_1^r \cdot M - (1 + 2 + \ldots + (M-1))\beta \cdot h / g = N_1^r \cdot M - ((M-1)M/2)\beta \cdot h / g$$

which implies that the number of users in service 1 is

$$N_1^r = N^r / M + ((M-1)/2)\beta \cdot h / g. \tag{8}$$

[9] The substitution of $M(M-1)/2$ for the series $(1+2+\ldots+M-1)$ follows Kraus and Yoshida (2002).

To find *M,* we first calculate for which arrival time the scheduling costs would equal the crowding costs of service 1. Since the number of services used is an integer, there is generally no service that arrives at this moment. Service *M* is then the first service to arrive after this time:

$$M = \text{Floor}\left[\frac{g \cdot N_1^r}{\beta \cdot h}\right] = \text{Floor}\left[\frac{1}{2} + \frac{\sqrt{8g \cdot N^r + \beta \cdot h}}{2\sqrt{\beta \cdot h}}\right], \tag{9}$$

where Floor[*x*] denotes the highest integer below *x*. The second expression for *M* in (9) is obtained by substituting (8) for N_1^r into the first one, and rewriting. We can next derive the NCP price for service 1 by inserting (8) and (9) into the crowding cost function. This price is also the equilibrium price for all services, since prices are constant over time:

$$P_i^{NCP} = C_{CR}[m] + C_{SD}[m] = g \cdot N_m^r + h(m-1)\beta = g \cdot N^r / M - \beta \cdot h(M-1)/2. \tag{10}$$

Although the price is the same in all used services, marginal social cost is not. The marginal external crowding cost in service *m* is $g \cdot N_m^r$. Service 1 has the highest externalities; service *M* the lowest. We find it convenient to separate the train-ticket price into two components: a train "fare" that equals marginal operating costs, and a time-varying "levy", ρ^r, that is used to manage the crowding externalities. In the first-best optimum, the levy equals marginal external crowding cost. Excluding fixed travel-time cost and marginal operating cost, the FB price is the sum of crowding cost, crowding levies, and scheduling cost:

$$P_i^{FB} = C_{CR}[m] + C_{SD}[m] + \rho^r[m] = 2g \cdot N^r / M - \beta \cdot h(M-1)/2. \tag{11}$$

The new user-equilibrium *M* differs from in (9) due to the internalisation of the externalities:

$$M = \text{Floor}\left[\frac{1}{2} + \frac{\sqrt{16g \cdot N^r + \beta \cdot h}}{2\sqrt{\beta \cdot h}}\right] \tag{12}$$

The N^r is different in the NCP and FB equilibria, because the number of users is endogenous and prices differ.

If road congestion is not priced, we are in a second-best situation. Then, the welfare-maximising time-invariant component of the levy can be expected to be negative (i.e. a subsidy). This reduces the number of car drivers, which increases welfare because marginal social cost on the road exceeds the private cost. In contrast, a private operator is likely to add a positive time-

invariant term to maximise its profits. With FB pricing of both modes, there is no extra time-invariant addition to the levy, since the time-variant levy equals the marginal externality.

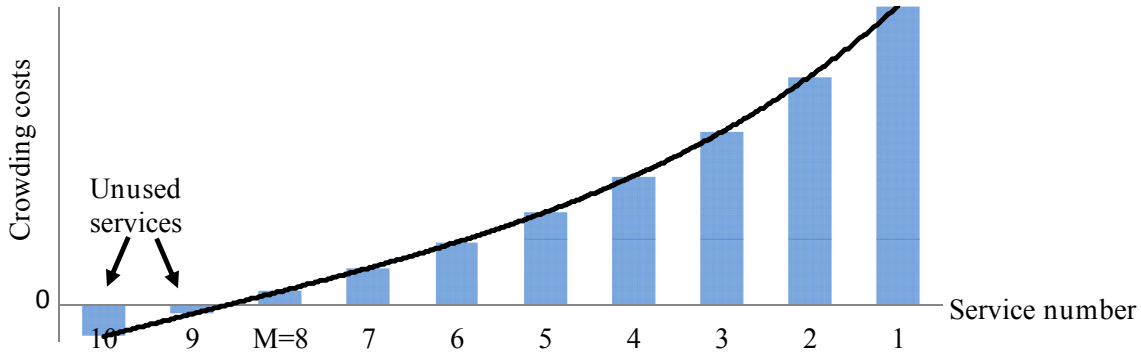
Pricing of crowding externalities increases the prices of train travel. This differs from bottleneck congestion, where tolling may lower the price for drivers with a high ratio $\mu_i \equiv \alpha_i/\beta$. But, even with homogeneous users, differences persist: FB pricing of a bottleneck would leave the price unchanged, because the queuing is a pure loss (Vickrey, 1969); with crowding congestion or dynamic flow congestion as in Chu (1995), FB pricing raises the price. Our crowding congestion, therefore, is similar to a discrete dynamic flow congestion technology.

4.2. Proportional heterogeneity and rail pricing

We have just established that ratio heterogeneity has little impact in our train model, since the travel time is fixed. With proportional heterogeneity, the α_i and β_i vary proportionally, and we can expect a greater impact. Users with the highest values use service 1 and arrive at t^*, since they care most about schedule delays. Users with the lowest values use service R. Consequently, even without congestion pricing, users arrive ordered by increasing β, while on the road this only happens with pricing. This self-ordering lowers the mean scheduling cost and price, and more so for more pronounced heterogeneity. This also puts an upward pressure on the number of used train services, since the earliest services are used by users with the lowest values of schedule delay, who are more willing to avoid crowding cost by taking an earlier train.

Crowding costs now change non-linearly over time, which is illustrated by the trend curve in Figure 4. A consequence of this is that we are unable to find closed-form solutions for the number of users per service and the number of services used. But we can find numerical solutions. In particular, for each pair of services m and $m+1$ there is a value of schedule delay that implies indifference between the two.

