Why congestion tolling could be good for the consumer:

The effects of heterogeneity in the values of schedule delay and time on the effects of tolling

Version of 19 June 2010

Vincent van den Berg^{a*}, Erik T. Verhoef$^{a \#}$

a: Department of Spatial Economics, VU University, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands
#: Email: everhoef@feweb.vu.nl
*: Corresponding author: email: vberg@feweb.vu.nl, tel: +31 20 598 6049, fax: +31 20 598 6004

Abstract

We analyse the efficiency and distributional impacts of congestion pricing, in the bottleneck model, with continuous heterogeneity in the values of time and schedule delay. With a heterogeneous value of schedule delay, tolling makes the arrival ordering more efficient, and this lowers scheduling costs. If there is not much more heterogeneity in the value of time than in the value of schedule delay, then first-best tolling decreases the generalised price for most users. We find that the consumer surplus losses or gains from first-best tolling are not strictly monotonic in the value of time, because the value of schedule delays also determines such gains or losses. The greatest losses are not incurred by drivers with the lowest value of time, but by users with an intermediate value of schedule delays and the lowest value of time consistent with that value of schedule delays. The lowest values of time are among those who gain most from a public pay-lane. With a private pay-lane, the lowest value of time loses least of all free-lane users; moreover, there are also many pay-lane users that lose more than the lowest value of time.

JEL codes: D62, H23, L11, R41, R48
Keywords: Traffic congestion, Road pricing, Heterogeneity, Distributional impacts, Bottleneck model

1. Introduction

Road pricing seems to be gaining increasing momentum as an instrument in dealing with traffic congestion. The concept is firmly based in micro economic theory: Pigou (1920) recognised that traffic congestion entails an external cost, and that efficiency requires a toll equal to marginal external congestion cost.

Despite the strong economic case for road pricing, practical applications remain scarce. An important reason why road pricing meets resistance is its redistributive effect. When pricing reduces travel times and increases monetary costs, an individual’s losses are smaller, or gains are larger, when the value of time is higher. Although there is no perfect correlation between value of time and income, this is often taken as implying that the poor lose and the rich gain (Layard, 1977). Still, as became clear from an early discussion between Foster (1974, 1975) and Richardson (1974), it is not certain how the benefits or losses from road pricing vary over income. Even if higher incomes have higher values of time on average, and consequently lose less due to tolling per driven kilometre. They also drive more and have higher car ownership and longer commutes. Hence, they lose longer and more often. The net balance of these opposing forces is an empirical matter. Foster concluded that road pricing may be progressive; Richardson maintained that regressiveness is more likely. In Mayeres and Proost (2001), it is because higher incomes drive more, why the loss from road pricing *increases* with income.

Cain and Jones (2008) show that car ownership (in Scotland) is relatively low in the lowest income quintile. Still, the lowest incomes who do own a car spend on average about 40% of disposable income on motoring. Therefore, even if road pricing has a limited impact on the average low-income household, it may still have a big impact on those who own a car.

Small and Yan (2001) and Verhoef and Small (2004) use static flow congestion. The distributive impacts of first-best pricing are as expected: drivers' losses decrease with the value of time. Yet, with second-best pricing on only some of the lanes of the highway, the distributional effect is not monotonous in the value of time. The biggest losses are incurred by users with the 'critical' value of time, who are indifferent between the tolled (pay-lane) and untolled (free-lane) link. On the free-lane, the cost of the higher travel times decreases with the value of time, while the gain from the pay-lane's reduced travel time increases with the value of time. Hence, the public pay-lane's relative efficiency increases with heterogeneity in the value of time. Relative efficiency is the welfare gain of a policy from the no-toll situation relative to the public first-best toll's gain. Conversely, Van den Berg and Verhoef (2010) note that, with bottleneck congestion, a pay-lane's relative efficiency decreases with this heterogeneity. Further, with a more heterogeneous value of time, no-toll equilibrium congestion externalities are smaller. Consequently, the welfare gain of first-best or pay-lane tolling is lower.

Empirical evidence, see for example Small, Winston and Yan (2005), shows that in reality values of time differ substantially, reinforcing the case for considering heterogeneity explicitly in road pricing analyses. But it is not only the value of time that determines the disutility of congestion and welfare effects from pricing. Dynamic models of traffic congestion emphasise the importance of schedule delay costs.

In Vickrey (1973), the values of time and schedule delays vary proportionally. All drivers gain due to first-best tolling—except the lowest values, who are unaffected. Arnott et al. (1988; 1994) use two group heterogeneity in the value of time and schedule delay. The imposition of a first-best toll may change the order of arrivals. In the no-toll equilibrium, drivers with a higher ratio of value of schedule delay to value of time arrive closer to the common most desired arrival time. With optimal tolling, the higher value of schedule delay do so—and this need not be the same group. Apart from removing all travel delays, the optimal toll then also reduces total schedule delay costs by making the arrival order more efficient. In terms of distributional impacts, a time-variant toll without rebate "benefits drivers with high unit travel time and schedule delay costs" (Arnott et al., 1994, p. 158). De Palma and Lindsey (2002) use the bottleneck model and study heterogeneity in the value of time with fixed values of schedule delay. With perfectly inelastic demand, all users lose due to first-best tolling—except the highest values, who are unaffected.[1]

We analyse *first-best tolling* and *public* and *private pay-lanes* when the values of schedule delay and time are heterogeneous. A type *i* driver faces a schedule delay if she does not arrive at her preferred arrival time (t^*). The value of earlier arrival (*schedule delay early*) is β_i and of later (*schedule delay late*) it is γ_i. There is no heterogeneity between the values of schedule delay early and late. We analyse proportional heterogeneity—ala Vickrey (1973)—and heterogeneity in the ratio (μ) of values of time (α) and schedule delay: $\mu_i = \alpha_i / \beta_i$. Proportional heterogeneity scales the values of time and schedule delay. It can be viewed as an effect from income: the higher income is, the higher the values. Ratio heterogeneity can, for example, be viewed as giving the effect of schedule constraints from work (e.g. office versus assembly-line worker) or from home-life (e.g. children or not) or from extra scarcity of time due to an hectic family or social life.

More ratio heterogeneity decreases marginal external costs from congestion and thus lowers the welfare gain from tolling. This welfare gain increases with proportional heterogeneity (i.e.

[1] The last conclusion on the very highest value of time follows personal communication by email with Robin Lindsey on 17 June 2010 and the results of Arnott et al. (1994) and Van den Berg and Verhoef (2010a).

where the values of time and schedule delay vary proportionally), this is regardless of the amount of heterogeneity in the ratio of the value of time to value of schedule delay.

The average generalised price can be lower with first-best tolling than without tolling, if the value of time is not too much more heterogeneous than the value of schedule delay. That tolling can be good for the average consumer is surprising, as the common thought is that tolling is harmful for consumers. The distributional effects of tolling in our model can also be surprising. The lowest values of time and schedule delay are among those who gain most from a public pay-lane. With a private pay-lane, the low values of schedule delay lose relatively little; free-lane users with larger values, and even some pay-lane users, face larger price increases.

The next section describes the demand and generalised price equations. Section 3 analyses the no-toll (NT) and tolling equilibria with M discrete user groups. Then, the case of continuous heterogeneity is studied. For this, Section 4 describes the numerical model set up. Section 5 studies the NT equilibrium, Section 6 the first-best public (FB) toll, and Section 7 pay-lanes. Section 8 gives the sensitivity analysis. Section 9 concludes. Table 1 summarises the policies.

Table 1: Abbreviations of the policies

Abbreviation	Description
NT	No toll equilibrium
FB	Welfare maximising public first-best time-variant toll
PL	Welfare maximising public pay-lane (i.e. with a tolled pay-lane and untolled free-lane), with a time-variant toll
PPL	Profit maximising private pay-lane, with a time-variant toll

2. The demand and generalised price functions

One road connects the origin and destination. All users have the same preferred arrival time (t^*), which we normalise to arrival time (t) is zero. Drivers face a trade off between *travel time* costs, because of the queue at the bottleneck, and *scheduling* costs, due to arriving before (*schedule delay early*) or after (*schedule delay late*) t^*. We use square brackets to indicate that something is a function of the variables listed inside brackets. We use round brackets for arithmetic. Equation (1) gives the generalised total price for a type *i* driver. It is the sum of *travel time costs* ($CT_i[t]$), *schedule delay costs* ($CSD_i[t]$), toll ($\tau[t]$), and operating costs (υ). Operating costs are the same for all users. Travel time is the sum of free-flow travel time (T_f) and travel delay ($T_D[t]$). The latter follows from the queue length ($q[t]$) by $T_D[t]=q[t]/s$. Here s is the bottleneck capacity. The analytical models ignore free-flow travel time and operating costs, the numerical models include them. The toll consists of the time-*variant* toll ($\tau_t[t]$) and time-*invariant toll* ($\overline{\tau}$).

$$P_i[t] = CT_i[t] + CSD_i[t] + \upsilon + \tau[t] = \mu_i \cdot \beta_i (T_D[t] + T_f) + Max(-\beta_i t, \gamma_i t) + \upsilon + \tau[t] \qquad (1)$$

The value of schedule delay *late* (γ_i) is a linear in the value of schedule delay *early* (β_i), following $\gamma_i=\eta\beta_i$. The relative size of the value of time is μ_i. The term *type* indicates all users with the same values of time and schedule delay. Types can be continuously or discretely distributed. For a deterministic equilibrium, the inequality $\mu_i>1$ must hold for all users (Arnott et al., 1990).

