Congestion tolling in the bottleneck model with heterogeneous values of time[1]

Vincent A. C. van den Berg[a*], Erik T. Verhoef[a b #]

a: Department of Spatial Economics, VU University, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands
b: Affiliated to the Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam, The Nederlands.
#: email: everhoef@feweb.vu.nl
*: Corresponding author: email: vberg@feweb.vu.nl, tel: +31 20 598 6049,

Abstract

When analysing the effects of transport policies it is important to adequately control for heterogeneity: previous studies note that ignoring heterogeneity biases the estimated welfare effects of tolling. This paper examines the effects of tolling, in a bottleneck model, with a continuously distributed value of time. With homogeneous users, first-best public tolling has no effect on prices. With heterogeneity it does: low values of time lose, and high values of time gain. The average congestion externality decreases with the heterogeneity in the value of time. Consequently, the welfare gain of first-best tolling also decreases. The more heterogeneous the value of time is, the lower the relative efficiency of a public pay-lane. This finding contrasts with the previous literature. Earlier studies, using static flow congestion, conclude that the relative efficiency increases with this type of heterogeneity. With more heterogeneity in the value of time, the relative efficiency of a private pay-lane is also lower, while that of a public time-invariant toll is higher. Our results suggest that the welfare gains of different tolling schemes are affected differently by heterogeneity. Further, the impact of heterogeneity on the effects of a policy also depends on the type of congestion considered.

Keywords: congestion pricing, bottleneck model, heterogeneity, value of time, second-best pricing
JEL codes: D62, H23, L11, R41, R48

1. Introduction

Traffic congestion is a major problem in most modern societies. A way to alleviate this problem is congestion pricing. When studying tolling, it is important to adequately incorporate heterogeneity: Arnott et al. (1988) find that ignoring heterogeneity may bias the calculated welfare effects of first-best tolling. This bias can be negative, zero, or positive, depending on the type and extent of heterogeneity. Lindsey (2004) notes that the congestion cost a user imposes on another user decreases with her value of time and increases with the other's value of time.

Dynamic congestion models usually assume that a driver faces a schedule delay if she does not arrive at her preferred arrival time (t^*). In the conventional linear specification, the cost per hour of earlier arrival than t^* (*schedule delay early*) is β, the cost per hour of later arrival (*schedule delay late*) is γ. The value of time is α. A number of studies have looked at the effect of heterogeneity in these parameters. Cohen (1987) examines fine tolling with two user group that differ in their t^*, α, β, and γ. Huang (2000) investigates pricing policies in a transit and road network with two groups of users. De Palma and Lindsey (2002) find that, in the bottleneck model with homogeneity, the morning and evening peak are mirror images of each

[1] This is a post-print of the article published in 2011 in Transportation Research Part B, 45(1), 60–78: for the published version see http://dx.doi.org/10.1016/j.trb.2010.04.003, the journal website is http://www.journals.elsevier.com/transportation-research-part-b-methodological/

other. But this symmetry breaks down with heterogeneity. Arnott et al. (1994) use two types of users and perfectly inelastic demand. First-best tolling raises the price for the low values of time. With a more heterogeneous value of time, this price increase is larger. If there is enough heterogeneity in α, then even if the tolls are redistributed equally, the low value of time can still be worse off.

In the empirical literature, estimation methods, such as probit and mixed logit, that control for unobserved heterogeneity are becoming increasingly popular. This suggests that in reality heterogeneity of preferences is common and important. Still, only a limited number of theoretical transport studies control for heterogeneity, and most that do control use two groups of users. As Verhoef and Small (2004) note, apart from the fact that two group heterogeneity seems a crude approximation of real heterogeneity, it may also cause analytical problems due to the possible existence of multiple pooled and separating equilibria.

This paper examines the welfare effects of first-best (fine) and second-best congestion pricing in a bottleneck model with continuous heterogeneity in the value of time and price sensitive demand. In the first second-best case we study, part of the road capacity is used as a pay-lane with the goal of maximising welfare (public pay-lane) or profits (private pay-lane).[2] In the second case, only one public time-invariant toll can be set for the entire peak. This time-invariant toll can only lower the externality by lowering total demand, while time-variant first-best tolling eliminates all queuing. To the best of our knowledge, the effects of these second-best policies have not been analysed before in the heterogeneous user bottleneck model with price sensitive demand.

Evans (1992) analyses a flow congestion model with drivers that differ in their value of time and valuation of the trip. If the values of time and of the trip are correlated (i.e. high values of time value the trip more), then congestion pricing improves welfare more with heterogeneity than with homogeneity. Small and Yan (2001) and Verhoef and Small (2004) look at private and public pay-lanes using flow congestion. They conclude that the relative efficiency of a public pay-lane increases with heterogeneity in α. Relative efficiency is the proportion of the first-best welfare gain, relative to the no-toll case, that a policy achieves. The pay-lane attracts the high-α-users, who gain more from travel time savings. The free-lane's higher travel time is less harmful for the low-α-users.

In our model, the relative efficiency of a public pay-lane decreases with the heterogeneity in the value of time (α). This is opposite to the conclusions of the previous studies. Further, with greater heterogeneity in α, the mean congestion externality is lower and there is less to gain from tolling, implying that the first-best toll's welfare gain is lower.

We assume that t^* is the same for all persons. Further, all drivers have the same values of schedule delay early (β) and late (γ). These simplifying assumptions influence our results. For example, Arnott et al. (1988) show that with heterogeneity in t^*, total scheduling costs are lower, while travel delay costs are unaffected. In Vickrey (1973), the values of time and schedule delay vary proportionally (i.e. all drivers have the same ratios of β/α and γ/α). All drivers are better off with first-best (FB) tolling than without tolling, with the exception of the very lowest values users who are unaffected. Xiao et al. (2009) use the same assumption regarding α, β, and γ as Vickrey (1973). Their one-step (coarse) toll has higher welfare gain with this heterogeneity than with homogeneity and is a Pareto-improvement.

Arnott et al. (1988) also analyse heterogeneity in the value of time and schedule delay, with a fixed β/γ ratio. In the no-toll (NT) case, users arrive ordered on α/β, with the users with the highest value arriving furthest from t^*. First-best (FB) tolling eliminates the queue. Now users arrive ordered on β, with the highest-β-users arriving closest to t^*. Thus, with a

[2] Our pay-lane case can also be interpreted as the situation where there are two roads connecting the origin and destination and only one of them can be tolled (e.g. a tolled highway and an untolled secondary road). The two interpretations are mathematically the same. Indeed, the pay-lane studies we discuss actually interpret the model as a two link case.

heterogeneous value of schedule delay, FB tolling makes the arrival ordering more efficient; decreasing average scheduling costs and increasing the FB welfare gain. In Van den Berg and Verhoef (2010), we analyse heterogeneity in values of time and schedule delay. This model is less tractable than the present one, and has fewer closed-form solutions. Still, numerical analysis indicates that a more heterogeneous value of time lowers the FB welfare gain and relative efficiencies of pay-lanes, even if the value of schedule delay is also heterogeneous.

The next section describes the generalised price and demand functions. Section 3 considers the no-toll (NT) equilibrium with a discrete value of time distribution. We then move on to continuous heterogeneity. Section 4 discusses the set up of the numerical models. Section 5 analyses the NT equilibrium. Section 6 investigates public first-best (FB) tolling. Section 7 examines a private monopolist controlling the road and setting a time-variant (PM) toll. Section 8 studies the public (PL) and private (PPL) pay-lane, and public time-invariant (TI) tolling. Table 1 summarises the policies. The sensitivity analysis is presented in Section 9 and Section 10 concludes. Appendix A defines some of the used symbols.

Table 1: Abbreviations of the analysed policies

Abbreviation	Description
NT	No toll
FB	Public first-best welfare-maximising time-variant toll
PM	Private profit-maximising time-variant toll, which is set by a monopolist
PL	Public welfare-maximising pay-lane, with a time-variant toll
PPL	Private profit-maximising pay-lane, with a time-variant toll
TI	Public time-invariant (uniform) welfare-maximising toll, which is set on the whole road

2. The generalised price and demand surface

Consider a single road with a bottleneck, connecting a single origin and destination. The generalised price ($P_i[t]$) in (1), for a user arriving at the destination at t, is the sum of the travel time costs ($CT_i[t]$), schedule delay costs ($CSD_i[t]$), toll $\tau[t]$, and car operating costs (υ). Operating costs are the same for each person. We use square brackets to indicate arguments of a function; round brackets are used for arithmetic. Travel time is the sum of the travel delay and free-flow travel time (T_f). Travel delay ($T_D[t]$) equals the queue length faced when arriving at t ($q[t]$) divided by capacity s. The coefficient α_i denotes the value of time of type i users, and is the only source of heterogeneity. We refer to a user with a value of time of α_i as a "type" i user, where types can be continuously or discretely distributed. To shorten the algebra, we define the parameters $\eta=\gamma/\beta$ and $\delta=\beta\cdot\gamma/(\beta+\gamma)$. We normalise t^* (the preferred arrival time) to zero for notational ease.

$$P_i[t] = CT_i[t] + CSD_i[t] + \tau[t] + \upsilon = \alpha_i \cdot (T_D[t] + T_f) + Max(-\beta \cdot t, \gamma \cdot t) + \tau[t] + \upsilon \qquad (1)$$

We assume that costs are linear in schedule and travel delay. For an equilibrium without mass departures, the inequality $\alpha_i>\beta$ must hold. If it does not hold for some early arriving users, then for those users the departure rate is infinite and a mass departure results (Arnott et al., 1990). Note that $\alpha_i<\beta$ can be viewed as the situation where users arriving before t^* prefer prolonging their trip over getting out of the car and entering work.

If demand is price sensitive for all values of time, an inverse demand surface (D_i) can be defined. We assume that it has the form of equation (2). The slope of i's function is $-B/b_i[\alpha_i]$. The $A+A_i$ is the constant of the demand function. A_i is a value of time specific addition to A. The integral of $b_i[\alpha_i]$ over all values of time is one. This is for algebraic ease, and is not a necessary assumption, because B can be freely chosen.