Figure 4: Crowding costs per service that would make each service attractive to use



Using these indifferent users it is possible to numerically solve the model. Rail-travel prices are now

$$P_i^{NCG} = C_C[m] + C_{SD}[m, \beta_i] = g \cdot N_m^r + h(m-1)\beta_i. \quad (13)$$

Crowding externalities are not affected by proportional heterogeneity, and hence service m's crowding levy still follows $g \cdot N_m^r$. The FB price is

$$P_i^{FB} = C_C[m] + C_{SD}[m, \beta_i] + \rho^r[m] = gN_m^r + h(m-1)\beta_i + gN_m^r. \quad (14)$$

5. Set-up numerical models

Our model is best illustrated using a numerical example. It is calibrated such that the NCP equilibrium has 9000 road users and 5000 train users. The bottleneck capacity is 4500 cars per hour; hence the road peak lasts 2 hours. The crowding coefficient, g, is 1/200. The fixed headway, h, between trains is 15 minutes. The mean value of time is €10.50, for the value of schedule delay it is €5.00. Free-flow car-travel time is 30 minutes. Fixed train-travel time is 45 minutes. The fuel costs of the car are €6.50. The rail fare equals per-user operating costs of €7.50. The fixed costs for the rail operator are zero, although Section 7 has a sensitivity analysis of the effect of fixed cost.

In the base-case parameterisation of proportional heterogeneity, the NCP equilibrium has a value of schedule delay that is uniformly distributed between €2.00 and €8.00, and the value of time is 2.1 times the value of schedule delay. With ratio heterogeneity, the value of time ranges between €5.50 and €15.50. Without congestion pricing, the distribution of users on the road and rail are calibrated to follow the same density function. Pricing changes the prices, and thus results in different equilibrium distributions of values of time and schedule delay for the modes.

Finally, the models are calibrated to ensure that the weighted average of the own-price elasticity, in the NCP equilibrium, is −0.5 for both car and train; the cross-elasticity of rail-travel demand with respect to the price of car travel is 0.1.[10]

[10] To achieve these goals, we adapt the calibration procedure of Van den Berg and Verhoef (2011a,b). For ease of composition we split up the own- and cross-derivatives of the inverse demand for mode j in: 1) a part that is the same for all types, and 2) a part that is also specific to type i: the derivative w.r.t. the own number of users becomes $B_i^j = b1^j / b2_i^j$ and the cross-derivative w.r.t the number of users of the other mode $E_i^j = e1^j / e2_i^j$. Note that cross derivatives of both modes should be equal, and thus $e1^j / e2_i^j = e1^k / e2_i^k = e1 / e2_i$. The constant of the inverse demand for mode j for type i is $P_i^{j,NCG} - e1_j \cdot N_k - b1_j \cdot N_j$, where $P_i^{j,NCG}$ is the NCP price and N_j the total number of users. To ensure the distribution of users, $b2_i^j$ and $e2_i^j$ equal the NCP distribution function of users. Finally, we set $b1_j$, $b1_k$, and $e1_j$=$e1_k$ such that the desired equilibrium elasticities result.

As discussed it is not possible to analytically solve the model. The numerical solution first approximates the densities of users in the car and train by cubic splines, and then alters these spline-densities iteratively using a fixed-point algorithm until the model converges to satisfy the user equilibrium condition of (1). All pricing regimes have the time-variant tolls and levies as derived in the previous two sections; the cases with an untolled alternative also have a time-invariant term, the optimisation of which takes place in a loop around the program that finds the equilibrium densities.

6. Numerical pricing model with proportional heterogeneity

We study a number of policies, which are summarised in Table 1. In the NCP regime, there is no congestion pricing and the rail fare equals marginal operating cost. In all regimes, we allow for time variation of the toll. With second-best road pricing, it is convenient to decompose the toll into a time-variant component—which eliminates the queuing—and a time-invariant term. With welfare maximisation by "pricing the road only" (CW), we find that this second term is negative, to attract users away from the suboptimally priced train. With "profit maximisation on the road only" (CP), the term is positive, to extract more revenues. The second-best schemes "train-welfare" (TW) and "train-profit" (TP) maximisation have similar set-ups. Finally, with first-best (FB) pricing, the road toll prices congestion externalities; the train has a levy that equals the crowding externalities plus a fare that equals marginal operating cost. Because we are unable to find closed-form solutions for the regimes, we present numerical results for the base-case parameterisation, and sensitivity analyses that explore the effects of heterogeneity.

Table 1: Description of the pricing regimes

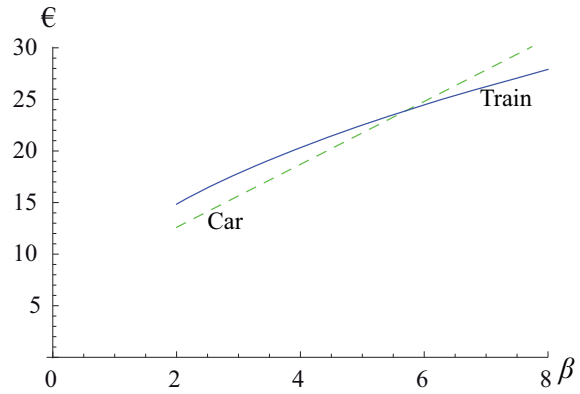
Policy	Meaning	Rail pricing	Road Pricing
NCP	No Congestion Pricing	Only a fare that equals marginal rail operating cost (v^t)	No-toll
TW	Train Welfare	Welfare-maximising second-best congestion levy + a fare that equals marginal-rail-operating cost	No-toll
TP	Train Profit	Rail-profit-maximising second-best congestion levy + a fare that equals marginal-rail-operating cost	No-toll
CW	Car Welfare	Only a fare that equals marginal-rail-operating cost	Welfare-maximising second-best congestion tolling
CP	Car Profit	Only a fare that equals marginal-rail-operating cost	Road-profit-maximising second-best congestion tolling
FB	First-best	First-best congestion levy + a fare that equals marginal-rail-operating cost	First-best congestion toll

6.1. Base case no-congestion-pricing (NCP) equilibrium

This section discusses the pricing regimes from Table 1 with proportional heterogeneity. Section 7 will consider the same regimes under ratio heterogeneity. Figure 5 shows the NCP prices by β

including fuel cost and free-flow travel time. The price of train travel is increasing in β and piecewise linear with kinks at the values that imply indifference between two services; however, these kinks are hard to detect visually. Users to the left and right of a kink use different services. Prices of car travel increase linearly with β, consistent with equation (6). Because the two modes are imperfect substitutes, they are used by users with all values of time and schedule delay.