Equation (2) gives the inverse demand function. $A+A_i$ is the constant in type *i*'s demand: A is common to all and A_i is type specific. The slope is determined by B and $b_i[\mu_i,\beta_i]$. The $b_i[\mu_i,\beta_i]$ is assumed to integrate to one, for algebraic ease.

$$D_i = A + A_i - \frac{B}{b_i[\mu_i, \beta_i]}\ n_i \qquad (2)$$

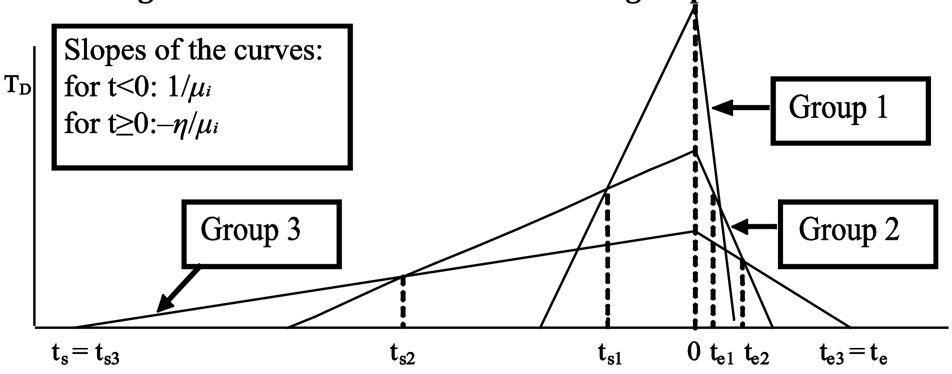
3. Equilibria with discrete heterogeneity

3.1 No-toll (NT) equilibrium with discrete heterogeneity

It is illustrating to start with the case of discrete heterogeneity, as this makes it easier to explain the effects of heterogeneity. This section assumes that each user group has a different ratio of values of time and schedule delay ($\mu_i \equiv \alpha_i/\beta_i$) and value of schedule delay (β_i). If two groups have the same μ_i, they share their NT arrival period—and although this would not change the model—it would complicate the mathematical notation. Subsequent sections relax this assumption. The groups are ordered on their μ_i. Arnott et al. (1988) find that Group 1, with the lowest μ_i, arrives closest to t*. Group M, with the highest $\mu_i \equiv \alpha_i/\beta_i$, arrive at the greatest distance from t^*.

A group's equilibrium price is constant during the period this group arrives. Outside this period the price is higher. A group's isocost curve gives the combinations of schedule delay and queuing time for which prices are constant over time. Obviously, a different isocost curve applies for a different cost level. The slope of an isocost curve is $1/\mu_i$ before t* and $-\eta/\mu_i$ after. Figure 1 gives the equilibrium isocost curves with three groups. At t_s and t_e, the first and last driver arrive; t_{si} and t_{ei} indicate when group i starts and ends to arrive. If i's equilibrium isocost curve is above the curves of the other groups and not below the horizontal axes, then at this moment only group i drivers arrive. Hence, Group 3 users arrive between t_s and t_{s2}, and between t_{e2} and t_e. Group 1 arrives between t_{s1} and t_{e1}. In equilibrium, the bottleneck operates at capacity during the entire peak. Thus, the peak duration (t_e–t_s) equals N/s. Here N is the number of users. At t_s and t_e Group M users arrive and face a zero queue length.

Figure 1: The *isocost* curves for three groups of drivers.



Using the above discussion, group M's price can be derived. Then, the price for group M–1 can be found, and so on for each group. Equations (3) and (4) give the generalised formulas for i's queuing and scheduling costs at t_{si} and t_{ei}. For arrivals closer to t* a group i user's scheduling cost is lower, whereas the queuing cost is higher. The scheduling costs of i increase with the number of users with a smaller ratio of value of time to value of schedule delay (μ_j), while queuing costs increase with the number of drivers with a larger μ_j. Both costs increase with i's value of schedule delay.

$$CSD_i[t_{si}] = CSD_i[t_{ei}] = \frac{\eta}{(1+\eta)} \frac{\beta_i}{s} \sum_{j=1}^{j=i} n_j \tag{3}$$

$$CT_i[t_{si}] = CT_i[t_{ei}] = \frac{\eta}{(1+\eta)} \frac{\beta_i}{s} \left(\mu_i \sum_{j=i+1}^{j=M} n_j / \mu_j \right) \tag{4}$$

$$P_i = CSD_i[t] + CT_i[t] = \frac{\eta}{(1+\eta)} \frac{\beta_i}{s} \left(\sum_{j=1}^{j=i} n_j + \mu_i \sum_{j=i+1}^{j=M} n_j / \mu_j \right) \tag{5}$$

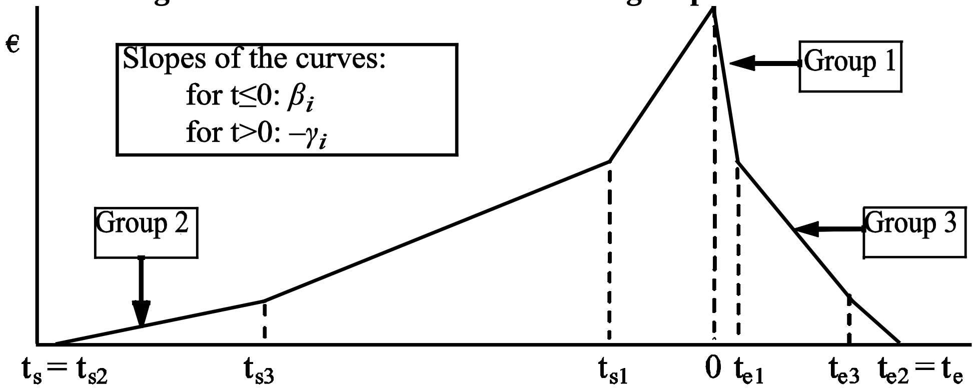
The NT price of i in (5) is the sum of i's queuing and scheduling costs. It increases with μ_i and β_i. Group M's price is $\eta \cdot N \cdot \beta_M / ((1+\eta)s)$. This is the same price as in the homogeneous user model with a value of schedule delay of β_M. Group M−1 gains from the heterogeneity. Group M's higher ratio $\mu_i \equiv \alpha_i / \beta_i$ induces them to build up the queue slower than Group M−1 drivers would. Hence, Group M drivers impose lower congestion costs than Group M−1 users. Group M−2 drivers enjoys an even larger price advantage, because Group M and M−1 drivers build up the queue slower than they.

3.2 Tolling equilibrium with discrete heterogeneity

With first-best tolling, the *time-variant* toll eliminates all queuing and the time-invariant toll is zero. As de Palma and Lindsey (2000) discuss, both a private and a public pay-lane's time-variant toll eliminates all queuing on the pay-lane, since queuing is always wasteful.

A queue-eliminating toll changes the arrival ordering: groups now arrive ordered on their value of schedule delay. The *highest-β-users* are group 1 and arrive closest to t^*. The *lowest-β-users* are group K and arrive furthest from t^* (Arnott et al, 1988). Figure 2 gives the FB toll for the example of Figure 1. Groups 2 and 3 have switched in the arrival order, since Group 3 has a higher β_i. Group 1 has the highest β_i and lowest μ_i. Hence, Group 1 arrives closest to t^* with and without tolling. The queue eliminating toll makes all users indifferent—as long as there is no queue—between all arriving moments in the period that their group arrives in the new ordering. Outside that period, their price is higher. The *time-variant* toll is by definition zero at t_s. Its slope is β_i before t^* and $-y_i$ after.

Figure 2: The toll schedule for three groups of drivers.