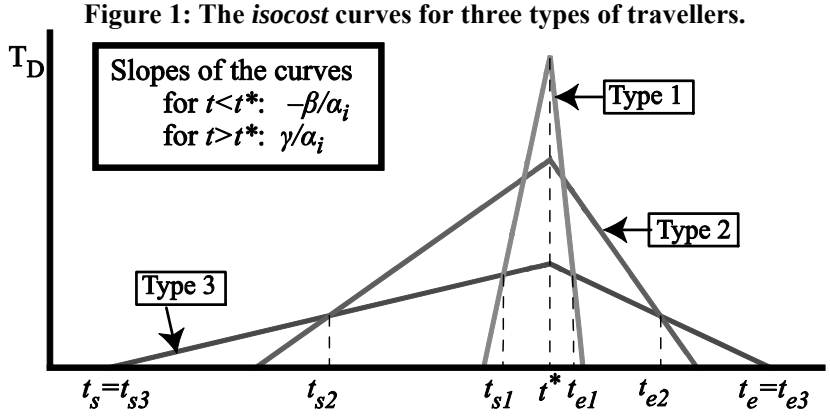
$$D_i = A + A_i - \frac{B}{b_i[\alpha_i]}\, n_i \qquad (2)$$

3. No-toll equilibrium with discrete heterogeneity

It is instructive to start our exposition with a discrete heterogeneity model, as the effects of heterogeneity are easier to understand in this case. This section discusses the NT equilibrium with M user groups with group specific values of α_i. Without loss of generality the analytical models ignore the free-flow travel time and operating costs. In the numerical models these two items are added again. The groups are ordered by decreasing values of β/α_i. The group with the lowest α_i is group 1, and the group with the highest α_i is group M. Arnott et al. (1988) find that, in the NT equilibrium, group 1 arrives closest to t^*, and group M the furthest from t^*. A group's equilibrium price is the same at all moments when drivers from this group arrive and not lower at any other moment. t_s is the point in time when the first driver arrives, at t_e the last driver arrives.

Following Arnott et al. (1988) we set up *isocost* curves; which give the combinations of travel delay (along the vertical axis) and schedule delay (along the horizontal axis) that result in a constant price over time. Figure 1 draws the equilibrium *isocost* curves with three user types. Clearly, different *isocost* curves apply for different cost levels. If a type's equilibrium *isocost* curve is above the curves of the other types and not below the horizontal-axis, then at this moment only this type arrives. If another type of user arrived at this time, her price would be higher than during her own arrival period. Let t_{si} and t_{ei} indicate when group i starts (for early arrivals) and finishes (for late arrivals) arriving. Then, type 3 users arrive between t_s and t_{s2}, and between t_{e2} and t_e. Type 2 users arrive between t_{s2} and t_{s1}, and between t_{e1} and t_{e2}. The upper envelope of the *isocost* curves gives the equilibrium delays.

Figure 1: The *isocost* curves for three types of travellers.



Equation (3) gives the equilibrium generalised price, and is derived in Appendix B. Here, n_{jNT} is the number of NT drivers with a value of time of α_j. The N_{NT} is the total number of NT users. In the second part of (3), we rewrite the formula using the discrete cumulative distribution function $F[\alpha_i]$ and density function $f[\alpha_i]=n_{iNT}/N_{NT}$. The first part in brackets in (3) multiplied by the term outside them (i.e. $F[\alpha_j]\cdot N\cdot\delta/s$) measures i's schedule delay costs at t_{si} and t_{ei}. Queuing costs at these moments are the second term in the brackets multiplied. Persons with a relatively high α_i arrive at the outside of the peak. Hence, schedule delay costs increase with α_i. Queuing costs *decrease* non-linearly with α_i, since the lower the value of time is, the closer one arrives to t^*. The net effect is always that i's price increases with α_i.

$$P_i = \frac{\delta}{s}\left(\sum_{j=1}^{j=i} n_{jNT} + \alpha_i \sum_{j=i+1}^{j=M} \frac{1}{\alpha_j} n_{jNT} \right) = \frac{\delta N_{NT}}{s}\left(F[\alpha_i] + \alpha_i \sum_{j=i+1}^{j=M} \frac{1}{\alpha_j} f[\alpha_j] \right) \qquad (3)$$

Group M users have highest value of time of all users. Their price is also the highest and

equal to the price in the homogeneous user model. Hence, this highest-α-group is unaffected by the heterogeneity. With heterogeneity in α, group M–1 users enjoy lower prices than with homogeneity; because group M's larger value of time induces them to build up the queue at a slower rate than group M–1 users would.

The price for a type i is unaffected by the distribution of the groups with a lower α_j. If all types with a value of time smaller than i turned into a type i, then i's price would be unaffected. Conversely, if a type i traveller changed into a type $i+1$ ($\alpha_i<\alpha_{i+1}$), then the price for type i would decrease. The effects of a type's value of time on the price of another type can be verified in Figure 1, by imagining the effects of an increase in α_2 (keeping it between α_1 and α_3). The *isocost* curve for group 2 would then become flatter, but would intersect the curve for group 3 at the same two points. Accordingly, the prices for group 3 would be unaffected, while group 1 benefits from shifting to a lower *isocost* curve.

4. Numerical setup

This section describes the set up of the numerical base case. The capacity of the bottleneck is 3600 cars per hour, and the number of NT users is 9000. This means that the peak lasts 2.5 hours. The free-flow travel time is 30 minutes. Operating costs per trip are €7.30. The scheduling cost parameters are β=€4 and γ=€15.6 per hour, implying δ=3.1837 and η=3.9. This value of η is the same as in Arnott et al. (1990).

Figure 2: The polynomial value of time distribution in the no-toll equilibrium

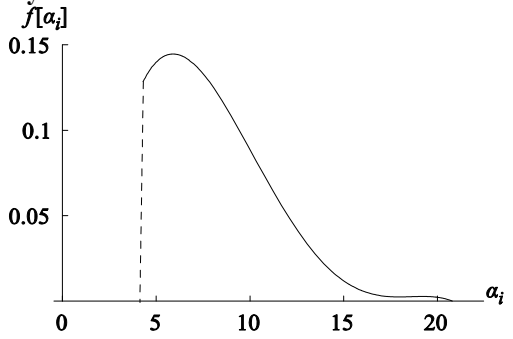
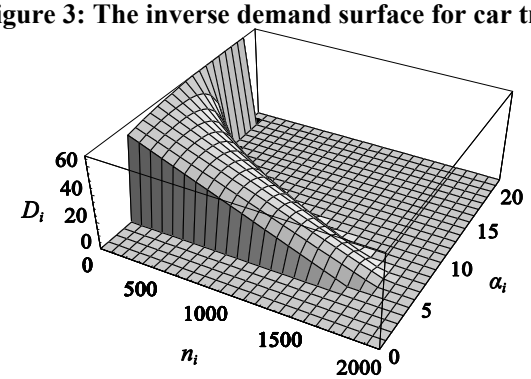


Figure 3: The inverse demand surface for car trips



We use a similar methodology as Verhoef and Small (2004) to set up demand. First, a fourth-order polynomial[3] is fitted on the lognormal value of time distribution of Van den Berg et al. (2010). We use a polynomial because it is easier to integrate, making the calculations in the numerical models easier. Still, the model works for any distribution with a probability density function (PDF) that can be twice integrated. Figure 2 plots the PDF ($f[\alpha_i]$) that is used

[3] The density function is given by the following polynomial $f[\alpha_i]=b_0+b_1\cdot\alpha_i+b_2\cdot\alpha_i^2+b_3\cdot\alpha_i^3+b_4\cdot\alpha_i^4$. The parameters are b_0=−0.177403; b_1 =0.135674; b_2 =−0.0188112; b_3 =0.00095716; b_4 =−0.0000166662.

in this paper. Note that the distribution looks log-normal. The estimated lognormal distribution ranges from zero to infinity. The lower part of the distribution has to be disregarded, because it violates the $\alpha_i>\beta$ assumption. The polynomial distribution has a maximum value per hour of travel time ($\bar{\alpha}$) of €20.80, the minimum ($\underline{\alpha}$) is €4.10. The average the value of time is €8.19, the median €7.70 and the mode €5.90.

The numerical inverse demand function is plotted in Figure 3. We choose the parameters for this demand function to accomplish three goals. First, the user distribution of Figure 2 results in the NT equilibrium. Second, the average total price elasticity is –0.4 in the NT equilibrium. This is the same elasticity that Verhoef and Small (2004) used. The total price includes the free-flow travel time and operating costs. Third, the aggregate number of NT users should be 9000.[4]

5. No-toll equilibrium with a continuously distributed α

5.1. The analytical model

This section analyses the NT equilibrium with a continuous value of time distribution. We first analyse the analytical models and then illustrate them with the results of the numerical model as calculated in Mathematica 5.0.

Section 3 derived the equilibrium price levels for discrete heterogeneity. The extension to continuous heterogeneity proved straightforward. One just has to replace the summation signs with integrals. The resulting price function is given in (4). Here, n_{jNT} is the number of NT users with a value of time of α_j. The N_{NT} is the total number of NT drivers. $F[\alpha_j]$ is the cumulative distribution function of α_j, and $f[\alpha_i]$ is the probability density function of α_i. The interpretation of the formula is otherwise the same as in the discrete case.