Figure 5: Prices in the NCP equilibrium including monetary costs and free-flow travel time



6.2. Congestion pricing and proportional heterogeneity

Table 2 shows the aggregate results of the different policies. The CW only prices the road and attains 95.8% of the first-best gain, while TW only prices the train and attains only 3.4%. A first reason for these differences is that the average NCP externality on the road is €10.00, whereas in the train it is €6.04; the lower externality in the train is partly due to the lower number of users. However, the main difference between bottleneck and crowding pricing is that bottleneck congestion has a pure waste from queuing, and tolling removes this, while for the train not all congestion is a pure loss. Therefore, externality pricing is more beneficial in the bottleneck model than in the crowding model.

Figure 6 shows the change in prices of train travel due to the policies compared to the NCP equilibrium. Figure 7 does this for car travel. In both figures, the right panel is for profit maximisation to allow for differences in scale due to profit maximisation. The higher β is, the more the price of train travel increases due to congestion pricing. These users use services that arrive closer to t^*, which have higher crowding costs and thus higher levies. The curves showing the change in the price of train travel due to rail pricing are piecewise linear with upward sloping sections followed by flat sections. The flat sections represent users who use the same service as before: for them only crowding costs and congestion levies change, which they value equally. The sloping sections are for switchers, who face higher schedule delays, which are more costly the higher β. All users with low values of time and schedule delay switch services, since they care relatively little about schedule delays and more about congestion levies.

The second-best TW scheme only prices rail users. It is less harmful for train users than the first-best (FB) policy, because it adds a subsidy to the time-variant congestion levy to attract road users. In fact, the TW lowers the prices of train travel for values of schedule delay below €4.85. These users arrive far from t^*, in relatively empty trains, where the sum of time-variant and time-invariant parts of the levy is negative.

Road pricing is more likely to be regressive: the higher α and β are, the higher the gain from pricing. In contrast, the price change for rail travel increases with β, which suggest that rail pricing might be progressive. This indicates that congestion pricing might have a very different political acceptability in different modes.

Table 2: Base case effects of the policies with proportional heterogeneity

	NCP	TW	TP	CW	CP	FB
Time-invariant train levy	-	−€2.97	€22.65	-	-	-
Time-invariant road toll	-	-	-	−€0.63	€24.51	-
Mean car-travel price (E[P^C])	€21.75	€21.75	€22.55	€19.70	€41.14	€20.32
Mean train-travel price (E[P^r])	€22.15	€21.98	€ 44.79	€22.04	€22.55	€24.54
Total toll revenue	-	-	-	€34,284	€136,059	€40,062
Total crowding levy revenue	-	€1,232	€61,783	-	-	€14,394
Consumer Surplus	€414,073	€414,788	€321,672	€434,714	€277,826	€416,803
Welfare	€414,073	€416,020	€383,455	€468,997	€413,885	€471,259
Number of car driver (N^c)	9000	8995.6	9,747.6	9447.1	5065.7	9423.8
Number of train users (N^r)	5000	5014.7	2,479.9	4918.7	5724.0	4656.8
Relative efficiency	0	0.034	−0.535	0.958	−0.003	1
Percentage welfare gain	-	0.47%	−7.39%	13.26%	−0.05%	13.81%
Percentage of NCP users who would be better off	-	81.4%	0%	100%	0%	49.5%

Figure 6: Change in train-travel prices due to the implementation of pricing regimes

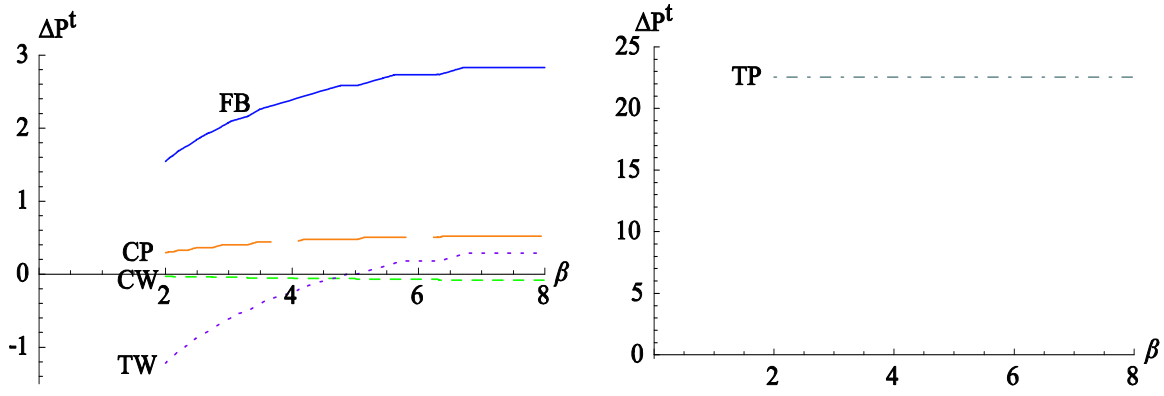
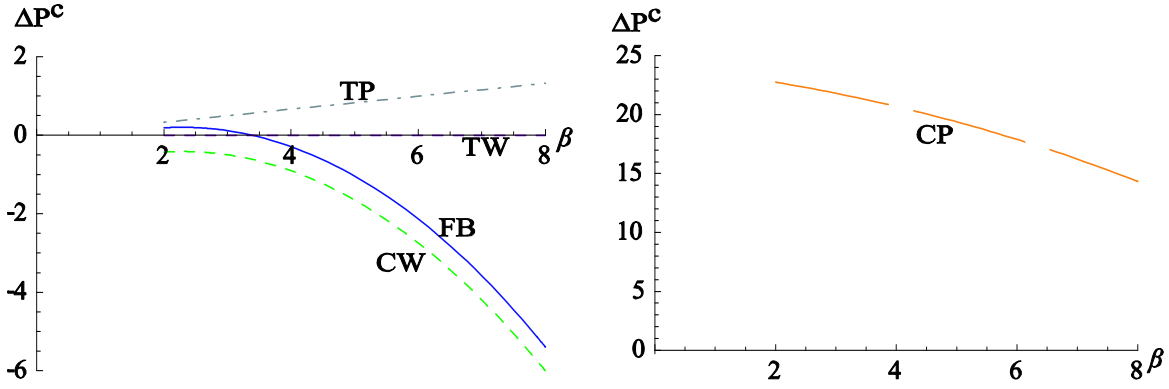