Using the above discussion, equations (6) and (7) for i's schedule delay cost and toll can be derived. For arrivals closer to t^*, scheduling costs are lower, whereas the toll is higher. The sum of scheduling cost and toll is constant during the period a group arrives. The resulting price formula is given in (8).

$$CSD_i[t_{si}] = CSD_i[t_{ei}] = \frac{\eta}{(1+\eta)} \frac{\beta_i}{s} \left(\sum_{k=i}^{k=K} n_k \right) \tag{6}$$

$$\tau[t_{si}] = \tau[t_{ei}] = \frac{\eta}{(1+\eta)} \frac{1}{s} \left(\sum_{k=1}^{k=i-1} \beta_k n_k \right) + \overline{\tau} \tag{7}$$

$$P_i = \tau[t] + CSD_i[t] = \frac{\eta}{(1+\eta)}\frac{1}{s}\left(\beta_i \sum\nolimits_{k=i+1}^{k=K} n_k + \sum\nolimits_{k=1}^{k=i-1} n_k \beta_k\right) + \overline{\tau} \tag{8}$$

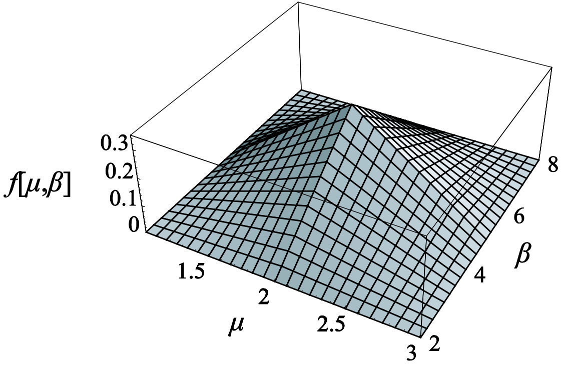
4. Numerical set-up

This section discusses the numerical set-up of the base case. The bottleneck's capacity is 3600 cars per hour. Operating costs are €7.30 per trip. Free-flow travel time is 30 minutes.

Figure 3 shows the NT density function of μ and β. It is based on two univariate symmetric triangular distributions: $f[\mu_i,\beta_i]=g[\beta_i]\cdot h[\mu_i]$. A symmetric triangular distribution is defined by its minimum and maximum. The minimum value per hour of schedule delay early ($\underline{\beta}$) is €2, the maximum ($\overline{\beta}$) is €8. The minimum ratio of value of time to value of schedule delay early ($\underline{\mu}$) is 1.01, the maximum ($\overline{\mu}$) is 3.01. The value of time is always larger than the value of schedule delay early. The weighted average value of time is €10.05. The relative size (η) of γ_i to β_i is 3.9. This is the same value as in Arnott et al. (1990).

The inverse demand function is created, following equation (2), so that three goals hold. First, the total number of NT users is 9000. Second, the weighted average of the NT equilibrium elasticity to total price is –0.4. Total price is the price including free-flow travel time and operating costs. Third, the discussed density function holds in the NT case.[2]

Figure 3: The multivariate distribution



5. Continuous heterogeneity no-toll (NT) equilibrium

Now we analyse the continuous heterogeneity no-toll case. This paper first studies the analytical models and then illustrates these by the numerical results as calculated in Mathematica 5.0. We were unable to find closed-form solutions for the tolling policies. Hence, in these cases we give the analytical results in so far we have them and then describe the numerical solution.

5.1 Analytical model for the no-toll (NT) equilibrium

The continuous heterogeneity price formula proofed a straightforward generalisation of the discrete version. The NT users arrive ordered on the ratio of value of time to value of schedule delay early (μ_i). The lowest-μ-users arrive closest to t^*. Schedule delays increase and queuing times decrease with μ_i, whereas they are independent of β_i. All NT users with the same μ_j behave in the same way. Thus, as (9) shows, we aggregate all users with the same μ_j to $m_{iNT}[\mu_i]$.

[2] To achieve these goals, we set $b_i[\mu_i,\beta_i]$ equal to the density function ($f[\mu_i,\beta_i]$), and A_i to the NT total price (P_{iNT}). The mean elasticity depends on B and the average total price. The average price is a function of the density function and total number of users. Hence, we can calibrate the aggregate elasticity with B. If $b_i[\mu_i,\beta_i]=f[\mu_i,\beta_i]$ and $A_i=P_{iNT}$, then equating inverse demands to prices, and rewriting, results in $N_{NT}=A/B$. Hence, we set the aggregate number of users with A. The A and B we use are 53.1841 and 0.0059094.

To get the continuous heterogeneity price equation (10) from the discrete version, we replace the summation signs by integrals and n_i by $m_{iNT}[\mu_i]$ (since there are now multiple types with the same μ_i). Equation (10) is rewritten to equation (11) by replacing $m_{iNT}[\mu_i]$ with $h[\mu_i]\cdot N_{NT}$ (i.e. μ's density function multiplied with the number of NT users). $H[\mu_i]$ is the cumulative distribution function of μ_i; N_{NT} is the number of NT users. We use a multivariate probability density function (PDF) of μ and β that is the product of two independent PDF's: $f[\mu_i,\beta_i]$ equals $h[\mu_i]g[\beta_i]$. The $g[\beta_i]$ and $h[\mu_i]$ are the univariate density functions of β and μ. Our model also works with a multivariate density function that is not based on two unconditional density functions. Then, however, our equations require more integrals and look more cluttered.

$$m_{iNT}[\mu_i] = \int_{\underline{\beta}}^{\bar{\beta}} n_{lNT}\, d\beta_l \tag{9}$$

$$P_{iNT} = CSD_i[t] + CT_i[t] = \frac{\eta}{1+\eta}\frac{\beta_i}{s}\left(\int_{\underline{u}}^{\mu_i} m_{jNT}[\mu_j]\, d\mu_j + \mu_i\int_{\mu_i}^{\bar{\mu}} \frac{m_{jNT}[\mu_j]}{\mu_j} d\mu_j\right) \tag{10}$$

$$P_{iNT} = \frac{\eta}{1+\eta}\frac{N_{NT}}{s}\beta_i\left(H[\mu_i] + \mu_i\int_{\mu_i}^{\bar{\mu}} \frac{h[\mu_j]}{\mu_j} d\mu_j\right) \tag{11}$$

Scheduling costs are given by the left term in the brackets in (11) multiplied by the term outside the brackets: by $\eta\cdot N_{NT}\cdot\beta_i\cdot H[\mu_i]/((1+\eta)s)$. The right term in the brackets multiplied gives scheduling costs. Prices increase non-linearly with the ratio $\mu_i \equiv \alpha_i/\beta_i$. Although schedule delays and queuing time are independent of β_i, prices increase linearly with β_i. A delay is more valuable with a higher β_i. The value of time equals $\mu_i\cdot\beta_i$. Thus, queuing time is more valuable with a higher β_i. As prices depend linearly on β_i, heterogeneity in β does not affect the average price.

5.2 Congestion externalities and heterogeneity

Following Lindsey (2004), equation (12) gives the congestion cost effect of a type *j* driver on a type *i*: it is the derivative of *i*'s price to the number of type *j* users. On all drivers with a larger μ_i than *j* (i.e. with $\mu_i \geq \mu_j$), type *j* causes a congestion effect of $(\beta_i/s)\cdot\eta/(\eta+1)$, which is independent of μ_i and μ_j. On all users with a smaller μ_i, *j* causes a smaller congestion effect. This smaller effect decreases with *j*'s ratio $\mu_i \equiv \alpha_i/\beta_i$ and increases with *i*'s μ_i.

The congestion effect of *j* also increases with *i*'s value of schedule delay, *j*'s value of schedule delay has no effect. A higher μ_j means that *j*'s isocost curve becomes flatter. This implies that *j* builds up the queue less quickly. This decreases the price for all users with a smaller μ_i than *j*, while prices for users with a larger μ_i than *j* are unaffected.

$$\partial P_{iNT} / \partial n_{jNT} = \begin{cases} \dfrac{\eta}{1+\eta}\dfrac{\beta_i}{s} & \mu_i \geq \mu_j \\ \dfrac{\eta}{1+\eta}\dfrac{\beta_i}{s}\dfrac{\mu_i}{\mu_j} & \mu_i < \mu_j \end{cases} \tag{12}$$

The marginal external cost of a type *i* driver in (13) is the integral of $n_j\cdot\partial p_j/\partial n_i$ over all *j*. Only the mean value of schedule delay affects externalities; the heterogeneity middles out,[3] as congestion effects depend linearly on β_i. The larger μ_i is, the smaller *i*'s externality: the *lowest-μ-users* cause the highest externality. The reason for this is that drivers that care relatively little

[3] Comparison of line two and three of equation (9) also shows this. In line two we find the term $\beta_l\cdot g[\beta_l]$, which contains the heterogeneity in β. In line three, the term is integrated out to $E[\beta]$; and *i*'s mec is a function of $E[\beta]$, μ_i, and the distribution of μ.

about travel delays—i.e. with a small μ_i –drive in the centre of the peak. For any distribution of μ with the same $E[\beta]$, the highest externality equals the externality in the homogeneous user model. With more ratio heterogeneity, there are higher *highest-μ-types*, who impose lower externalities than all other types, and/or more *high-μ*-users, who cause low externalities. The externalities of the lowest-μ-users cannot exceed the maximum externality. Hence, with more ratio heterogeneity, the mean externality is lower. This result was also found by Van den Berg and Verhoef (2010) and is hence not discussed further. Since, the mean externality decreases with ratio heterogeneity, the mean prices also decreases.