Formula (4) can be rewritten as (5) by inserting the $H[\alpha_i]$ function from (6) for the part between brackets in (4). The $H[\alpha_i]$ increases with the value of time (α_i), and ranges between one for the highest-α-driver and some positive fraction for the lowest-α-driver. $H[\alpha_i]$ can be viewed as giving the fraction that type i's generalised price is of the highest-α-drivers' price of $\delta N_{NT}/s$, thus allowing for a intuitive reinterpretation of (4).

$$P_i = \frac{\delta}{s}\left(\int_{\underline{\alpha}}^{\alpha_i} n_j d\alpha_j + \alpha_i \int_{\alpha_i}^{\bar{\alpha}} (n_{jNT}/\alpha_j)\, d\alpha_j \right) = \frac{\delta}{s} N_{NT}\left(F[\alpha_i] + \alpha_i \int_{\alpha_i}^{\bar{\alpha}} \frac{f[\alpha_j]}{\alpha_j} d\alpha_j \right) \qquad (4)$$

$$P_i = \delta/s\, N_{NT}\, H[\alpha_i] \qquad (5)$$

$$H[\alpha_i] = \left(F[\alpha_i] + \alpha_i \int_{\alpha_i}^{\bar{\alpha}} (f[\alpha_j]/\alpha_j) d\alpha_j \right) = \left(\int_{\underline{\alpha}}^{\alpha_i} n_j d\alpha_j + \alpha_i \int_{\alpha_i}^{\bar{\alpha}} n_j/\alpha_j\, d\alpha_j \right) / N_{NT} \qquad (6)$$

The price in (5) increases with the value of time. Queuing costs decrease with α_i and schedule delay costs increase. Still, the decrease in queuing costs is always smaller than the increase in schedule delay costs. If this were not true, the price would decrease with α_i. But this cannot be true, as then a low-α-user could always move to the time that a high-α-user arrives and face the same schedule delay and travel delay. Then, the low-α-user would again face a lower price than the high-α-user, because travel delays are less costly with a lower α_i.

Equation (7) gives the derivative of j's price with respect to the number of type i drivers. This derivative is what Lindsey (2004) calls the *congestion cost effect*. The *congestion effect* of i is the same on all type j users with a larger value of time than i's own. Surprisingly, this *congestion effect* is equal to the effect of all drivers in the homogeneous user model. On

[4] To achieve these goals we first set $b_i[\alpha_i]$ equal to $f[\alpha_i]$ and A_i equal to the NT total price. The mean total price elasticity is a function of the mean price, number of users, and B. The mean price depends on the value of time distribution and number of users. Hence, the elasticity is directly calibrated by B. Finally, A is set such that the number of users is 9000. If $b_i[\alpha_i]=f[\alpha_i]$ and $A_i=P_{iNT}$, then equating inverse demands to prices, and rewriting, results in $N_{NT}=A/B$ and $n_{iNT}=b_i[\alpha_i]\cdot A/B$. Thus, A determines the number of drivers and $b_i[\alpha_i]$ the number of users for each value of time (n_{iNT}). The values used for B and A are 0.0050477 and 45.4291.

smaller value of time drivers, *i*'s *congestion effect* is smaller, and it decreases with the relative size of α_i to α_j. Note that these results are consistent with the discussion in Section 3.

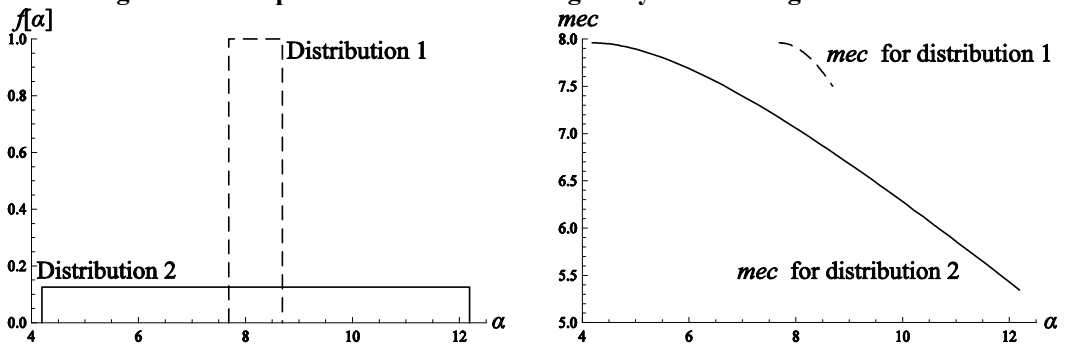
Multiplying (7) by n_j and integrating the result gives the marginal external costs (mec_i) of a type *i* driver (i.e. the marginal externality) in (8). A driver's *mec* is the sum of the congestion effects she imposes on all other drivers. The highest-α-drivers cause the lowest externality of $E[\alpha]\cdot\delta\cdot N_{NT}/(s\cdot\bar{\alpha})$. The $E[\alpha]$ is the (weighted) average of the value of time. $\bar{\alpha}$ is the maximum value of time. The lowest-α-drivers cause the highest externality of $\delta\cdot N_{NT}/s$. The smaller α_i is, the larger *i*'s externality. The marginal externality of a lowest-α-driver equals the one of all users with homogeneity. The marginal external cost of a lowest-α-driver is above the own price, while the opposite holds for a highest-α-driver. In comparison, with homogeneity, the *mec* equals private cost for all, and marginal social costs are therefore twice private costs.

$$\partial P_j / \partial n_i = \begin{cases} \delta/s & \alpha_j \geq \alpha_i \\ (\delta/s)\alpha_j/\alpha_i & \alpha_j < \alpha_i \end{cases} \tag{7}$$

$$\text{mec}_i = \int_{\underline{\alpha}}^{\bar{\alpha}} (\partial P_j / \partial n_i) n_j d\alpha_j = \frac{\delta}{s} N_{NT} \left(1 - F[\alpha_i] + \frac{1}{\alpha_i}\int_{\underline{\alpha}}^{\alpha_i} \alpha_j f[\alpha_j] d\alpha_j\right) \tag{8}$$

The average externality decreases the heterogeneity in α. This follows from the fact that a lowest-α-driver causes the highest congestion effects that equal those with homogeneous users; while all other users cause lowers congestion effects on users with smaller values of time than them. This result is best further explained using two uniform value of time distributions. The left panel of Figure 4 plots two value of time distributions, the right panel shows the resulting *mec*'s. Besides the different distributions, the set up underlying the figure is the same as in the numerical base case. If the amount of heterogeneity (i.e. the spread) increases, there are new users with values of time below the old lowest value; in the figure with α_i*'s* between €4.19 and €7.69. The congestion effects on these *new low-α-users* caused by the other users are low, since the effect on *j* by *i* decreases with α_i and increases with α_j. There are also *new types* with values of time exceeding the old highest value. These *new high-α-types* cause low congestion effects as their values of time are relatively high. Consequently, the larger spread is, the lower average externality. Finally, since the externalities decrease with the heterogeneity in α, the average travel price also decreases.

Figure 4: Example of the effect of heterogeneity on the marginal externalities



5.2. Numerical base case results

Figure 5 shows the numerical no-toll (NT) schedule delay costs, queuing costs and their sum. The capacity of the bottleneck is 3600 cars per hour and there are 9000 users. Thus, the peak starts at $t_s = -\delta\cdot N_{NT}/(s\beta) = -1.99$ hour (i.e. 1.99 hours before t^*) and ends at 0.51 hours.

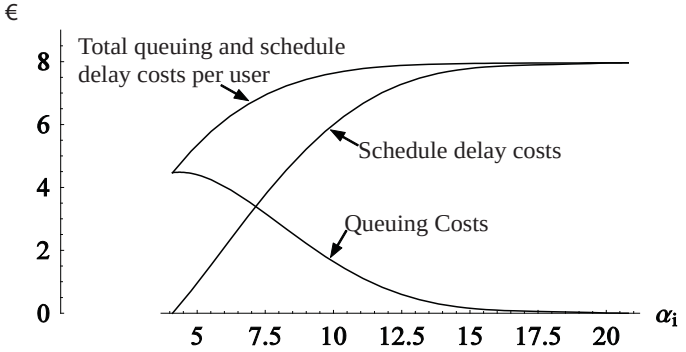
For an α_i of €7.15 per hour, schedule delay costs equal queuing costs. At this value, the cumulative distribution function of the value of time is 0.43. Total travel delay costs are

€24,584 and total scheduling costs are €35,001. There is a difference here with the homogeneous user model, where these totals are equal. In the public discussion of congestion the travel time costs usually dominate, whereas the schedule delay costs are less emphasised. Conversely, in our model, the schedule delay costs are more important for most persons, as well as on aggregate. Still, from a policy perspective travel delays might be more important, as these can be influenced more effectively through tolling.

Type i's total price is a concave increasing function of α_i. Queuing costs *decrease* with α_i, because the higher one's value of time, the closer one arrives to the outside of the peak. The distribution of α has between about €15 and €20.80 per hour a flat tail with very low densities. This causes the price function in Figure 5 to be rather flat in this range.

To sum up, this section found that the externalities of all except the lowest value of time decrease with heterogeneity in α. Accordingly, the mean NT price also decreases with this heterogeneity.

Figure 5: Decomposition of the travel price for each value of time



6. First-best public toll

Arnott et al. (1988) find that, also with heterogeneity, the social optimal toll eliminates all queuing, since queuing is always wasteful. All NT users could arrive in the same period without queuing if people arrived at the bottleneck in a flow equal to capacity. Arnott et al. (1994) note that the result of the homogeneous user model, that the price is the same in the NT and FB equilibria, breaks down with heterogeneity. We extend their analysis by using continuous heterogeneity and price sensitive demand. This section derives the closed-form solutions for the FB price and number of users per value of time. The numerical subsection finds that tolling increases the price for most users, but that the highest values of time gain.

6.1. The analytical model

To arrive at the first-best public (FB) optimum, the toll is set so that everyone is indifferent as to what moment they arrive between t_s and t_e when travel delays are absent. This requires the slope of the toll schedule to be β for early arrivals and $-\gamma$ for late arrivals. N_{FB} indicates the number of FB users. The toll $\tau[t]$ can be split into a time-*invariant* part $\overline{\tau}$ and a time-*variant* part $\tau_t[t]$. The optimal time-*variant* toll is zero at t_s and t_e and has a maximum for arrivals at t^*. A no-queuing equilibrium is achieved by the time-variant toll of (9). With this toll, the price is time-invariant provided there is no queuing. Hence, there is no incentive to cause a queue. The optimal FB time-invariant toll is zero. If it is non-zero, it only distorts prices and does not alleviate any externality: the price (including toll) now equals social costs.