Figure 7: Change in car-travel prices due to the implementation of pricing regimes



6.3. Sensitivity analysis on the amount of proportional heterogeneity

Figures 6 and 7 suggest that heterogeneity matters. A natural follow up question is how sensitive the results are the degree of heterogeneity, which we measure by the spread of the uniform distribution of the value of schedule (β). Without congestion pricing, the average price for car travel is unaffected by the spread, which is in accordance with equation (6). The mean price of train travel decreases with the spread due to the self-ordering of users: train users with higher values arrive closer to t^* when delays are low; the opposite holds for users with low values.

Table 3 shows consumer surplus and prices in the NCP equilibria for distributions with spreads of 0, 3, 5, 6, and 7. A spread of 0 means homogeneous users. All distributions of β have a mean of €5.00. Consumer surplus decreases with the spread, but this is a result of the calibration, it does not reflect a meaningful effect of heterogeneity.

Table 4 gives the results of FB pricing for different amounts of heterogeneity. FB prices of car travel decrease, and welfare increases, with the degree of proportional heterogeneity, because the gain from the more efficient arrival order increases. This is consistent with the single mode case of Van den Berg and Verhoef (2011b). The FB price of train travel also decreases with the spread. This is partly because the ordering on β reduces scheduling cost; and the larger the spread is, the larger the reduction. Another reason is that road tolling lowers the average price of car travel, thereby lowering demand for train travel.

Table 3: Proportional heterogeneity and the no-congestion-pricing (NCP) case

Spread of the value of schedule delay	Consumer surplus	E[P^C]	E[P^r]
0	€418,250	€21.75	€23.19
3	€416,241	€21.75	€22.69
5	€414,814	€21.75	€22.34
6 (Base case)	€414,073	€21.75	€22.15
7	€413,260	€21.75	€21.94

Table 4: Heterogeneity in the value of schedule delay and first-best FB pricing

Spread of the value of schedule delay	E[P^C]	E[P^r]	Consumer Surplus	Toll revenue	Levy revenue	Welfare	%ΔW	N^c	N^r
3	€21.06	€25.39	€409,819	€43,004	€15,708	€468,531	12.6%	9258	4672
5	€20.56	€24.84	€414,433	€41,046	€14,842	€470,321	13.4%	9369	4662
6 (Base case)	€20.32	€24.54	€416,803	€40,062	€14,394	€471,259	13.8%	9424	4657
7	€20.08	€24.22	€419,138	€39,076	€13,902	€472,116	14.2%	9479	4651

Table 5 gives percentage welfare gains for the other policies. Table 6 reports the corresponding relative efficiencies. The welfare gain of second-best CW road pricing increases with the spread. Still, its relative efficiency decreases slightly. This differs from with second-best pricing of two roads that are perfect substitutes in Van den Berg and Verhoef (2011b), where the relative efficiency increases with the degree of proportional heterogeneity. The difference lies in that train and car are imperfect substitutes. With perfect substitutes, users with the highest α and β flock to the priced link, since it is most beneficial to them. With imperfect substitutes, the priced link is used by all values of time and schedule delay, even though for many values the unpriced link has a lower price and tolling raises the car-travel price. Furthermore, with imperfect substitutes, there are also users with high values who continue to use the unpriced train, even though it has a higher price. These effects lower the relative efficiency; and, accordingly, for two modes, the relative efficiency of CW pricing decreases with the amount of proportional heterogeneity if the degree of substitutability is low enough.

Table 5: Effect proportional heterogeneity on percentage welfare gains

Spread of β	TW	TP	CW	CP	FB
0	0.62%	-7.41%	11.01%	-1.55%	11.42%
3	0.50%	-7.41%	11.97%	-0.87%	12.56%
5	0.48%	-7.40%	12.85%	-0.34%	13.38%
6	0.47%	-7.39%	13.23%	-0.05%	13.81%
7	0.45%	-7.39%	13.57%	0.25%	14.24%

Table 6: Effect of proportional heterogeneity on relative efficiencies

Spread of β	TW	TP	CW	CP
0	0.054	-0.649	0.966	−0.136
3	0.040	-0.590	0.963	−0.069
5	0.036	-0.553	0.960	−0.025
6 (Base case)	0.034	-0.532	0.958	−0.003
7	0.032	-0.519	0.953	0.017

Conversely, the relative efficiency of private CP road pricing increases with proportional heterogeneity. The difference between the CW and CP arises because with profit maximisation the number of drivers drops substantially and primarily high-values drivers continue to use the road. Hence, the effect of the imperfect substitutes is less pronounced with profit maximisation.

Moreover, more proportional heterogeneity increases the mean values of time and schedule delay on the private road, making its travel time and schedule delay savings more valuable. Finally, more heterogeneity lowers the prices of train travel, increasing the competition that the private road faces, and thereby lowering the mark-up that the firm can ask.

The relative efficiency of private TP (rail-only) pricing increases with the spread. This occurs predominantly because the FB percentage-gain increases, which makes the TP's welfare loss relatively smaller. TP's welfare only increases slightly. With a larger spread, the lowest values of schedule delay are lower. This enables users with low values to use earlier train services. This easier shift to earlier services limits the market power of the operator, inducing a lower time-invariant levy. TW second-best rail pricing is hardly affected by the heterogeneity.