$$\begin{aligned}\text{mec}_i &= \int_{\underline{\beta}}^{\overline{\beta}}\int_{\underline{\mu}}^{\overline{\mu}} (\partial P_j / \partial n_{iNT}) n_{jNT}\ d\mu_j d\beta_l \\ &= \frac{\eta}{1+\eta}\frac{N_{NT}}{s}\left(\int_{\underline{\beta}}^{\overline{\beta}}\left(\int_{\mu_i}^{\overline{\mu}} h[\mu_j]\ d\mu_j + \frac{1}{\mu_i}\int_{\underline{\mu}}^{\mu_i} \mu_j h[\mu_j]\ d\mu_j\right)\beta_l g[\beta_l] d\beta_l\right) \\ &= \frac{\eta}{1+\eta}\frac{N_{NT}}{s}\left(1 - H[\mu_i] + \frac{1}{\mu_i}\int_{\underline{\mu}}^{\mu_i} \mu_j h[\mu_j]\ d\mu_j\right)E[\beta]\end{aligned} \tag{13}$$

5.3 Base case numerical model for the no-toll (NT) equilibrium

In the numerical NT base case, average queuing and scheduling costs are €3.97 and €4.97. The average total price (including free-flow travel time and operating costs) is €21.27. Total consumer surplus is €239,332. Average travel time is 54 minutes. Minimum travel time equals the free-flow travel time of 30 minutes, the maximum is 77 minutes.

Figure 4: Schedule delay costs (left) and queuing costs (right) per user

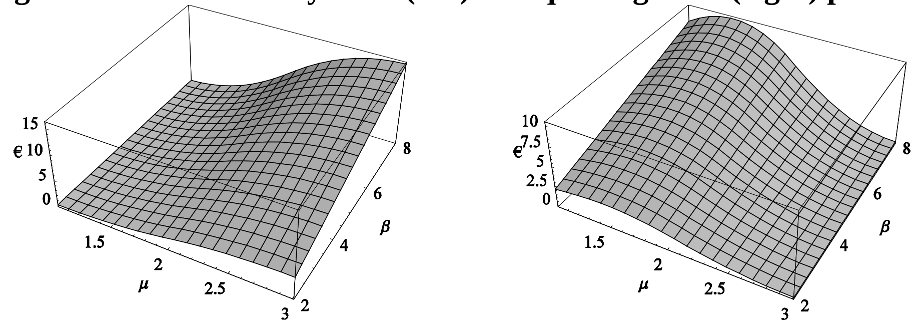


Figure 5: No-toll equilibrium price excluding free-flow travel time and operating costs

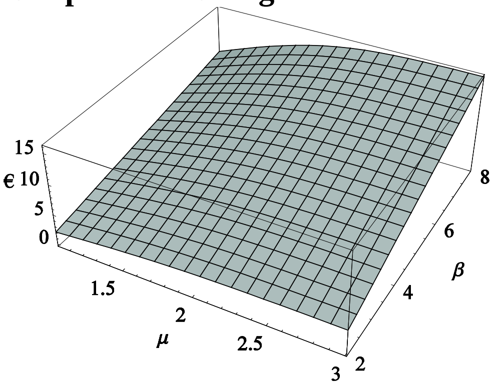


Figure 4 shows that scheduling and queuing costs increase linearly with β. Scheduling costs increase non-linearly with μ. Queuing costs decrease with μ for all, but the *lowest-μ-users.* The *lowest-μ-drivers* face the longest travel delays. Yet, with their low values of time, these delays are not that costly. Figure 5 depicts the NT prices. Prices increases linearly with the value of schedule delay; they increase non-linearly with μ_i.

Summarising, the average externality and travel price decrease with ratio heterogeneity. Since NT prices depend linearly on β_i, proportional heterogeneity does not affect average NT price.

6. Continuous heterogeneity and first-best public (FB) tolling

The first-best FB toll maximises welfare, which is the sum of total consumer surplus and toll revenues. The optimal time-*variant* toll eliminates all queuing. It slopes upward with t by β_i before t^*, and downward by $-\gamma_i$ after t^*. The optimal time-*invariant* toll is zero. FB tolling changes the arrival ordering. Before, users arrived ordered on μ_i. Now, they arrive ordered on β_i. All high-β-users now arrive close to t^*, whereas before some arrived far from t^* (Arnott et al, 1988). Thus, not only does tolling eliminate queuing, it also decreases scheduling costs.

6.1 Analytical model for the first-best public (FB) equilibrium

We derive the FB price equation (14) from the discrete version, by replacing the summation signs by integrals. $q_{jFB}[\beta_j]$ is the number of FB users with a value of schedule delay of β_j. It is the integral over μ_i of the number of users with β_j. In the FB equilibrium, all users with the same value of schedule behave in the same way regardless of their μ_i. The toll i faces increases with i's value of schedule delay: the higher β_i is, the closer i arrives to t^* where the toll is the highest. The effect on scheduling costs of β_i is ambiguous: schedule delays decrease with β_i, whereas the value of a delay increases with β_i. Nevertheless, FB prices always increase with β_i. Different from in the no-toll (NT) equilibrium, FB prices are independent of μ_i.

$$P_{iFB} = CSD_i[t] + \tau_t[t] = \frac{\eta}{(1+\eta)}\frac{1}{s}\left(\beta_i \int_{\beta_i}^{\overline{\beta}} q_{jFB}[\beta_j]\, d\beta_j \; + \int_{\underline{\beta}}^{\beta_i} \beta_j\, q_{jFB}[\beta_j]\, d\beta_j\right) \qquad (14)$$

6.2 Numerical base case model for the first-best public (FB) equilibrium

We could not find a closed-form solution of the FB equilibrium. Still, for a given starting distribution of the FB users, there is a solution for prices, total number of users, and implied distribution of the number of users. Unless, however, we exactly chose the equilibrium distribution as the starting distribution, the implied distribution and starting distribution are not equal. We could directly calculate NT prices, because we assume the NT distribution of users.[4]

Because queuing is eliminated, prices depend on the value of schedule delay only, and no longer on α. Hence, Figure 6 depicts the distribution of price components as a function of β alone. A given schedule delay is more costly with a higher β. But high-β-users arrive at more central moments, causing schedule delay costs to fall with β over a substantial range of Figure 6. These users have to pay high tolls to obtain these arrival times. The result is that the price always rises with β. This seems a rather general result. A user with a lower β could always travel at the same moment as a higher-β-driver, paying the same toll, but incurring lower scheduling cost. Because low-β-drivers choose to drive at a different moment, their price must be even lower.

[4] Still, this discussion does suggest a simple solution method. Starting with some distribution of FB users (we use the NT distribution), calculate prices and distribution of demands implied by these prices. This distribution is not equal to the starting distribution, as demands and prices are not in equilibrium. We approximate the distribution of demand between iterations by a cubic spline with 200 points. This spline is used as the next starting distribution. By continuing this procedure until convergence, using the new distribution of demand as the starting distribution, we find the FB equilibrium. The convergence criterion is a maximum absolute difference in the number of users of 10^{-20}% between iterations. It is possible to use the demand in the previous iteration as the starting distribution and not the spline. Then, however, the starting distribution's equation grows exponentially complex with the number of iterations. This makes the integrations slow and often causes them to break down.

Figure 6: Schedule delay costs and tolls per user

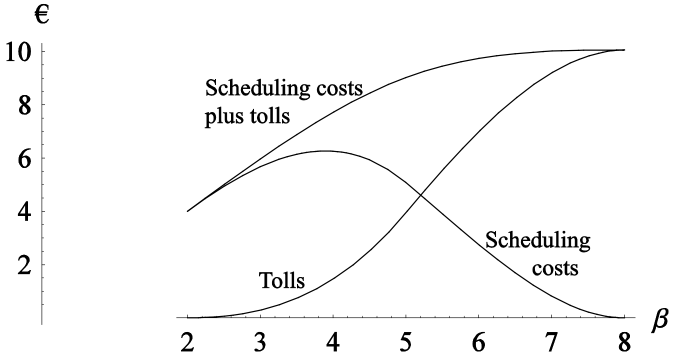
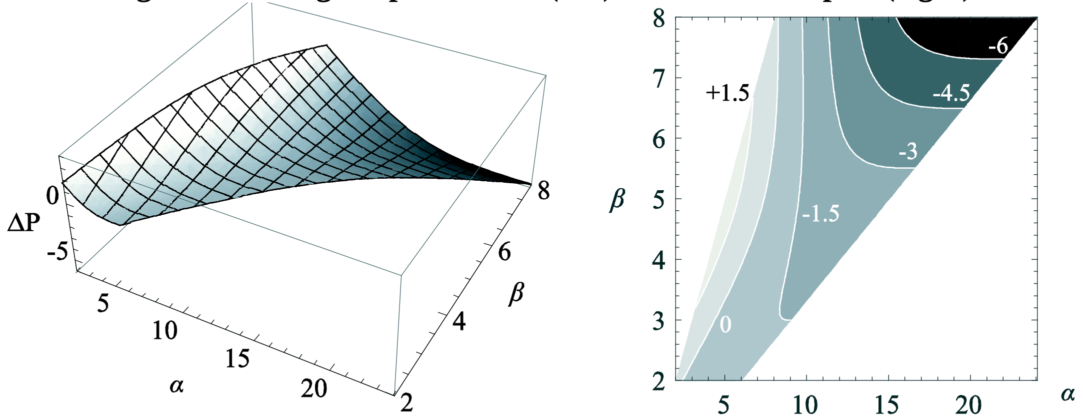


Figure 7: Change in price in 3D (left) and in contourplot (right)



Note: the contourlines are, from left to top right, change in price is 1.5, 0, –1.5, –3, –4.5 and –6 euros. The value of time (α_i) equals $\beta_i \cdot \mu_i$.