Given the discussed toll and elimination of the queue, prices are given by (10). The price is for all drivers equal to $\delta N_{FB}/s$. Without tolling there were large differences in generalised prices. Both the average scheduling cost and average toll are now $1/2 \cdot \delta N_{FB}/s$. The toll equals, at each arrival moment, the difference between marginal social costs and private costs.

$$\tau_t[t] = \begin{cases} \delta N_{FB} / s + \beta t & t < 0 \\ \delta N_{FB} / s - \gamma t & t \geq 0 \end{cases} \tag{9}$$

$$P_{iFB} = CSD_{FB}[t] + \tau_t[t] = \frac{\delta}{s} N_{FB} \tag{10}$$

For low-α-drivers FB prices are much higher than NT prices. Thus, these users are hurt by FB tolling. Hence, their demand and consequently total demand decrease. This lowers the price for high-α-drivers, thereby making them better off. With perfectly inelastic demand, the welfare losses of the low-α-drivers are larger, since they cannot change their behaviour. In this case, the highest-α-drivers' price is unaffected by the FB toll, while all other users lose.

To find the FB equilibrium, we use that $P_{iFB}[N_{FB}]-P_{iNT}[N_{NT}]=D_{iFB}[n_{iFB}]-D_{iNT}[n_{iNT}]$ must hold for all i. Here n_{iNT} and n_{iFB} stand for the number of type i users in the no-toll (NT) and first-best (FB) cases. Inserting i's demand function (2) and price functions (5) (NT case) and (10) (FB case) into the above equality results in (11).

$$-\left(B / b_i[\alpha_i]\right)(n_{iFB} - n_{iNT}) = \delta \cdot N_{FB} / s - \delta \cdot N_{NT} \cdot H[\alpha_i] / s \tag{11}$$

Equality (11) can be rewritten into (12). This equation gives the number of type i drivers conditional on the total number of FB drivers. The $H[\alpha_i]$ function remains as defined above, and therefore only reflects ratios of NT price and not of FB prices.

$$n_{iFB} = n_{iNT} + b_i[\alpha_i]\, N_{NT} \left(\frac{\delta}{s\, B} H[\alpha_i] - \frac{\delta}{s\, B} b_i[\alpha_i]\right) N_{FB} \tag{12}$$

The closed-form solution for the total number of FB case users, in (13), can be obtained by integrating (12) over all values of time and rewriting the result. From (13), it is clear that the number of users decreases due to the toll. In the formula N_{NT} (number of NT users) is multiplied by a term that is smaller than one, since the integral of $H[\alpha_i]\cdot b_i[\alpha_i]$ is smaller than one. $H[\alpha_i]$ ranges between zero and one. The $b_i[\alpha_i]$ integrates to one. Hence, the integral of $H[\alpha_i]\cdot b_i[\alpha_i]$ is smaller than one. Equation (13) also shows that the number of users always decreases due to FB tolling, unless if there is no heterogeneity in α. The economic reason for the decrease in users is that FB tolling increases the price for most values of time.

$$N_{FB} = N_{NT} \left(\frac{1 + \frac{\delta}{s\, B} \int_{\underline{\alpha}}^{\overline{\alpha}} H[\alpha_i]\, b_i[\alpha_i]\, d\alpha_i}{1 + \frac{\delta}{s\, B}} \right) < N_{NT} \tag{13}$$

Inserting (13) back into (12) gives the equilibrium number of type i users. From (12) it is easy to find who gains from the toll and who loses. If $N_{NT}\, H[\alpha_i]=N_{FB}$, the number of type i users is the same in the FB and NT case, because then the last two elements in (12) cancel each other out. If the number of type i users remains constant, this implies that i's price remains constant. If $H[\alpha_i]\cdot N_{NT}$ is larger than N_{FB}, which is the case for the higher-α-users, these users gain from FB tolling, while the lower-α-users lose.

6.2. *Numerical results*

Figure 6 shows the effect of FB tolling on prices excluding free-flow travel time and operating costs. For an α of €11.08 per hour, the price is the same before and after the introduction of the toll. Higher values of time gain, whereas lower value lose. This means that 26.2% of the NT users gain from FB tolling; for 73.8% prices increase.

The left part of Figure 7 shows the average change in consumer surplus by value of time (α). The average change is calculated by dividing the change in surplus for a value of time by the number of NT users with that value. The right part of Figure 7 plots the average change in *consumer welfare* after toll revenues are returned equally to all NT users (i.e. it is the sum of the consumer surplus and received toll revenue refund).[5] Before the refund, all users with a value of time below €11.10 are made worse off, whereas after it all drivers are better off.

Toll revenue is €34,244. Aggregate consumer surplus loss is €8875. Hence, welfare (i.e. consumer surplus plus toll revenues) increases by €25,369. The decrease in the total number of users is 200, which is a modest reduction of 2.2 percent. Finally, travel time decreases on average by 27 minutes, and all travel delays are removed.

Consumer surplus losses for the 50% of *lower-α-drivers* are quite large. This might make the scheme politically difficult to accept, even though aggregate welfare increases. This also makes it less likely that a FB tolling scheme would pass in a democratic society, if only because the majority of drivers would have a strong motivation to oppose the toll.

To conclude, FB tolling eliminates all queuing. The high values of time enjoy a price decrease due to this, whereas the medium and lower values face a price increase.

Figure 6: Generalised price for a user in the NT and FB toll cases

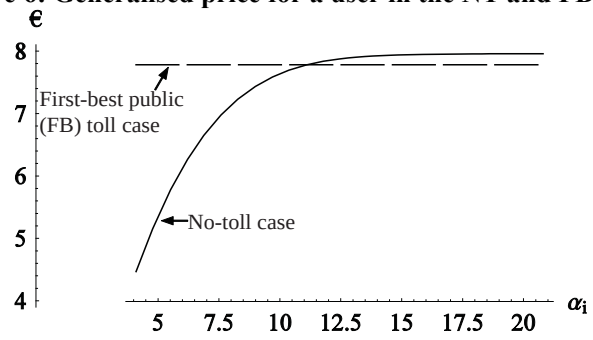
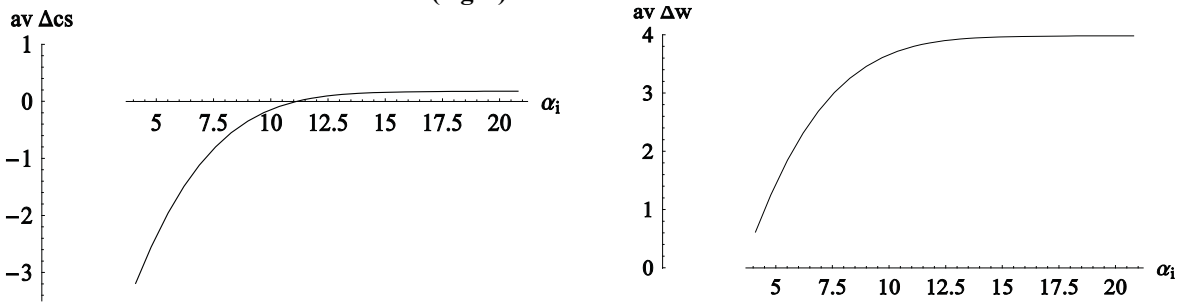


Figure 7: Average change in consumer surplus (left) and consumer welfare when the toll revenue is returned (right) due to the introduction of the FB toll



7. A private monopoly controlling the road

7.1. The analytical model

We now turn to the case where the road is controlled by a monopolist who sets a profit-maximising time-variant PM toll. The monopolist internalises the queuing costs of the users, since any decrease in queuing costs can be met by an equal increase in the toll and thus add to profit. Hence, a monopolist eliminates all queuing by using the same formula for the time-variant toll as the FB operator. On top of this a time-invariant toll is set, which maximises total toll revenues.

[5] Average surplus change is av $\Delta cs_i=(cs_{iFB}-cs_{iNT})/n_{iNT}$, average consumer welfare change is av $\Delta w_i=(cs_{iFB}+(\text{toll revenue}):f[\alpha_i]-cs_{iNT})/n_{iNT}$.

The time-variant toll equals the difference between marginal social costs and private costs. The mean time-variant toll is $\delta N_{PM}[\overline{\tau}_{PM}]/(2s)$. The toll revenue in (14) includes the mean time-variant toll and the time-invariant toll ($\overline{\tau}_{PM}$).

$$\Pi_{PM} = (\overline{\tau}_{PM} + \delta N_{PM}[\overline{\tau}_{PM}]/(2s))N_{PM}[\overline{\tau}_{PM}] \tag{14}$$

The price formula (15) for the PM case is similar to the FB price equation. The solution of this equilibrium uses the same procedure as for the FB case. Equation (16) is the resulting formula for the total number of users (N_{PM}). Inserting (16) into profit formula (14) and maximising it to $\overline{\tau}_{PM}$ results in equation (17) for the optimal time-invariant toll. The optimal $\overline{\tau}_{PM}$ decreases with the heterogeneity in the value of time: with more heterogeneity, no-toll prices are lower, and hence users are less willing to pay the monopolistic toll.

$$P_{iPM} = CSD_{PM}[t] + \tau_{PM}[t] = \delta N_{PM}/s + \overline{\tau}_{PM} \tag{15}$$

$$N_{PM}[\overline{\tau}_{PM}] = \frac{N_{NT}\left(1 + (\delta/(s\,B))\int_{\underline{\alpha}}^{\overline{\alpha}} H[\alpha_i]\,b_i[\alpha_i]\,d\alpha_i\right) - \overline{\tau}_{PM}/B}{1 + \delta/(s\,B)} \tag{16}$$

$$\overline{\tau}_{PM} = \frac{s\,B^2 N_{NT} + \delta\,B\int_{\underline{\alpha}}^{\overline{\alpha}} H[\alpha_i]\,b_i[\alpha_i]\,d\alpha_i\;\;N_{NT}}{2\,s\,B + \delta} \tag{17}$$

7.2. The numerical model

Figure 8 shows the PM prices. They are much higher than FB and NT prices, because of the high average toll of €26.10, of which €24.00 is from the time-invariant toll. This high toll might seem surprising. Yet, with the low price elasticity, a high monopoly price is to be expected. Verhoef and Small (2004) also found a high PM toll, although in their model the difference between the PM and FB tolls is smaller. The number of users drops to 4755, a decrease of 47%. The average time-variant toll (€2.10) is lower than its FB counterpart (€3.89) since there are fewer PM users. Total consumer surplus drops by 72% to €57,148. Welfare decreases by 11% to €204,431.