6.4. Conclusions on proportional heterogeneity

Time-variant tolling on the road makes the arrival order more efficient by lowering scheduling costs. In the train, this extra efficiency gain does not occur, since users always arrive ordered by β. Due to the extra efficiency gain on the road, the welfare gain of first-best pricing increases with the degree of proportional heterogeneity. The relative efficiency of "welfare maximisation by only pricing the road", however, can decrease with proportional heterogeneity; whereas with two perfect substitute roads, this relative efficiency increases. Since crowding-congestion pricing does not alter the arrival order in the train, proportional heterogeneity has little effect on the relative performance of schemes that only price the train.

7. Numerical pricing model with ratio heterogeneity

We now turn to "ratio heterogeneity", which involves variations in the ratio of the value of time (α) to the value of schedule delay (β), with a fixed value of schedule delay: $\mu_i \equiv \alpha_i/\beta$. In the NCP case, the value of time is uniformly distributed between €5.50 and €15.50. Section 2 already noted that road externalities decrease with ratio heterogeneity, which lowers the gain from road tolling. The mean price of rail travel is unaffected by ratio heterogeneity, because the travel time of the train is fixed. The price of train travel for a user is now constant across services; with proportional heterogeneity, this price differed over services due to the differences in β_i.

7.1. Base case no-congestion-pricing (NCP) and first-best equilibria with ratio heterogeneity

Figure 8 gives the NCP prices including free-flow travel time and monetary costs. The price of rail travel increases linearly with the value of time when accounting for the fixed train travel time. The price of car travel increases concavely, and is highest for drivers with the highest ratio, which is in accordance with equation (3). In this calibration, prices for train travel are higher than

for car travel for all users. Nevertheless, the two modes are imperfect substitutes, and thus all values of time are represented in the train.

Again, road pricing gives a higher welfare gain than rail pricing. As Table 7 shows, CW pricing attains 96% of the first-best FB gain. The TW policy attains only 2%. The train has about 35% of the users, and consequently this large difference in welfare gain reflects more than just differences in the percentage of users who face congestion pricing.

Figure 8: Prices including monetary costs and free-flow travel time

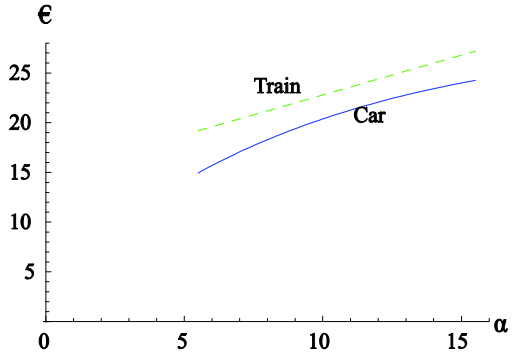


Table 7: Base case effects of the policies with ratio heterogeneity

	NCP	TW	TP	CW	CP	FB
Time-invariant train levy	-	−€2.02	€23.72	-	-	-
Time-invariant road toll	-	-	-	−€0.68	€22.17	-
Mean car-travel price (E[P^C])	€20.43	€20.47	€21.16	€20.99	€39.40	€21.64
Mean train-travel price (E[P^T])	€23.19	€24.30	€46.87	€23.19	€23.69	€26.13
Total toll revenue	-	-		€37,843	€120,877	€43,603
Total crowding levy revenue	-	€6,588	€64,949	-	-	€15,717
Consumer Surplus	€398,705	€392,906	€303,318	€394,216	€265,242	€374,120
Welfare	€398,705	€399,494	€368,267	€432,059	€386,119	€433,441
Number of car driver (N^c)	9000	9035.5	9757.5	8887.7	4860.4	8859.2
Number of train users (N^r)	5000	4881.6	2475.9	5018.3	5677.6	4726.0
Relative efficiency	0	0.023	−0.876	0.960	−0.362	1
Percentage welfare gain	-	0.20%	−7.63%	8.37%	−3.16%	8.71%
Percentage of NCP users who would be better off	-	0%	0%	30.3%	0%	13.8%

Congestion pricing is less beneficial to the average user with ratio heterogeneity than with proportional heterogeneity or homogeneity. This is consistent with the two group example in Figure 1. With proportional heterogeneity, consumer surplus increases due to TW, CW, and FB policies. With ratio heterogeneity, only the second-best road pricing scheme CW increases consumer surplus. The percentage of NCP (No-Congestion-Pricing) users who would face a lower price with pricing is also lower with ratio heterogeneity than with proportional heterogeneity. Table 7 shows that, with ratio heterogeneity, no user gains from TW pricing; with

proportional heterogeneity in Table 2, all car drivers and 47% of the rail users gain. Likewise, with ratio heterogeneity, 30% of NCP users gain from CW pricing; with proportional heterogeneity, all users gain. Finally, for all schemes, the welfare gain is lower with ratio heterogeneity than with proportional heterogeneity.

Figures 9 and 10 plot the price changes, compared to the NCP case, for train and car users. For train travel, the curves are flat lines, because only the cost of the fixed travel time depends on α_i. This contrasts with proportional heterogeneity, where these curves are piecewise linear.

All rail users are affected to the same extent by pricing. Hence, pricing does not have distributional effects between rail users. Conversely, on the road there is a clear distributional pattern: road pricing is less harmful for a driver with a higher the value of time (α), because it lowers the car travel time. The rail-only pricing TW and TP raise travel times on the road, and are therefore more harmful for drivers with a higher α. That a policy that increases travel time is less harmful with a lower value of time is logical, and was also found by Liu and Nie (2011), who study pricing in a large network and with two modes. It is worth noting that the signs of the slopes of the curves for drivers in Figures 7 and 10 are the same. In this regard, the distributional effects for drivers are thus similar for the two types of heterogeneity. But, as explained above, the signs of the effects (gains or losses) may differ.