We use the difference $\Delta P_i = P_{iFB} - P_{iNT}$ as a measure of the loss from FB tolling for type *i* users. Figure 7 shows this difference, as a three-dimensional plot in the left panel and as contour plot in the right one. The figure depicts the differences over values of time ($\alpha_i \equiv \mu_i \cdot \beta_i$) and not over ratio of value of time to value of schedule delay (μ), as the results in α-space are easier to interpret.

It is striking that so many user types win. Still, this result is less surprising once it is realised that in the bottleneck model with homogeneous users, first-best tolling has no effect on prices. In our model, besides eliminating travel delays, the toll reduces aggregate schedule delays. This additional benefit causes the change to be more favourable.

As a result, the average generalised price decreases by 3.4% to €8.64 due to FB tolling. Aggregate consumer surplus increases by 1.4% to €242,571. Welfare (i.e. total consumer surplus plus toll revenues) raises by 17.7% to €281,798. The consistent implication—but not less surprising when comparing the results to those for the textbook static model of traffic congestion—is that aggregate use increases by 0.6% to 9057. Despite the longer duration of the peak, average schedule delay cost decreases by 13.1% to €4.32, which is of course due to the more efficient order of arrivals. In short, the average traveller benefits from first-best congestion pricing, even before toll revenues are recycled. Moreover, a majority of drivers, of 55%, benefits.

The contour plot of Figure 7 shows that, for a given β, the benefit increases with α. This seems to confirm common wisdom on the distributional impacts of road pricing. Yet, dependence of the gains on β causes an interesting twist. First, the greatest losses are not incurred by the drivers with the smallest α. Instead, *intermediate users* with an average β and lowest possible α for that β lose most. The reason is that their low $\mu_i \equiv \alpha_i / \beta_i$ allowed these intermediate users to arrive close to t^* in the no-toll equilibrium, benefiting from low schedule delay cost. With pricing, they lose this advantage, while the elimination of travel delays brings them little gain. As a result, the intermediate users incur the greatest losses from congestion pricing.

One might observe that this result, that it is not the lowest-α drivers who lose most, is somewhat exaggerated by the fact that β is restricted to be below α, so that the lowest α is not present for this intermediate β where losses seem largest. But even then, the contour plot shows that there are many instances where drivers with a higher α have greater losses than some drivers with a lower α *and* a lower β. At the same time, in the upper right corner there are many instances where drivers with a higher α gain less than some drivers with a lower α but a higher β. In other words, taking into account heterogeneity in the value of schedule delays makes the losses and gains of first-best congestion pricing to be not perfectly correlated with the value of time.

7. Continuous heterogeneity and the pay-lane

With a pay-lane, a share (ρ) of capacity is made a separate lane, and to use this lane one has to pay a toll. The remainder of road is the untolled free-lane. We also refer to the pay-lane as lane 1 and the free-lane as lane 2. Note that our pay-lane model could also be interpreted as the situation where there is a tolled road and a separate untolled road. The two interpretations are mathematically the same.

The pay-lane's *time-variant* toll eliminates all queuing on the pay-lane. The private operator adds a time-invariant toll that maximises total toll revenue. The public operator adds a negative time-invariant toll (i.e. subsidy) that maximises welfare. The subsidy attracts extra drivers to the public pay-lane. Hence, the pay-lane peak last longer than on the free-lane, and schedule delays are higher. The higher the subsidy is, the higher total schedule delays. The optimal subsidy is at the point where, for a marginal subsidy increase, the welfare gain from lessening queuing equals the loss from higher schedule delays. Because of the positive time-invariant PPL toll, the private pay-lane is used by relatively fewer drivers than the free-lane. Hence, the free-lane has the longer peak and schedule delays.

7.1 Analytical pay-lane model

The pay-lane's price equation (15) is basically the same as in the FB case. $q_{j1}[\beta_j]$ is the number of pay-lane users with a value of schedule delay of β_j. The free-lane price equation (16) is basically the same as the NT price formula. The $m_{j2}[\mu_j]$ is the number of free-lane users with μ_j. Finally, there is a critical $\alpha^*[\beta]$ curve that separates free-lane and pay-lane users. The users on the curve are indifferent about using the pay-lane or free-lane. All the users that, for their β_i, have a higher value of time than the curve drive on the pay-lane.

$$P_{i1} = CSD_i[t] + \tau_t[t] + \overline{\tau} = \frac{\eta}{(1+\eta)} \frac{1}{s\,\rho} \left(\int_{\underline{\beta}}^{\beta_i} \beta_j \; q_{j1}[\beta_j] \; d\beta_j + \beta_i \int_{\beta_i}^{\overline{\beta}} q_{j1}[\beta_j] \; d\beta_j \right) + \overline{\tau} \tag{15}$$

$$P_{i2} = CSD_i[t] + CT_i[t] = \frac{\eta}{1+\eta} \frac{\beta_i}{s(1-\rho)} \left(\int_{\underline{u}}^{\mu_i} m_{j2}[\mu_j] \; d\mu_j + \mu_i \int_{\mu_i}^{\overline{\mu}} \left(m_{i2}[\mu_j] / \mu_j \right) d\mu_j \right) \tag{16}$$

7.2 Base case numerical model for the public pay-lane (PL)

We were unable to find closed-form solutions for the pay-lane equilibria. The numerical solution is more difficult than for the FB toll, and is discussed in the Appendix. The optimal time-*invariant* toll is −€5.36. Hence, arrivals at the outside of the pay-lane peak receive a subsidy of

€5.36. The mean time-*variant* toll is €6.25. Consumer surplus increases 7.0% from the NT case to €256,142. Welfare increases with 8.7% to €260,237. The number of users increases, with 3.4%, to 9309.8. Of these users, 4619.1 use the pay-lane, and 4690.7 the free-lane. Thus, although the pay-lane has only a third of capacity, almost equal amounts of traffic use the pay-lane and free-lane. As a consequence, the pay-lane has a higher mean schedule delay.

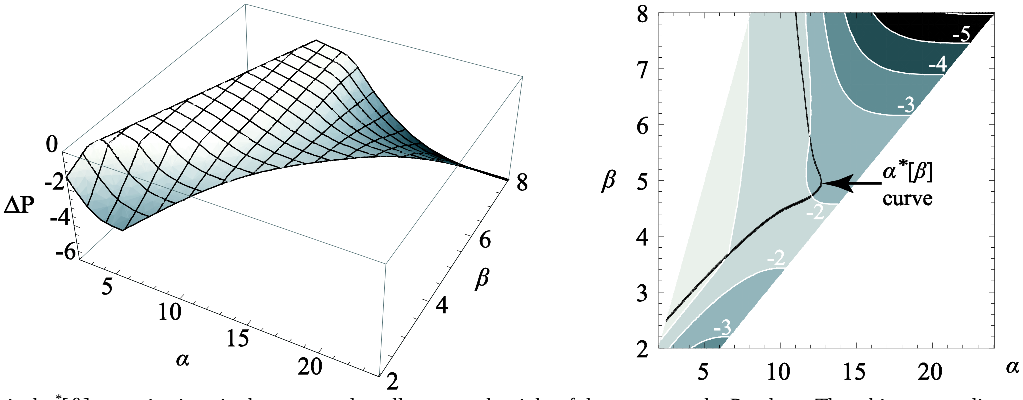
Figure 8 shows the difference between PL and NT prices. Again the results are shown as a function of the value of time (α) and not as a function of μ. In Figure 8's contourplot, the $\alpha^*[\beta]$ curve separates the pay-lane and free-lane users. All users to the right of the curve use the pay-lane. Surprisingly, not only the high values of time and schedule delay use the pay-lane, but also the low values of time and schedule delay. Moreover, all *lowest-β-users* use the pay-lane. This is counterintuitive: one would expect that only the highest values of time and schedule delay would use the tolled lane. The *low-β-users* arrive at the outside of the pay-lane peak. They face *negative* tolls and large schedule delays. With their low values of schedule delays, these large delays are not costly. Hence, they can enjoy the negative tolls and attain a large price decrease. Having *low-β-users* use the pay-lane improves the PL's welfare gain, as the pay-lane's higher schedule delays are imposed on the *low-β-users*.