Figure 8: Generalised price for a traveller in the NT, FB toll and PM cases

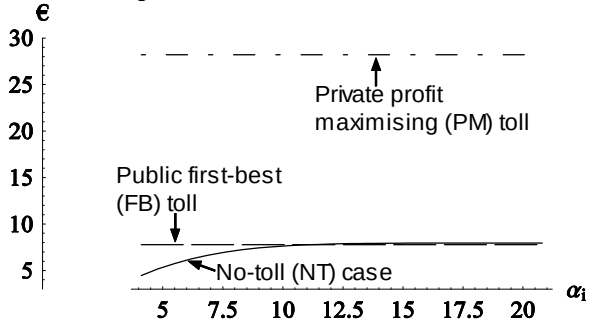
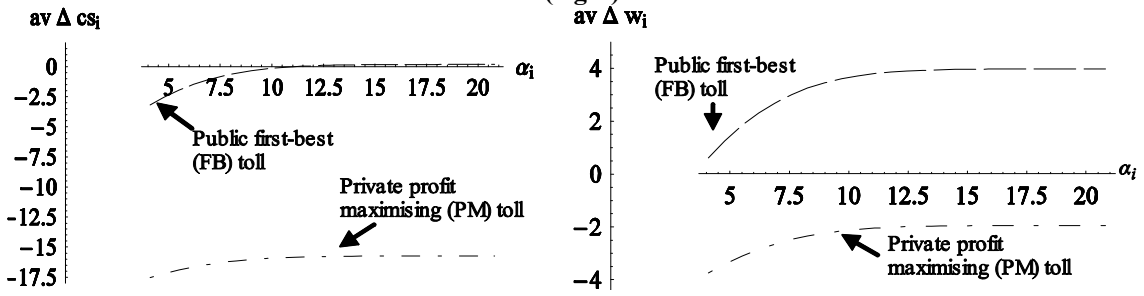


Figure 9 (left) shows the differences in average consumer surplus between the PM and NT equilibria. The right part shows the average consumer welfare changes after toll recycling, which is of course less likely for a private operator. The PM curves are not exactly downward shifted copies of the FB curves, because the price elasticity differs over the values of time. All values of time lose due to monopolistic tolling; although *lower-α-drivers* lose most. This is due to the same reason they lost in the FB case: time-variant tolling eliminates all queuing, but increases the price for *lower-α-drivers.*

Figure 9: Average change in consumer surplus (left) and consumer welfare when the toll revenue is returned (right) due to the PM toll



8. Second-best tolls with a continuously distributed value of time

This section studies some second-best policies. Firstly, public (PL) and private pay-lanes (PPL) are analysed. With a pay-lane, part of the capacity is separated, and to use this pay-lane a toll has to be paid. The remaining capacity, the free-lane, is toll free. A political advantage of this policy is that it does not force people to pay a toll if they want to travel. Another second-best policy we analyse is the public time-invariant (TI) toll. This option circumvents a practical problem of the other tolls that have to continuously change over time. Thus, the TI toll is easier to implement, maintain and understand. For instance, in the Oslo toll ring scheme a fixed toll is asked during the entire day (Odeck and Bråthen, 1997). In the London congestion charging scheme a fixed toll applies between 7 am and 6 pm.[6] An example of a more flexible toll is the HOT-lane (pay-/carpool-lane) on the I-15 in California, where the toll can be changed every six minutes (Brownstone et al., 2003). Also on the I-394 HOT-lane in Minnesota the toll can be changed frequently.[7]

Unlike for the previous policies, for the second-best policies we were unable to find closed-form solutions. Hence, for these policies we discuss the analytical results in so far they are available and then describe the numerical solution methods.

8.1. The analytical model for the pay-lane cases

Braid (1996) uses a homogeneous user bottleneck model to study a public pay-lane. He shows that the public pay-lane has a time-variant toll, to eliminate queuing, and a negative time-invariant toll (i.e. a subsidy). The pay-lane has no queuing, whereas the free-lane does. Hence, when the time-invariant toll is zero, the pay-lane has lower social marginal costs. Thus, it is attractive to draw extra users from the free-lane, by setting a time-invariant subsidy. Our public pay-lane (PL) model differs from Braid's (1996) in that we add heterogeneous users. However, the same logic applies: it is optimal to eliminate all pay-lane queuing, and a negative time-invariant toll attracts extra users to the pay-lane.

Because the PL operator sets a negative time-invariant toll, the peak starts earlier and ends later on the pay-lane than on the free-lane. Hence, maximum and average schedule delays on the public pay-lane are higher. The optimal subsidy follows from a trade off between the extra schedule delay costs it causes and the travel delay reductions that can be achieved by having drivers use the pay-lane instead of the free-lane.

If a profit-maximising firm operates the pay-lane (PPL), then the optimal toll again contains the same time-variant toll equation, plus a time-invariant term that maximises revenue. De Palma and Lindsey (2000) note that a PPL operator has the incentive to eliminate queuing on its pay-lane because any decrease in queuing costs can be met by an increase in toll. The PPL profit-maximising time-invariant toll is lower than in the monopolist PM case,

[6] This is according to the website of Transport for London on 18 January 2010, see www.tfl.gov.uk/tfl/roadusers/congestioncharge/whereandwhen/

[7] This follows the information on www.mnpass.org as retrieved by us on 1 March 2010.

since the pay-lane faces the competition of the free-lane. Because of the positive time-invariant toll, there are fewer pay-lane users and more free-lane users in the PPL case than in the PL case. Hence, average schedule delay costs are higher on the PPL's free-lane. The peak is shorter on the PPL's pay-lane than on its free-lane.

The pay-lane has a share ρ of capacity and the free-lane the remainder. The numerical model uses 1/3 for ρ. The pay-lane is indicated by lane 1 and the free-lane by lane 2. The total numbers of users on each lane are N_1 and N_2; the total number of users is N_p. Subscript p indicates a pay-lane equilibrium. Although the pay-lane's time-variant toll follows the same formula as the FB toll, the actual levels differ as the ratio of number of users to capacity differs.

The pay-lane is only used by values of time that equal or are larger than the critical value of time (α^*). The free-lane is used by values not exceeding α^*. For low-α-users the pay-lane's shorter travel time is not worth the toll. A PLL's α^* is higher than a PL's, as a smaller share of drivers uses the pay-lane. The α^* users can arrive at any moment on the pay-lane or arrive at the outside of the free-lane peak. If there is a queue, the higher one's value of time, the further from t^* one arrives. On the free-lane, α^* drivers face no queue and have the highest price. The pay-lane price is the same for all users (if free-flow travel is zero). In equilibrium, the price for α^* drivers is equal on the free-lane and pay-lane.

Equation (18) gives the equilibrium price in a pay-lane case. The price function for the pay-lane drivers, the upper part of (18), is basically the PM formula (15). The difference is that capacity is now a fraction ρ of total capacity. The free-lane price function, the lower part of (18), is basically the same as the NT price function in (5). In the second part of this equation, we inserted the $K[\alpha_i]$ function. This function can be viewed as giving i's free-lane generalised price as a fraction of the α^* users' price. Hence, this $K[\alpha_i]$ function has, for the pay-lane equilibrium, the same interpretation as the $H[\alpha_i]$ function has for the NT equilibrium.

$$P_{ip} = \begin{cases} P_{i1} = (\delta/(s\cdot\rho))\cdot N_1 + \bar{\tau}_p & \bar{\alpha} \geq \alpha_i \geq \alpha^* \\ P_{i2} = \dfrac{\delta}{s(1-\rho)}\left(\int_{\underline{\alpha}}^{\alpha_i} n_{j2}\, d\alpha_j + \alpha_i \int_{\alpha_i}^{\alpha^*} \dfrac{n_{j2}}{\alpha_j} d\alpha_j \right) = K[\alpha_i]\ \dfrac{\delta}{s(1-\rho)} N_2 & \underline{\alpha} \leq \alpha_i \leq \alpha^* \end{cases} \tag{18}$$

Differentiating j's free-lane price equation (18) with respect to n_{i2} gives i's congestion effect on j. This formula is given in (19). The total marginal externality of i in (20) is found by integrating i's congestion effect multiplied by n_{j2} over all values of time of the free-lane users. The situation on the pay-lane has no direct effect on the free-lane externalities. Formula (20) shows that users with a value of time of α^* cause the lowest externalities. The externality of i decreases with α_i and increases with the number of free-lane users. The mean free-lane congestion externality decreases with the heterogeneity.

$$\partial P_j / \partial n_i = \begin{cases} \delta/(s(1-\rho)) & \alpha_j \geq \alpha_i \\ (\delta/(s(1-\rho)))\alpha_j/\alpha_i & \alpha_j < \alpha_i \leq \alpha^* \end{cases} \tag{19}$$

$$\text{mec}_i = \int_{\underline{\alpha}}^{\alpha^*} n_{j2}\, \partial P_{j2}/\partial n_{i2}\, d\alpha_j = \frac{\delta N_{NT}}{s(1-\rho)} \left(\int_{\alpha_i}^{\alpha^*} n_{j2} d\alpha_j + \frac{1}{\alpha_i} \int_{\underline{\alpha}}^{\alpha_i} (\alpha_j\, n_{j2}) d\alpha_j \right) \tag{20}$$

The first issue that makes it impossible to find closed-form solutions is that the distribution of n_{i2} is not known. The second is that there is no closed-form solution for α^*. For a given α^* and time-invariant toll, there is a solution for the prices and number of users for each value of time on the pay-lane. For a given α^* and free-lane equilibrium's value of time density function, there is a solution for the free-lane's prices and total number of users. The derivations of these solutions are shown in appendix C. The α^* and free-lane's density function are, however, not known. Hence, we use the numerical solution method of appendix D.