Figure 9: Change in train-travel prices due to the pricing regimes

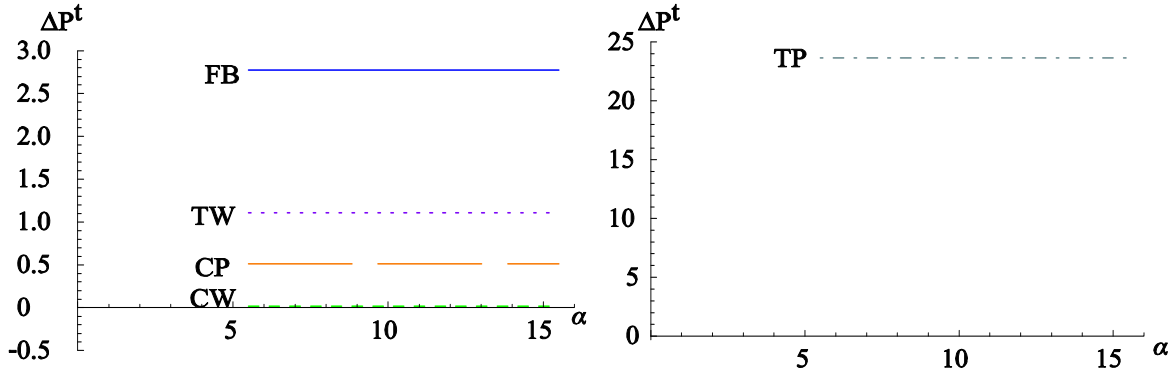
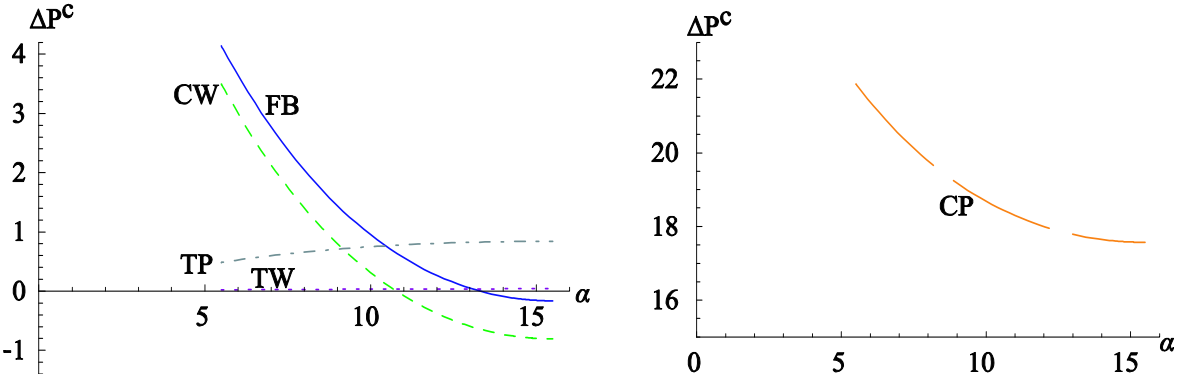


Figure 10: Change in car-travel prices due to the pricing regimes



7.2. Sensitivity analysis on the amount of ratio heterogeneity

Table 8 analyses how the degree of ratio heterogeneity affects consumer surplus and the average prices. The base-case spread of the uniform distribution of α is 10; the other spreads are 10.8, 9.2, 6, and 0.[11] The mean value of time is always €10.05. Consumer surplus increases slightly with the spread, which is also what we found with the proportional heterogeneity. But again this is due to the calibration, and does not represent a meaningful effect of heterogeneity. The average price of car travel decreases with the spread of α, because the congestion externalities decrease. The mean price of train travel is unaffected, because the train travel time is fixed.

Table 8: Ratio heterogeneity and the no-congestion-pricing (NCP) case

Spread of the value of time	CS	$E[P^C]$	$E[P^r]$
0	€418,250	€21.75	€23.19
6	€405,734	€20.78	€23.19
9.2	€400,052	€20.52	€23.19
10 (base case)	€398,705	€20.43	€23.19
10.8	€396,925	€20.34	€23.19

Table 9 depicts the effects of the spread of α on first-best pricing. In contrast to proportional heterogeneity, ratio heterogeneity causes FB pricing to raise the average price for car travel (see Table 8). This was already implied by equations (3) and (4). Still, the price increase for rail travel is larger. This is, again, because congestion pricing is less beneficial with crowding congestion than with bottleneck congestion. The percentage FB welfare gain decreases with the spread. This is, as Section 2 argued, because externalities on the road decrease, which means that there is less to gain from tolling.

Table 9: Ratio heterogeneity and first-best pricing

Spread of the value of time	$E[P^C]$	$E[P^r]$	Toll revenue	Fare revenue	CS	Welfare	$\%\Delta W$	N^c	N^r
6	€21.71	€26.09	€44,461	€15,656	€384,481	€444,598	9.6%	8946.0	4712.9
9.2	€21.65	€26.13	€43,770	€15,706	€376,097	€435,573	8.9%	8876.1	4723.5
10 (base case)	€21.64	€26.13	€43,603	€15,717	€374,120	€433,441	8.7%	8859.2	4726.0
10.8	€21.60	€26.13	€43,448	€15,726	€371,768	€430,942	8.6%	8843.4	4727.9

Tables 10 and 11 investigate the effects of the spread of the value of time on the performance of the other policies. We again see that the gain from FB pricing decreases with the spread. The relative efficiencies of schemes that only price the road decrease with the degree of ratio heterogeneity. This corresponds with Van den Berg and Verhoef (2011a), where the relative efficiency of "single-link pricing" reduces with the degree of this heterogeneity. The welfare gain

[11] For a deterministic bottleneck equilibrium $\alpha_i > \beta$ must hold for all i. Hence, the maximum spread is 10.99. However, for a spread above 10.8, the numerical model is unstable. Therefore, the used maximum spread is 10.8.

of TP is hardly affected by ratio heterogeneity. Still, its relative efficiency is more negative with a higher degree of ratio heterogeneity, because the first-best gain decreases, decreasing the denominator of the relative efficiency. The welfare gain and relative efficiency of TW marginally decrease with the spread of α, as road externalities are smaller, and thus there is less to gain from attracting car drivers away from the road.