All users are better off under the PL than in the NT case. Also in Braid (1996)'s homogeneous user model this occurs. Still, with heterogeneity in the value of schedule delay, users gain more, because of the self-ordering on β.

In the static PL model of Verhoef and Small (2004), drivers with the critical values of time lose most due to second-best pricing. In our model, there is to be no local maximum of ΔP_i near $\alpha^*[\beta]$, although it does mark a contour where ΔP_i becomes suddenly steeper for increasing value of time (α). An interesting question is why ΔP_i is falling with value of time for free-lane users; while, in the static model, ΔP_i is increasing with the value of time for them. The answer is that the subsidy on the pay-lane reduces the duration of the peak on the free-lane. This reduces travel delays; and the higher the value of time is, the larger the gain from this. Consistent with this, with a profit-maximising pay-lane (PPL), there is a positive time-invariant toll and local maxima of ΔP_i at $\alpha^*[\beta]$.

Although, users with a high μ_i and β_i gain most; the lowest values of time and schedule delay also gain. The users with an intermediate β_i and a low μ_i gain the least. In contrast, the conventional view is that a pay-lane is bad for the lowest values of time (before revenue recycling), since these users cannot afford the pay-lane and have to use the free-lane.

Figure 8: Differences between PL and NT prices in 3D (left) and contourplot (right)



Note: the critical $\alpha^*[\beta_i]$ curve is given in the contourplot, all users to the right of the curve use the Pay-lane. The white contourlines are, from left to top right, change in price is: –1, –2, –3, –4, and –5 euros. The value of time (α_i) equals $\beta_i{\cdot}\mu_i$.

7.3 Base case numerical model for the private pay-lane (PPL)

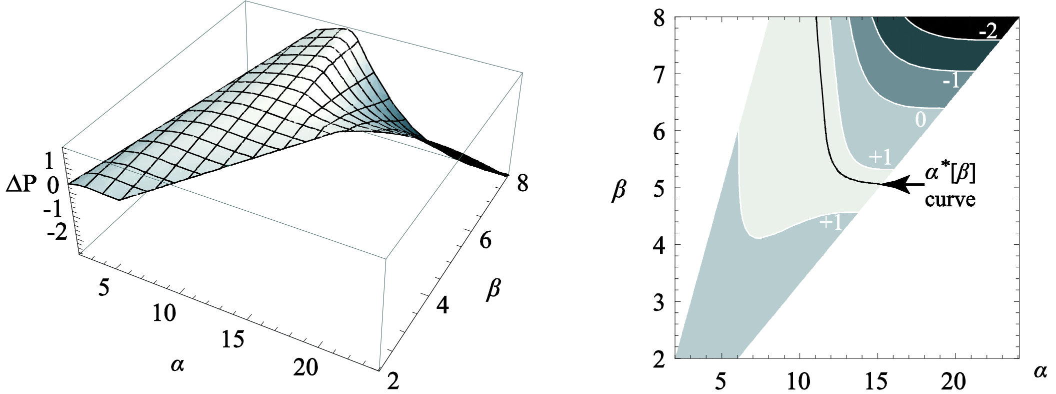
Figure 9 depicts the differences between PPL and NT prices. Due to the positive time-invariant toll, the peak lasts shorter on the pay-lane than on the free-lane. The PPL's $\alpha^*[\beta]$ curve, in Figure 9, has a more expected shape than the PL's: only the high values of time and schedule delay use the pay-lane. For the users that are indifferent between the pay-lane and free-lane prices increase most. Hence, there are pay-lane users that are hurt considerably by the PPL. For free-lane users that are further from the $\alpha^*[\beta_i]$ indifference curve the PPL is less detrimental. On the pay-lane, as a type is further away from the $\alpha^*[\beta_i]$ curve the PPL becomes rapidly more beneficial. Nevertheless, for most pay-lane users the PPL raises the price.

The PPL has some distributional surprises. It is not the *low values of time and schedule* delay that lose most, but the middle group that is (almost) indifferent between the pay-lane and free-lane. Further, this last group also contains pay-lane users. There are thus pay-lane users for who the PPL is more harmful than for the lowest-β-users. These results are similar to Verhoef and Small (2004)'s, where also the intermediate users lose most.

The time-invariant PPL toll is €3.87 and the mean time-variant toll is €4.34. The number of users decreases with 1.7% to 8846.8, because the private pay-lane (PPL) raises the price for most users. Nevertheless, the PPL increases welfare by 4.1%. Its relative efficiency is 0.23.

The PL might be seen as strange as it subsidises some pay-lane users: in practise, tolls are not negative. If the public pay-lane's time-invariant toll is constrained to zero—preventing negative tolls—welfare is much lower. Then, the distributional effects are akin to those of the PPL in Figure 9, although the critical $\alpha^*[\beta_i]$ curve would be to left and lower. The $\alpha^*[\beta_i]$ users lose most from this third-best policy, the low-β-users are hardly affected, whereas the high β and μ pay-lane users gain. The effects of this third-best pay-lane are similar to those in Verhoef and Small (2004)'s public pay-lane.

Figure 9: Difference between PPL and NT prices in 3D (left) and contourplot (right)



Note: the critical $\alpha^*[\beta_i]$ curve is also given in the contourplot, all users above the curve travel on the Pay-lane. The values for the difference in price ($P_{iPL}-P_{iNT}$) of the contourlines are, from top to bottom, change in price is: −2, − 1, 0, and 1 euros. The value of time (α_i) equals $\beta_i{\cdot}\mu_i$.

7.4 Concluding the pay-lane models

A public pay-lane lowers prices of all users. Prices with a private pay-lane are higher than no-toll prices for all free-lane users and for some pay-lane drivers. The public pay-lane is not only used by the high values of time and schedule delay, but also by the low values of schedule delay. The *low-β-users* arrive on the outside of the peak. They face large schedule delays and negative tolls (i.e. subsidies). The lowest values of time and schedule delay are among those who gain the most from the PL. Moreover, these users lose relatively little with the PPL.

8. Sensitivity analysis

This section focuses on the effect of different distribution of μ and β on the policies. We study five cases: *Homogeneity*, the B*ase case*, *Less ratio heterogeneity*, and a *Uniform distribution.* In all cases the mean value of schedule delay early is €5 and the mean relative size of the value of time is 2.01. The *Base case*'s spread of the triangular distribution of μ is 2, and of β it is €6. With the *Less ratio heterogeneity,* the spread of triangular distribution of μ is reduced to 1. With *Less proportional heterogeneity,* the spread of β equals €2. With the *Uniform distribution*, we test whether our results depend on our triangular distribution. The *Uniform distribution* has the same variance as the *Base case*. Finally, in the *Homogeneity* case, all users have the same parameters.[5]

8.1 Effect of heterogeneity on the no-toll (NT) case

Table 2 studies the effect of ratio and proportional heterogeneity on the NT equilibrium. The mean price is lower in the *Base case* than in the *Less ratio heterogeneity* case. This suggests that the mean price lowers with ratio heterogeneity. The mean congestion externality decreases with ratio heterogeneity, thereby lowering queuing costs. As section 5 predicted, proportional heterogeneity has no effect on average NT prices. The *Base case* and *Less proportional heterogeneity* case have the same average externality and price.

By design, all five cases have the same NT consumer surplus. The advantage of this is that the effect of tolling is more comparable over cases than when surplus would differ across cases.

Comparison of the *Base case* triangular and *Uniform distribution* indicates that the choice of NT user distribution form has no significant effect on aggregate results. The average price only differs a cent between the two cases, whereas mean queuing and scheduling costs do not even differ a cent. With both distributions, prices non-linearly increase with μ_i and linearly with β_i.

Table 2: Effect of heterogeneity in the no-toll (NT) equilibrium distribution of users

	Homogeneity	Base case	Less ratio heterogeneity	Less proportional heterogeneity	Uniform distribution[a]
Spread of the μ distribution	-	2	1	2	1.414
Spread of the β distribution	-	6	6	2	4.243
Mean schedule delay cost	€4.97	€4.97	€4.97	€4.97	€4.97
Mean travel delay cost	€4.97	€3.97	€4.44	€3.97	€3.97
Mean total price	**€22.27**	**€21.27**	**€21.74**	**€21.27**	**€21.26**
Mean marginal external cost	€9.95	€8.95	€9.42	€8.95	€8.94
Total NT consumer surplus	€239,332	€239,332	€239,332	€239,332	€239,332

a: This uniform distribution has the same variance and mean as the *Base case* triangular distribution of NT users.

8.2 Effect of heterogeneity on first-best public (FB) tolling

Table 3 tabulates the results of the sensitivity analysis for FB tolling. The differences between the results for the *Uniform* and *Base case* distribution are again small. Tolling is more beneficial for consumers in the *Base case* than with *Homogeneity*. The price decreases due to FB tolling in the *Base case*; whereas under *Homogeneity*, prices are unaffected. Still, the welfare gain is higher under homogeneity, since toll revenues are considerably higher.