8.2. The analytical model for the public time-invariant toll

In practice it is unlikely that a fully time-variant toll can be set. Tolling schemes usually have one fixed toll during the entire peak or day. The public time-invariant (TI) toll is constant during the entire day. In the homogeneous user bottleneck model, TI tolling gives a far lower welfare gain than FB tolling (Arnott et al., 1990). The TI toll cannot eliminate queuing; it only reduces it by lowering demand. The toll is set to balance the welfare loss from tolling the small externality causers too much, and the loss of tolling large externality causers too little. Equation (21) gives the TI price. Again there seems to be no closed-form solution for the TI equilibrium. Therefore, Appendix D also discusses the numerical solution for this case.

$$P_{iTI} = (\delta / s)\left(\int_{\underline{\alpha}}^{\alpha_i} n_{jTI} d\alpha_j + \alpha_i \int_{\alpha_i}^{\overline{\alpha}} (n_{jTI} / \alpha_j)\, d\alpha_j \right) + \overline{\tau}_{TI} \tag{21}$$

8.3. The numerical models for the second-best policies

Table 2 summarises the results in the numerical base case of the policies. The average toll is highest in the PM case, followed by the TI case, the private pay-lane (PPL), the FB case and the public pay-lane (PL). The relative efficiency of the PL is 0.37 and of the PPL it is 0.12. Relative efficiency is the welfare change of a policy compared with the NT case relative to the FB's welfare gain. Under the PL, the total number of users is 1% higher than in the NT case. The PM toll's relative efficiency is –0.91. Hence, in our model, the welfare loss from a PM toll is almost as large as the gain from FB tolling.

Table 2: The numerical results for the policies

	NT	FB	PPL	PL	TI	PM
Total number of users	9000	8800	8781	9086	7962	4755
Total number of pay-lane users	-	-	2291	4039	-	-
Total number of free-lane users	-	-	6490	5047	-	-
Critical value of time	-	-	€9.85	€8.18	-	-
Total consumer surplus	€204,431	€195,555	€194,596	€208,400	€160,015	€57,148
Average toll	-	€3.89	€5.57	€ 1.34	€ 6.02	€26.10
Average time-variant toll	-	€3.89	€3.04	€ 5.36	-	€2.10
Time-invariant toll	-	-	€2.53	−€4.02	€6.02	€24.00
Total toll revenue	-	€34,245	€12,760	€5399	€47,906	€124,104
Social welfare	€204,431	€229,800	€207356	€213,799	€207,921	€181,252
Relative efficiency	0.00	1.00	0.12	0.37	0.14	−0.91

The PL's welfare gain is more than three times the PPL's. In the Verhoef and Small (2004) static flow congestion model, welfare with a PPL is lower than in the NT equilibrium. This illustrates an important difference between flow and bottleneck congestion: the pay-lane is more beneficial with bottleneck congestion. In the de Palma and Lindsey (2000) homogeneous-users model, the relative efficiencies of the PL and PPL policies are somewhat higher than in this paper, namely 0.57 and 0.29. Social welfare under the *public* TI toll is comparable to welfare under the *private* pay-lane and far below the PL's gain. This shows the benefit of setting a time-variant toll.

Figure 10 plots the prices in the second-best cases and compares these with the FB and NT prices. The prices are highest for the TI toll (although they are still lower than the PM prices) and lowest for the PL. The left part of Figure 11 shows the average changes in consumer surplus due to the policies. The right part gives the average consumer welfare changes when the toll revenues are returned equally to all NT drivers. Before the revenues are recycled, all drivers are better off under the PL than with the FB toll. Most values of time enjoy a higher

consumer surplus under the PL than in the NT case. Only the 31% lowest-α-drivers in the NT case, with α_i<€6.31, are worse off under the PL.

For the pay-lane policies, Verhoef and Small (2004) conclude that the middle group (with a value of time around α^*) have the largest losses, the lower-α-drivers lose less, and higher-α-drivers gain. In their paper, the average consumer surplus change curves for the PL and FB have two linear sections which meet at α^* (i.e. they are piecewise linear). In our paper, the average change curve for the FB case is concave; for the PL, it is partly concave and partly convex. Moreover, in our study, all change curves have continuous derivatives.

Under the PPL, before the toll revenues are returned, all users lose relative to the NT case. Still, if revenues are returned, all gain. The medium values of time types (6 to 9 euros per hour) are hurt most. The average consumer surplus change curve of the PPL case is more comparable in shape to the curve found by Verhoef and Small (2004), in that the middle group loses most. A difference is that, in our bottleneck model, it is not the α* users who lose most, but intermediate free-lane users, although the α* drivers are among those who lose most. The *lower-α-drivers* (α<€6.55) are better off under the PPL than under the FB toll. Still, after toll recycling all users are best off under the FB toll.

A reason to prefer a pay-lane policy might be because the lower values of time lose less under a (private) pay-lane than under a FB toll. A second reason could be that the government cannot finance capacity expansion, and a private operator is asked to build a new pay-lane. Of course, in this case, the calculation of the welfare change should account for the effects of capacity expansion, which, as Arnott et al. (1988) find, are also affected by heterogeneity.

Figure 10: Generalised price (excluding operating costs and free-flow travel time) in the NT, FB, PL and PPL cases (left) and in the NT, FB and TI toll cases (right)

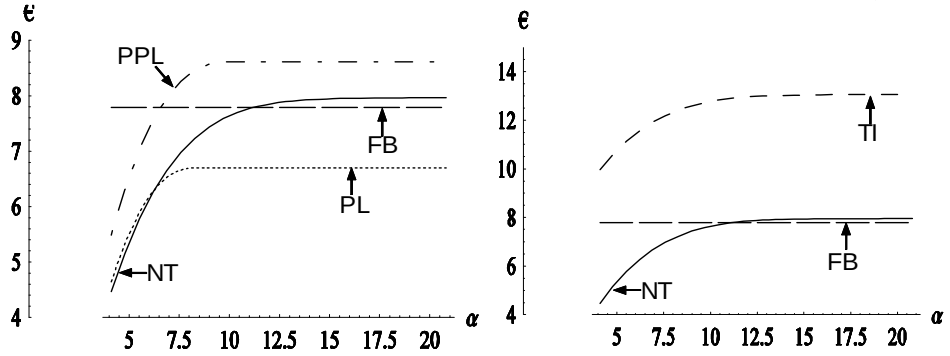
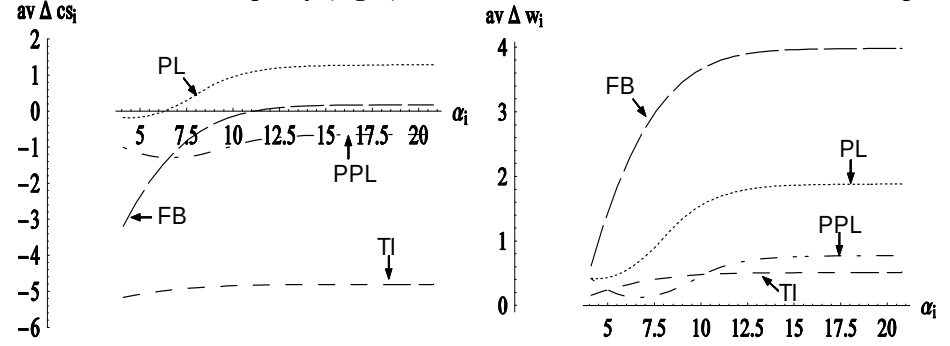


Figure 11: Average change in total consumer surplus (left) and consumer welfare when toll revenues are returned equally (right) due to the introduction of the second-best policies



8.5. Conclusion

With static congestion, drivers that are indifferent between the pay-lane and free-lane are hurt most by a pay-lane. With the bottleneck model's private pay-lane, there are free-lane users

that lose more than the indifferent users. With our public pay-lane, the indifferent users gain more than all free-lane users, while the price for the low-α free-lane users increases. If it is only possible to set a public time-invariant (TI) toll, welfare is much lower than in the FB case. TI tolling cannot eliminate queuing; it only limits congestion by lowering demand.

The congestion externality on the free-lane decreases with heterogeneity in α. The higher a free-lane user's value of time, the smaller her externality is; but the higher the price she faces.

9. Sensitivity Analysis

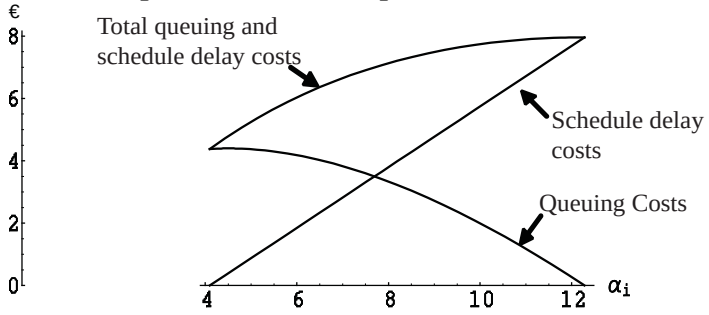
This section presents some sensitivity analyses. First, we check whether the results in the no-toll (NT) case depend on the value of time distribution used. Second, we study how the effects of FB tolling vary over different distributions of the value of time in the NT situation. Third, we look at the effects of different distributions of the value of time and price elasticities on the relative efficiencies of the second-best policies. Finally, the relative efficiencies of the pay-lane policies are shown for different shares of total capacity for the pay-lane.