Table 10: Effect of ratio heterogeneity on percentage welfare gains

Spread of α	TW	TP	CW	CP	FB
0	0.62%	-7.41%	11.01%	-1.55%	11.42%
6	0.33%	-7.59%	9.24%	-2.58%	9.58%
9.2	0.23%	-7.62%	8.53%	-3.05%	8.88%
10	0.20%	-7.63%	8.37%	-3.16%	8.71%
10.8	0.19%	-7.65%	8.56%	-3.26%	8.57%

Table 11: Effect of ratio heterogeneity on relative efficiencies

Spread of α	TW	TP	CW	CP
0	0.054	-0.649	0.966	-0.136
6	0.034	-0.792	0.965	-0.269
9.2	0.026	-0.859	0.961	-0.343
10 (base case)	0.023	-0.876	0.960	-0.362
10.8	0.022	-0.893	0.958	-0.381

7.3. Conclusions on ratio heterogeneity

The welfare gains of all pricing schemes decrease when the degree of ratio heterogeneity (in α_i/β) increases. Moreover, the relative efficiency of pricing only the road or train also decreases. With two roads, this type of heterogeneity had the same effect on the performance of first-best and second-best pricing. This contrasts with Section 4, where the effect of proportional heterogeneity differs between a two-roads and two-modes network.

8. Some further sensitivity analyses

8.1 Operational cost

Tabuchi (1993) analysed road and rail pricing when the rail operator's costs have a variable and fixed component; we assumed that it only has variable costs. To test the effects of this limitation, we redefined our models such that rail only has fixed cost. The fixed costs were set such that the average-operating cost in the NCP equilibrium equals the one in the foregoing simulations.

The effects of congestion pricing, and how heterogeneity affects these, remain roughly the same. The most notable difference is that, with fixed cost and second-best road-pricing (CW), it is beneficial to push users to the train since the fare is based on average operation cost instead of marginal operation cost. This result was also found by Tabuchi (1993). In the previous sections,

the opposite effect (see, e.g., Braid, 1996) was found: the priced road had a negative term in its toll to attract users from the crowded train.

Alternatively, there could be a variable cost to providing train services (instead of the cost per user). The number of services would then be a discrete capacity variable, which would be set following cost-benefit analysis. It would then also be necessary to consider the optimisation of road capacity. See, for example, Arnott et al. (1988, 1994), who study on the effects of capacity expansion in the bottleneck model under heterogeneity; Arnott and Yan (2000) on (second-best) capacity setting in a road and rail network; and Yang et al. (2001) on frequency and fare setting of two bus operators when there is competition from the untolled car and heterogeneity in the value of time.

8.2 Price elasticities

The effects of variations in the own-price elasticities are in line with the results in the earlier literature. In particular, private pricing is increasingly harmful with less elastic demand, because the company's market power increases. Changing the cross-elasticity is more interesting. It could be that the limited relative efficiency of second-best TW train pricing that we found stemmed from a low cross-elasticity, since this makes it difficult to attract drivers away from the unpriced road.

In the ratio heterogeneity model, the highest cross-elasticity of rail demand w.r.t. the price of car travel consistent with utility maximisation is 0.379 (see also footnote 4). In the base case, this elasticity was 0.1. This almost quadrupling raises the TW's relative efficiency by 314% to 0.095. Nevertheless, it remains small and far below the relative efficiency of road-only pricing. With proportional heterogeneity, the maximum cross-elasticity is 0.37. Here the relative efficiency for TW increases by 201% to 0.10. In conclusion, the limited cross-elasticity is not the most important reason for the low gain from rail pricing.

8.3 Crowding costs

Another reason for the low gain of rail pricing could be a relatively low value of crowding. Although it seems likely that crowded trains indeed cause a discomfort, we have little empirical information on the value of crowding for our model. To gain insight, we double the crowding coefficient g. This less than doubles total crowding costs, because users respond by arriving more spread out over the day. In the ratio heterogeneity model, the mean NCP price increases by €3.27 to €26.46; but the relative efficiency of TW pricing only increases by 11% to 0.025. The relative efficiency of the private TP increases from −0.876 to −0.38. With proportional heterogeneity, the doubling of g raises TW's relative efficiency by 17% to 0.04.

To conclude: the value of crowding congestion does have an effect, but the primary reason for the low gain from rail pricing relative to road pricing seems to be the difference between bottleneck and crowding congestion. With bottleneck congestion, pricing removes the pure waste that is queuing. The train has no such entirely wasteful congestion. The crowding congestion model, therefore, behaves like a discrete version of a dynamic flow congestion model.

9. Conclusion

This paper analysed congestion pricing in a road and rail network with heterogeneous users, where the train and car are imperfect substitutes. We separately studied "ratio heterogeneity", in the ratio of value of time to value of schedule delay, and "proportional heterogeneity", which varies both these values in fixed proportions.

With proportional heterogeneity, road pricing makes the arrival order more efficient, thereby lowering scheduling costs. Accordingly, the welfare gain of road pricing increases with this type of heterogeneity. The relative efficiency of private road-only pricing rises with proportional heterogeneity, while for welfare maximisation the relative efficiency decreases. This contrast with earlier research for two roads, which found that the relative efficiency of "single-link pricing" increases with the proportional heterogeneity (Van den Berg and Verhoef, 2011b). The difference was found to be caused by the fact that two roads are perfect substitutes, while car and train are not in our setting.

On the rail link, users always arrive in order of increasing value of schedule delay. Thus, pricing cannot improve the arrival order. The welfare gain of rail-congestion pricing is hardly affected by proportional heterogeneity. Still, the relative efficiency of private rail-only-pricing slightly decreases with this type of heterogeneity.

A general conclusion is that the gain of congestion pricing can be lower or higher with heterogeneity depending on the types and extents of the heterogeneity. Nevertheless, pricing always leads to a welfare improvement. Perhaps more important for policy are the distributional effects, which differ substantially between road and rail. In the train, pricing has no distributional effects or is more harmful for users with higher values of time and schedule delay. On the road, generally, users with a higher value are more likely to benefit from road pricing. Conversely, private rail pricing is more damaging for road users with higher values of time or schedule delay, because it increases congestion on the road, and this is more costly with higher values. The distributional effects thus need not follow the expected pattern, where a higher value of time or schedule delay makes a user better off with congestion pricing.