With more ratio heterogeneity, externalities are lower, and thus there is less to gain from tolling. Therefore, consumer surplus, toll revenues, and welfare are higher with less ratio

[5] We do not present sensitivity analyses on price elasticities or pay-lane capacity share. Van den Berg and Verhoef (2010) discuss these analyses. Their results are in line with the homogeneous user literature. Also in this case, we found that these sensitivity analyses give unsurprising results.

heterogeneity. In the *Less ratio heterogeneity* case, the FB welfare gain is 19.5% of NT welfare, whereas in the *Base case* it is 17.7%, and with *Homogeneity* 18.7%.

With FB tolling, scheduling costs are lower in *Base case* than with the *Less proportional heterogeneity* case, since the gain of a more efficient arrival ordering increases with proportional heterogeneity. Accordingly, FB welfare gain and consumer surplus increase with this heterogeneity. If there is not too much more heterogeneity in the value of time than in the value of schedule delay, then FB tolling can be good for the average consumer. In the *Base case*, 55% of the NT users would gain; in the *Less ratio heterogeneity* case, 66% of NT users would gain, and with the *Less proportional heterogeneity* case 39%. The share of NT users that would gain seems to increase with proportional heterogeneity and decrease with ratio heterogeneity.

Vickrey (1973) considers the case where the values of time and schedule delay vary proportionally. He finds that first-best pricing is a strict Pareto improvement. Conversely, in our paper some users lose. First, because we have price-sensitive demand. The types of users that gain most raise their demand. This increases the price for the users that hardly gained in Vickrey (1973). Second, we extend Vickrey's heterogeneity with ratio heterogeneity. As noted, the number of users that gains from an FB toll decreases with the ratio heterogeneity.

Table 3: Effect of heterogeneity on the first-best public (FB) toll

	Homogeneity	Base case	Less ratio heterogeneity	Less proportional heterogeneity	Uniform distribution
Spread of the μ distribution	-	2	1	2	1.414
Spread of the β distribution	-	6	6	2	4.243
Mean schedule delay cost	€4.97	€4.32	€4.36	€4.70	€4.32
Mean FB toll	€4.97	€4.32	€4.36	€4.70	€4.32
Mean total price	€22.27	€21.01	€21.07	€21.75	€21.00
Number of FB users	9000	9054.6	9122.3	8922.8	9055.3
Toll revenues	€44,770	€39,137	€39,746	€41,970	€39,094
Total FB consumer surplus	€239,332	€242,571	€246,180	€235,352	€242,606
Welfare under the FB	€284,102	€281,708	€285,926	€277,323	€281,701
Percentage welfare gain from the NT case	**18.7%**	**17.7%**	**19.5%**	**15.9%**	**17.7%**
Percentage NT users that would have a lower price with FB tolling	Price is unchanged	55%	66%	39%	53%

8.3Effect of heterogeneity on the public (PL) pay-lane

Table 4 gives the sensitivity analysis for the PL (public pay-lane). In all five cases, all users gain due to the PL's time-invariant subsidy. Van den Berg and Verhoef (2010) have no heterogeneity in β. They find that many users lose from the PL. This indicates that, also in the current model, if the value of schedule delay is almost homogeneous, there will be users that are disadvantaged.

The *Less ratio heterogeneity* case has a higher welfare gain and relative efficiency than the *Base case*, indicating that these measures decrease with ratio heterogeneity (i.e. heterogeneity in the ratio of value of time to value of schedule delay). In the *Base case*, the time-invariant toll is higher than in the *Less ratio heterogeneity* case. Thus, in the *Base case*, the pay-lane has relatively fewer users compared with the free-lane. Hence, free-lane queuing is worse in the *Base case*. As Van den Berg and Verhoef (2010) discuss, it is apparently social optimal to allow more wasteful queuing with more heterogeneity, because of the lower congestion externalities, and this decreases the PL's relative efficiency.

The more proportional heterogeneity there is, the more efficient tolling makes the arrival ordering, thereby increases PL's welfare gain. Surprisingly, the PL's relative efficiency also increases with proportional heterogeneity. One would expect that this relative efficiency would decrease with this heterogeneity, as the pay-lane imposes the more efficient arrival order on only a part of the users. The public pay-lane has a larger maximum and mean schedule delay than the free-lane. These large delays are faced by the *lowest-β-users*. The *high-β-users* have the lowest delays. With more heterogeneity, the mean value of schedule delay of the pay-lane's *high-β-users* is higher, making the schedule delay savings the PL offers them more valuable. The mean value of schedule delay of the pay-lane's *low-β-users* decreases with heterogeneity in *β,* making the extra schedule delays the PL imposes less costly. This explains why the PL relative efficiency increases with the heterogeneity in *β*.

The differences between the *Base case* and *uniform* distribution are larger than in the NT and FB cases. Nevertheless, the differences are still minor. The time-invariant toll is a cent lower with the *uniform* distribution; the PL's percentage welfare gain is 0.22 percentage point lower.

Table 4: Effect of heterogeneity with the public (PL) pay-lane

	Homogeneity	Base case	Less ratio heterogeneity	Less proportional heterogeneity	Uniform distribution[a]
Spread of the μ distribution	-	2	1	2	1.414
Spread of the β distribution	-	6	6	2	4.243
Time-invariant part of the toll	−€6.37	−€5.36	−€5.40	−€5.38	−€5.35
Mean time-variant part of the toll	€7.27	€6.25	€6.29	€6.72	€6.31
Number of users	9302.8	9309.8	9356.4	9205.5	9293.9
Toll revenues	€3917	€4095	€4153	€5680	€4439
Total PL consumer surplus	€255,706	€256,142	€258,692	€250,408	€255,260
Welfare under the PL	€259,623	€260,237	€262,845	€256,089	€259,699
Relative efficiency	**0.453**	**0.493**	**0.505**	**0.441**	**0.481**
Percentage welfare gain from the NT case	**8.48%**	**8.73%**	**9.82%**	**7.00%**	**8.51%**
Percentage NT users that would have a lower price with the PL	100%	100%	100%	100%	100%

a: This uniform distribution has the same variance and mean as the *Base case* triangular distribution of NT users.

8.4 Effect of heterogeneity on the private (PPL) pay-lane

Table 5 shows the final sensitivity analysis for the PPL. It indicates that the PPL welfare gain and relative efficiency decrease with heterogeneity in μ. Van den Berg and Verhoef (2010) suggest that, with a more heterogeneous value of time, it is profit maximising to allow more wasteful queuing and schedule delay on the free-lane, lowering the relative efficiency and welfare gain.

The PPL's welfare gain and relative efficiency are lower in the *Less proportional heterogeneity* case than in the *Base case*, indicating that these measures increase with proportional heterogeneity. With a more proportional heterogeneous, the average value of schedule delay on the pay-lane is higher, and on the free-lane it is lower. This makes the schedule delay savings on the pay-lane more valuable and the free-lane's extra schedule delays less costly. This reasoning follows the same train of thought as Verhoef and Small (2004) use to explain why, with static flow congestion, the pay-lane's relative efficiency increases with heterogeneity in the value of time.

With homogeneity, all users are worse off with a PPL than in the NT case. Conversely, in all heterogeneity cases there are some users who gain. Just as with first-best tolling; the share of NT users that would face lower prices with the PPL decreases with heterogeneity in *β*. In opposition to FB tolling, the PLL's share that would gain slightly decreases with ratio $\mu \equiv \alpha/\beta$ heterogeneity. This seems to be because with *Less ratio heterogeneity*, the free-lane congestion externalities are worse. This makes the free-lane a less attractive good, enabling the PPL operator to ask a higher profit maximising time-invariant toll. Hence, with welfare maximization, the share of users that gains from tolling decreases with ratio heterogeneity; with profit maximization it increases.

The differences between the *Base case* triangular and uniform distribution are larger than for the NT or FB equilibrium. Again, however, the differences are not over all result changing.

Table 5: Effect of heterogeneity with the private (PPL) pay-lane

	Homogeneity	Base case	Less ratio heterogeneity	Less proportional heterogeneity	Uniform distribution
Spread of the μ distribution	-	2	1	2	1.414
Spread of the β distribution	-	6	6	2	4.243
Time-invariant part of the toll	€3.03	€3.87	€3.97	€3.14	€3.96
Mean time-variant part of the toll	€4.64	€4.34	€4.48	€3.97	€4.37
Number of users	8855.9	8846.8	8855.4	8836.1	8837.4
Toll revenues	€16,201	€17,964	€18,619	€16,288	€18,133
Total PPL consumer surplus	€231,729	€231,279	€231,724	€230,701	€230,790
Welfare under the PPL	€247,930	€249,243	€250,343	€246,989	€248,923
Relative efficiency	**0.192**	**0.234**	**0.236**	**0.202**	**0.226**
Percentage welfare gain from the NT case	**3.59%**	**4.14%**	**4.60%**	**3.19%**	**4.00%**
Percentage NT users that would have a lower price with the PPL	none	6.0%	5.7%	0.6%	5.0%

8.5 Concluding the sensitivity analysis

Arguably the most striking result is that the results seem robust to changes in the distributions of μ and β. Particularly striking is how close the results are for the *Uniform* and the *Base case* triangular distribution. Reducing the heterogeneity in μ makes externalities bigger, raising the gain from pricing. Reducing the proportional heterogeneity lowers the gains from the reordering of arrival times, lowering the gain from pricing. A larger share of NT drivers benefits from first-best pricing with less ratio heterogeneity; a smaller share gains with less proportional heterogeneity. The distributional effects of first-best and second-best tolling are also rather robust. For all cases, the patterns resemble those discussed for the base case. The main difference is that with more ratio or proportional heterogeneity, losses and gains are larger.