9.1. NT equilibrium with a uniform value of time distribution

Figure 12 shows the decomposition of NT prices for a uniform value of time distribution, just as Figure 5 did for the base case. The uniform distribution has the same mean (€8.19) and minimum (€4.10) as the base case. The shapes of the curves in Figures 5 and 12 are comparable. Queuing costs decrease and scheduling costs increase with the value of time.

A difference between the two figures is that the schedule delay costs are linear in the value of time for the uniform distribution. NT schedule delay costs are linear in the value of the CDF (cumulative distribution function) of the value of time. Since a uniform distribution has a linear CDF, scheduling costs are now linear in α_i. A second difference is that queuing costs are now concavely decreasing with α_i. In the base case, queuing costs are convexly decreasing with α_i for high to intermediate values of time, and concavely decreasing for the lower values. Still, also, with the uniform distribution, prices increase concavely with α_i. This suggests that the general shape of the NT price curve is robust to the density function used.

Figure 12: Decomposition of the NT price for a uniform distribution of time



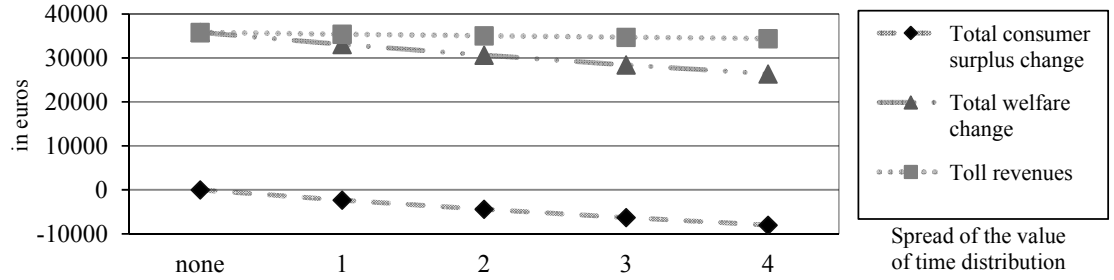
9.2. Welfare effects of an FB toll with different value of time distributions

Figure 13 depicts the total consumer surplus change, welfare gain and toll revenue from FB tolling for different NT distributions of the value of time (α). The column on the far left shows the results for the homogeneous user model. The other columns are for uniform distributions with different spreads, but with the same mean of €8.19. The figure tests how heterogeneity affects the first-best FB toll. The minimum value of time must exceed €4, for the $\alpha_i > \beta$ assumption to hold. Hence, the maximum spread is €4.18. Consequently, using mean-preserving spread alterations limits the amount of heterogeneity.

For the homogeneous user bottleneck model there is no loss in consumer surplus from FB tolling; accordingly, the welfare gain equals the toll revenues. FB consumer surplus and toll

revenues decrease with the spread. With more heterogeneity in α, the mean congestion externality and NT prices are lower, which means that there is less to gain from tolling.

Figure 13: Welfare changes from a FB toll for different value of time distributions



The FB welfare gain with homogeneity is 17.5% of NT welfare, whereas with a *uniform distribution with a spread of 4* it is 12.9% and in the base case (not shown in Figure 13) 12.4%. Hence, heterogeneity in the value of time has a large effect on FB tolling.

Consumer surplus losses in the base case are above those of the uniform distributions, presumably because the base case distribution has more heterogeneity. On the whole, however, the results from the base case distribution and uniform distributions are comparable. This suggests that the main results of our model do not depend heavily on the type of distribution, although they do depend on the amount of heterogeneity.

Figure 14a: Relative efficiencies for different value of time distributions in the NT case

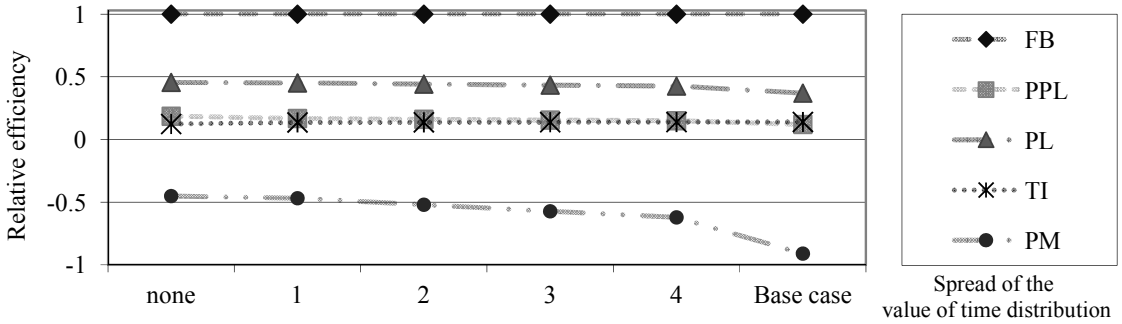
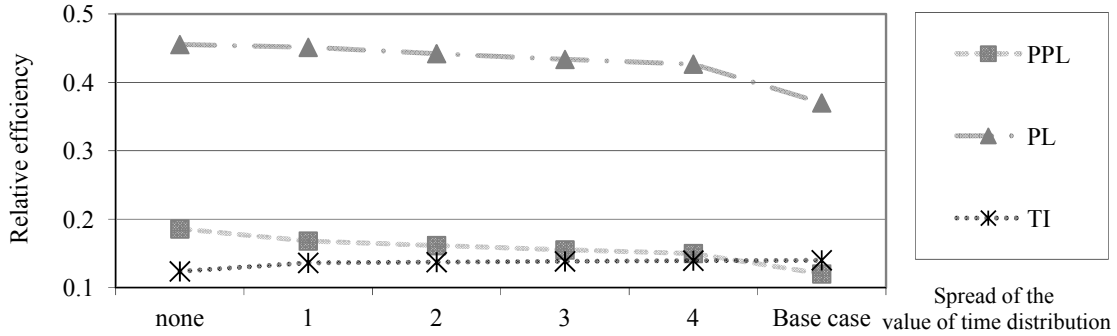


Figure 14b: Relative efficiencies of the policies excluding the PM and FB



9.3. Second-best policies with different value of time distributions

Figure 14a studies the relative efficiencies of policies for different NT value of time distributions. Figure 14b is the same figure as Figure 14a but without the curves for the PM and FB cases: the inclusion of these curves affects the visibility of the patterns for the other

curves. The column on the far left gives the results for homogeneous users. The middle columns give the results for uniform distributions of the value of time with the same mean of €8.19 but different spreads. The more to the right one moves, the larger the spread. The column on the far right displays the results for the base case.

With more heterogeneity in α, the welfare gain and relative efficiency of the PL, PPL, and PM toll are lower. The PPL's relative efficiency is 35% lower in the base case than with homogeneity (0.12 versus 0.19); for the PL, the relative efficiency is 19% lower (0.37 versus 0.46). These relative efficiencies decrease even though the first-best welfare gain also decreases. The TI's welfare gain is hardly affected by heterogeneity, but since the FB welfare gain decreases with heterogeneity, the TI toll's relative efficiency slightly increases.

9.4. Detailed analysis of the sensitivity analysis for the public pay-lane

This subsection investigates why the PL's relative efficiency decreases with heterogeneity. This is an unexpected conclusion: the intuitive result is that because the average value of time on the pay-lane increases with heterogeneity, the queue elimination should be more valuable. It also contradicts the insights of Verhoef and Small (2004).

To get a better insight, Table 3 gives some characteristics of the PL equilibria for the four uniform value-of-time distributions. As expected, the average value of time on the pay-lane increases with the spread. A surprising result is that for a larger spread, the optimal time-invariant subsidy is smaller. Thus, with more heterogeneity, it is optimal to attract fewer drivers to the queue-free pay-lane. The number of pay-lane users decreases with heterogeneity, whereas the number of free-lane users increases. Accordingly, the free-lane's share of the number of users and average free-lane queuing time also increase.

The question is: why is it optimal, with more heterogeneity, to allow more congestion on the free-lane? If the time-invariant toll (i.e. subsidy) is increased, the free-lane's highest-α-users switch to the pay-lane. These users gain relatively little from this, since they already arrived at the outside of the free-lane peak and faced short queues. This is an important difference from a flow-congestion PL, where the users who switch receive substantial travel time gains. A subsidy increase lowers free-lane queuing costs and raises total scheduling costs. The optimal subsidy is at the point where, for a marginal subsidy increase, the welfare gain from reducing the queue equals the loss from the longer schedule delays.

Table 3: Characteristics of the PL equilibria for four uniform value of time distributions

Spread of the distribution	**1**	**2**	**3**	**4**
Relative efficiency PL case	0.46	0.45	0.44	0.43
Welfare PL case	€219,068	€217,503	€216,198	€215,110
Welfare NT case	€204,633	€204,633	€204,633	€204,633
Welfare FB case	€237,711	€235,269	€233,052	€231,007
Critical value of time	€8.26	€8.37	€8.51	€8.68
Time-invariant toll	−€4.86	−€4.57	−€4.29	−€4.01
Mean time-variant toll	€5.70	€5.59	€5.47	€5.36
Number of pay-lane users	4300	4211	4126	4043
Number of free-lane users	4934	4976	5020	5065
Average hours of queuing time on the free-lane	0.41	0.43	0.45	0.48
Average value of time on the free-lane	€7.73	€7.28	€6.85	€6.44
Average value of time on the pay-lane	€8.73	€8.98	€9.28	€10.44
Average hours of schedule delay early on the free-lane	0.82	0.83	0.83	0.84
Average hours of schedule delay early on the pay-lane	1.43	1.40	1.37	1.34

Suppose that the total number of users and α^* were the same for all uniform distributions. Then, if the spread of the NT value of time distribution were to increase, there would be *new*

types of drivers with α_i's exceeding the old highest value of time, and *new types* with α_i's below the old lowest value. The number of the *old type* of users, who had α_i's lying within the *old spread* (i.e. the *pre-spread-increase* spread), would have to decrease since there now are *new type* user and the number of users is constant.