An interesting extension of our model would be to consider heterogeneity in the value of crowding. Then, users with high crowding values would arrive early in relatively empty trains, and low-value users would arrive close to the preferred arrival time. Hence, crowding

externalities will be lower with heterogeneity, thus lowering the gains from rail pricing. Accordingly, the effect of value-of-crowding heterogeneity in the rail model might be similar to that of ratio heterogeneity in the bottleneck model. Then it would also be interesting to introduce first and second class coaches, since users could self-select by value of crowding, thereby lowering total costs. If we used a crowding multiplication of the value of time—following some of the empirical literature—instead of separate crowding and travel time costs, the effect of heterogeneity on the performance of externality pricing in the train might be similar to that on road pricing. Another interesting extension would be to study a larger network, to see how the policies of this paper perform in a more realistic setting.

Acknowledgements

We thank the reviewers, Robin Lindsey, Stef Proost, Jan Rouwendal, Erik Kroes, Mrs. Ellman, and Muhammed Sabir for their helpful comments. This research was financially supported by Transumo and ERC (AdG Grant #246969 OPTION). The usual disclaimer applies.

References

Arnott, R., de Palma, A., Lindsey, R., 1988. Schedule delay and departure time decisions with heterogeneous commuters. *Transportation Research Record*, 1197, 56–67.

Arnott, R., de Palma, A., Lindsey, R., 1994. The welfare effects of congestion tolls with heterogeneous commuters. *Journal of Transport Economics and Policy*, 28 (2), 139–161.

Arnott, R., Yan, A., 2000. The two-mode problem: Second-best pricing and capacity. *Review of Urban and Regional Development Studies*, 12 (3), 170–199.

Braid, R.M., 1996. Peak-load pricing of a transportation route with an unpriced substitute. *Journal of Urban Economics*, 40 (2), 179–197.

Chu, X., 1995. Endogenous trip scheduling: the Henderson approach reformulated and compared with the Vickrey approach. *Journal of Urban Economics*, 37 (3), 324–343.

Cohen, Y., 1987. Commuter welfare under peak period congestion: Who gains and who loses? *International Journal of Transport Economics*, 14 (3), 239–266.

de Palma, A., Lindsey, R., 2002. Congestion pricing in the morning and evening peaks: a comparison using the Bottleneck Model. In: *Proceedings of the 39th Annual Conference of the Canadian Transportation Research Forum: 2002 Transportation Visioning - 2002 and Beyond*, Vancouver, Canada, 9–12 May 2002, pp. 179–193.

Gonzales, E.J., Daganzo, C.F., 2012. Morning commute with competing modes and distributed demand: user equilibrium, system optimum, and pricing. *Transportation Research Part B*, 46 (10), 1519−2153.

Hendrickson, C., Kocur, G., 1981. Schedule delay and departure time decisions in a deterministic model. Transportation science, 15 (1), 62–77.

Huang, H.-J., 2000. Fares and tolls in a competitive system with transit and highway: the case with two groups of commuters. *Transportation Research Part E*, 36 (4), 267–284.

Kraus, M., 1991. Discomfort externalities and marginal cost transit fares. *Journal of Urban Economics*, 29 (2), 249–259.

Kraus, M., 2003. A new look at the two-mode problem. *Journal of Urban Economics*, 54 (3), 511–530.

Kraus, M., Yoshida, Y., 2002. The commuter's time-of-use decision and optimal pricing and service in urban mass transit. *Journal of Urban Economics*, 51 (1), 170–195.

Layard, R., 1977. The distributional effects of congestion taxes. *Economica*, 44 (175), 297–304.

Li, Z., Hensher, D.A., 2011. Crowding and public transport: a review of willingness to pay evidence and its relevance in project appraisal. *Transport Policy*, 18 (6), 880–887.

Lindsey, R., 2004. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes. *Transportation Science*, 38 (3), 293–314.

Liu, Y., Nie, Y., 2011. Welfare effects of congestion pricing and transit services in multi-class multi-modal networks (version of 14 November 2011). In: *Transportation Research Board, the 91th Annual Meeting of the Transportation Research Board*. Washington D.C., USA.

Rouwendal, J., Verhoef, E.T., 2004. Second best pricing for imperfect substitutes in urban networks. *Research in transportation economics*, 9, 27–60.

Small, K.A., Verhoef, E.T., 2007. The Economics of Urban Transportation. London, Routledge.

Small, K.A., Yan, J., 2001. The value of "value pricing" of roads: second-best pricing and product differentiation. *Journal of Urban Economics*, 49 (2), 310–336.

Tabuchi, T., 1993. Bottleneck congestion and modal split. *Journal of Urban Economics*, 34 (3), 414–431.

van den Berg, V.A.C., Verhoef, E.T., 2011a. Congestion tolling in the bottleneck model with heterogeneous values of time. *Transportation Research Part B*, 45 (1), 60–70.

van den Berg, V.A.C., Verhoef, E.T., 2011b. Winning or losing from dynamic bottleneck congestion pricing? The distributional effects of road pricing with heterogeneity in values of time and schedule delay. *Journal of Public Economics*, 95 (7–8), 983–992.

Verhoef, E.T., Small, K.A., 2004. Product differentiation on roads: constrained congestion pricing with heterogeneous users. *Journal of Transport Economics Policy*, 38 (1), 127–156.

Vickrey, W.S., 1969. Congestion theory and transport investment. *American Economic Review (Papers and Proceedings)*, 59 (2), 251–260.

Vickrey, W.S., 1973. Pricing, metering, and efficiently using urban transportation facilities. *Highway Research Record*, 476, 36–48.

Wardman, M., Whelan, G.A., 2011. 20 Years of rail crowding valuation studies: evidence and lessons from British experience. *Transport Reviews*, 31 (3), 379–398.

Wu, W.-X., Huang, H.-J., 2010. Equilibrium and modal split in a competitive highway/transit system under different road-use pricing regimes. *Workingpaper of the Beijing University of Aeronautics and Astronautics*, version of 2 december 2010.

Yang, H., Kong, H.Y, Meng, Q., 2001. Value-of-time distributions and competitive bus services. *Transportation Research Part E*, 37 (6), 411–424.