The relative efficiencies of the pay-lanes are higher in the *Base case* than with *Homogeneity*. If the base case had a more ratio heterogeneity or a less proportional heterogeneity, then the relative efficiencies could be higher with homogeneity than with heterogeneity. The FB welfare gain is lower with the *Base case* than with *Homogeneity*. Nevertheless, if the calibration of the model would be different (e.g. more proportional heterogeneity), this could change. Whether a policy is more or less beneficial with heterogeneity than with homogeneity depends on the empirical question what distribution heterogeneity has: how much more heterogeneous is the value of time than the value of schedule delay?

9. Conclusion

This paper analysed how, in the bottleneck model with price sensitive demands, ratio and proportional heterogeneity affect congestion tolling. Proportional heterogeneity scales the value of time as the values of schedule delay proportionally; it can be viewed as giving the effect of income. Ratio heterogeneity is in the ratio of value of time to value of schedule delay.

More ratio heterogeneity lowers the average no-toll equilibrium congestion externality. Hence, there is less to gain from tolling, and the welfare gain of tolling is lower. More proportional heterogeneity, increases the gain of tolling; since the gain from the more efficient arrival ordering tolling causes is higher, meaning that tolling lowers mean schedule delays costs more.

With a pay-lane, only a part of the road is tolled, the remainder (the free-lane) remains toll free. The public pay-lane is used by highest values of time and schedule delay *and* by the lowest values of schedule delay. These *low-β-users* arrive on the outside of the pay-lane peak. They enjoy the negative tolls that then apply, whereas the large schedule delays then are not that costly. The private pay-lane has a positive time-invariant toll that maximises revenue. It is only used by the highest values of time and schedule delay. The relative efficiency of a pay-lane increases with proportional heterogeneity, it decreases with ratio heterogeneity.

If there is not too much more heterogeneity in the value of time than in the value of schedule delay, then the mean generalised price can be lower with first-best tolling than without tolling. The share of no-toll equilibrium users that would face lower prices with first-best tolling increases with proportional heterogeneity, while it decreases with heterogeneity in the ratio (μ) of value of time and value schedule delay. For a private pay-lane, the share that would gain reduces with both proportional and between heterogeneity. The first-best toll and private pay-lane are never a Pareto improvement. For most distributions of the heterogeneity, all users gain from the public pay-lane; only if the value of schedule delay is (almost) homogeneous, some users lose.

The distributional effects of tolling can be surprising. The gains or losses from first-best tolling are not strictly monotonically increasing in the value of time, because these also depend on the value of schedule delay. Conversely, with only heterogeneity in the value of time, the gains and losses from first-best pricing are monotonic. First-best tolling is most harmful for users with an average value of schedule delay and a slightly larger value of time, while it raises the price less for the lowest values of time and schedule delay. The distributional effects with a public pay-lane are similar to those of the first-best optimum. A difference is that the lowest values of time and schedule delay are among those who gain most. Further, low-value-of-time users lose relatively little with a private pay-lane, while higher value of time and schedule delay free-lane users and even some pay-lane users lose more.

We focused on which values of time and schedule delay gain or lose from tolling. Yet, the question how real world socio-economic groups are affected remains unanswered. Our results suggest that distributional impacts not only depend on income; but also on, for instance, the type of (un)employment, family structure, and trip purpose. If two drivers have the same income, but one is an assembly worker, and the second is unemployed or alternatively an office worker. Then, the commuting assembly worker probably has a higher value of schedule delays, as her job has strict starting hours. Clearly, pricing then has different effects on the two. Conversely, if the assembly worker travels to visit friends and the unemployed to visit a doctor, then the unemployed might have higher values of time and schedule delays. Even if two drivers are observably the same, they might still have very different values. For example, with two commuting office workers, who work at the same firm and have the same income. One worker might have a stricter boss or more scheduling constraints from the home life (e.g. small children), implying an higher value of schedule delay. While, the second might have a better car audio system, making driving more pleasant and thus travel time less costly. To map the results of our study on a real world population, one seems to need information on differences between groups of drivers and unobserved heterogeneity inside those groups.

Acknowledgements

This research was financially supported by Transumo. Transumo (TRANsition SUstainable MObility) is a Dutch platform for companies, governments and knowledge institutes that cooperate in the development of knowledge with regard to sustainable mobility. We are grateful for the comments of Eva Gutiérrez-i-Puigarnau and Ferdinand Paraguas.

Literature

Arnott, R., de Palma, A. and 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., 1990. Economics of a bottleneck. *Journal of urban economics* 27(1), 111–130.

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.

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

Cain, A., Jones, P.M., 2008. Does urban road pricing cause hardship to low-income car drivers?: an affordability-based approach. Transportation Research Record 2067, 47–55.

de Palma, A. and Lindsey, R., 2000. Private toll roads: Competition under various ownership regimes. *The Annals of Regional Science* 34(1), 13–35.

Foster, C.D., 1974. The regressiveness of road pricing. *International Journal of Transport Economics* 1(2), 186–188.

Foster, C.D., 1975. A note on the distributional effects of road pricing**.** *Journal of Transport Economics Policy* 9, 186–188.

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

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.

Mayeres, I., Proost, S., 2001. Marginal tax reform, externalities and income distribution. *Journal of Public Economics* 79(2), 343–363.

Pigou, A.C. (1920). The Economics of Welfare. Mac-millan, London.

Richardson, H.W., 1974. A note on the distributional effects of road pricing. *Journal of Transport Economics Policy* 8(1), 82–85.

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

Small, K.A., Winston, C., Yan, J., 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica* 73(4), 1367–1382.

van den Berg, V. and Verhoef, E.T., 2010. Congestion tolling in the bottleneck model with heterogeneous values of time. *Transportation Research Part B,* in press.

Verhoef, E.T. and Small, K.A. (2004). Product differentiation on roads. *Journal of Transport Economics and Policy* 38(1), 127–156.

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

Appendix: Numerical solution method for a pay-lane equilibrium

This appendix discusses the numerical solution procedure for the pay-lanes. Three iterative procedures are used: the second is around the first and the third procedure around the second. We begin with some starting time-invariant toll, critical $\alpha^*[\beta]$ curve and distribution of users. The first iterative procedure searches for the distributions for which the prices and inverse demands are equal. The starting distribution is used to calculate prices and the demands implied by these prices. The next iteration uses this iteration's demand, approximated by a cubic spline, as the starting distribution. The convergence criterion is a maximum absolute percentage change in the number of pay-lane and free-lane drivers of 10^{-12}% from one iteration to the next.

For now pay-lane and free-lane prices for $\alpha^{*}[\beta]$ curve users are not equal. Hence, we seek a new curve for which the price differences are smaller. After this, the first iterative procedure is repeated again. The convergence criteria for the second procedure is a maximum absolute percentage difference between pay-lane and free-lane prices for $\alpha^{*}[\beta]$ curve users of 0.075%.

The third procedure searches for the optimal time-invariant toll. For the public pay-lane, we start with calculating welfare for three time-invariant tolls (−€1.50, −€2.00 , and −€2.50). Next, we fit a second-order polynomial to the tolls and their welfare. This polynomial is maximised to find the predicted toll ($\overline{\tau}^{p}$). Next iteration's toll is $\overline{\tau}^{t+1} = (1-ss)\cdot\overline{\tau}^{t} + ss\cdot\overline{\tau}^{p}$, where *ss* is the step size. Then, welfare resulting from this new toll is calculated. We fit a polynomial on the last two tolls (i.e. −€2.00 and −€2.50) and this iteration's toll and the welfares corresponding to these tolls, to find the next iteration's toll. If there is no increase in welfare from one iteration to the next, the step size is halved. Both consumer surplus and toll revenue seem globally concave in the time-invariant toll. Therefore, this procedure should converge to the optimal toll.

We repeat this third procedure until the absolute in welfare between iterations is below 0.25. After convergence, we use the time-invariant toll with the highest welfare in the series of calculations as the optimal toll. The procedure for the PPL is basically the same as for the PL, only now toll revenues are maximised instead of welfare.