A driver's congestion effect depends on the total number of users and the relative size of her value of time. Thus, the effect of an α^* driver on all *old type* users (i.e. those with α_i's that are in the *old pre-spread-increase* spread) remains the same. However, because there are fewer of these *old type* users, the total externality of an α^* driver on them is lower. In contrast, on the *new type* free-lane users, with the lowest values of time, the α^* driver's congestion effect is lower than her effect on the other free-lane users.

This implies that, if the number of users and α^* are constant, the externality of α^* users decreases with heterogeneity. This lessens the queuing cost decrease from a marginal subsidy increase. Yet, the excessive scheduling costs due to the subsidy are unaffected by the size of the spread. Accordingly, if the spread increases, the excessive scheduling effect of a marginal subsidy increase becomes larger than the queuing cost decrease effect. To get these marginal effects back in balance, the subsidy must be decreased. This increases the number of free-lane users, thereby raising free-lane congestion externalities and causing the two effects of the subsidy to be in balance again.

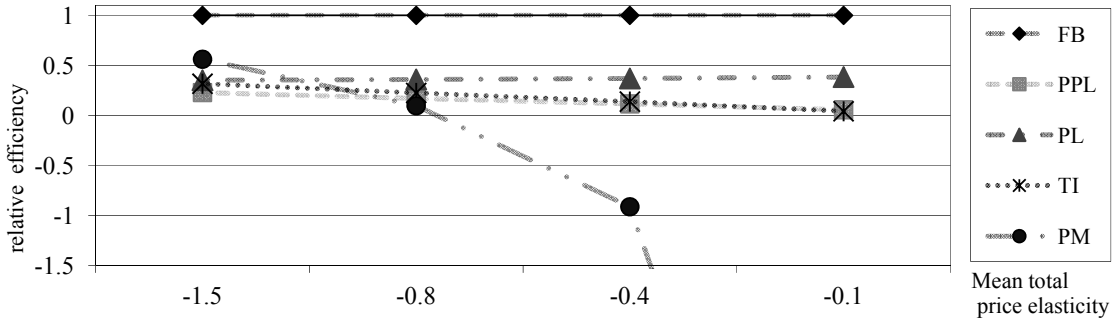
With more heterogeneity, it is thus optimal to allow more wasteful queuing on the free-lane. This, finally, is why, with more heterogeneity, the PL welfare gain is lower in absolute terms and relative to the FB welfare gain. Note that we do find, as Verhoef and Small (2004) did, that with more heterogeneity the average pay-lane user gains more because the mean value of time is higher. But this welfare enhancing effect is dominated by the detrimental effect of the longer optimal free-lane queuing.

Similarly, the relative efficiency of the PPL is lower with more heterogeneity because it is profit maximising to allow a smaller share of the drivers onto the pay-lane. This is achieved by increasing the time-invariant toll. Consequently, with more heterogeneity there is more queuing and scheduling delay on the PPL's free-lane, which lowers the relative efficiency.

9.5. Second-best policies with different price elasticities

The previous section found that a pay-lane achieves less of a welfare gain with more heterogeneity. In this section, Figure 15 looks at the relative efficiencies of the policies for four different mean total price elasticities in the NT equilibrium. The more inelastic demand is, the higher the PM toll and the larger the welfare loss. This is because with more inelastic demand, the monopolist has more market power. For the same reason, the PPL's welfare gain increases with the absolute elasticity of demand. Still, the change is less extreme, since the competition of the free-lane limits the pay-lane's market power. The TI toll's relative efficiency is higher with more elastic demand. As demand is more inelastic, the relative efficiency of the PL increases slightly. With more inelastic demand, consumer surplus *gain* from a PL is larger relative to the FB consumer surplus *loss*.

Figure 15: Relative efficiencies for different mean price elasticities in the NT situation



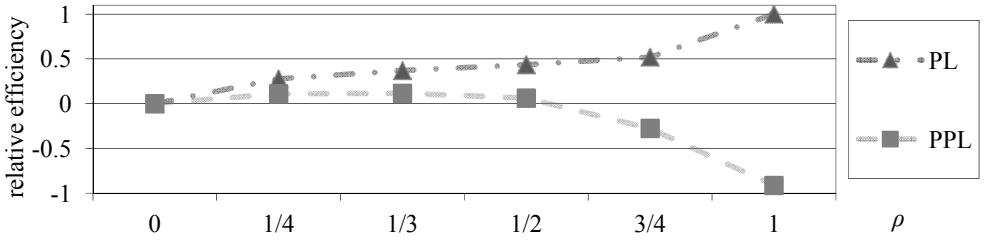
For the most elastic demand, welfare is higher in the monopolistic PM case than under any of the second-best policies. This may seem surprising, because usually a monopoly is bad for welfare. Still, it is consistent with the results from the static flow congestion models, which show that monopolistic tolling approaches FB tolling as demand becomes more elastic. The other results in this section on the effects of the elasticity of demand also correspond with those of the static flow models.

9.6. The pay-lane policies with different shares of total capacity

As a final sensitivity analysis, Figure 16 shows the relative efficiencies of the PL and PPL for different shares of capacity for the pay-lane. A ρ (share) of zero implies the NT case, and a ρ of one means either a PM or FB toll. The figure shows that welfare increases with the PL's capacity share. For the PPL, maximum welfare is attained around a share of a quarter to a third. For larger shares, the PPL's market power becomes too strong, and the negative effect of mark-up pricing dominates the positive effect of the reduction in queuing.

In Verhoef and Small (2004), the PPL is detrimental for welfare for all shares of capacity (at the base case price elasticity), and the relative efficiency decreases with the share. Using a homogeneous user bottleneck model, de Palma and Lindsey (2000) found a similar curve as ours for the PPL. Still, in their paper the PPL is welfare enhancing for larger shares of capacity than in our paper. With heterogeneity, time-variant tolling causes a consumer surplus loss, whereas with homogeneity it has no such effect. The welfare gain from tolling is lower with heterogeneity, and hence a PPL is only welfare improving for lower shares of capacity.

Figure 16: Relative efficiencies of the pay-lane policies with different shares of capacity



9.7. Concluding the sensitivity analysis

We may summarise this section's main findings as follows: firstly, the FB welfare gain decreases with heterogeneity, because the NT mean externality decreases. The welfare gain and relative efficiency of a public (PL) or private (PPL) pay-lane also decrease with heterogeneity. A private pay-lane can be welfare enhancing if its share of capacity is not too large. The monopolistic PM toll can be welfare increasing if the NT demand elasticity is a very elastic –1.5. However, it seems doubtful that travel demand is that elastic in reality.

10. Conclusion

This paper analysed the welfare effects of tolling in the bottleneck model with continuous heterogeneity in the value of time and price sensitive demand. In the no-toll equilibrium, the generalised price is largest for the highest values of time, and prices concavely increase with the value of time. If a public first-best (FB) toll is set, the queue is eliminated. Then, the price (excluding free-flow travel time) is the same for all values of time, and only slightly below the price faced by the highest value of time in the no-toll case. Therefore, the lower values of time face a large price increase, and their demand and consumer surplus decrease. Hence, with a first-best public toll, total demand, consumer surplus, toll revenue, and welfare are lower with heterogeneity in the value of time than without.

With a public pay-lane (PL), a time-variant toll is set to eliminate all queuing on the pay-lane. To this is added a negative time-invariant toll (subsidy) to attract extra users to the queue-free pay-lane. Due to this subsidy, total consumer surplus is higher under a public pay-lane than under the no-toll and first-best toll equilibria. On a road controlled by a (PM) private monopolist, queuing is also eliminated by time-variant tolling. A private pay-lane (PPL) operator uses the time-variant toll to eliminate the queue on its pay-lane. On top of the time-variant toll, a private operator sets a time-invariant toll that maximises total profit.

With increased heterogeneity in the value of time, the mean congestion externality is lower. Consequently, the average price in the no-toll equilibrium is lower. Hence, there is less to gain from queue-eliminating tolling, and the welfare gain of such a toll decreases. This is the case when the entire road is priced (i.e. the first-best public (FB) and profit maximising (PM) cases) and when only a part is priced (i.e. a pay-lane).

Probably the biggest surprise of our study is that the relative efficiency of the public pay-lane declines with heterogeneity in the value of time. This is opposite to the findings of previous studies using static flow congestion. The reason for the result is that, in our bottleneck model, it is welfare maximising to allow more queuing on the free-lane with more heterogeneity.

If only a time-invariant public (TI) toll can be set, welfare is substantially lower than with a time-variant toll. The welfare effect of this policy is hardly affected by heterogeneity in the value of time. Since the welfare gain of first-best tolling decreases with heterogeneity, the relative efficiency of the time-invariant toll increases. Finally, for a private pay-lane, the welfare gain and relative efficiency decrease with heterogeneity in the value of time; because, with more heterogeneity, it is profit maximising to allow more wasteful queuing and excessive schedule delay on the free-lane.

This paper makes the interesting suggestion that the effect of heterogeneity on the welfare effects of policies depends on the type of congestion (flow or bottleneck). If one ignores heterogeneity in the value of time, the welfare effects of policies may be biased. This bias is different for different policies and types of congestion. In reality, it seems likely that congestion consists of both bottleneck and flow congestion. It is therefore interesting to examine the effects heterogeneity has in a model that accounts for both types of congestion. A second valuable extension is the inclusion of heterogeneity in the value of schedule delay.

Acknowledgements

This paper is part of the project *Betrouwbaarheid van transportketens* of 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 thank Eric Kroes and Eva Gutiérrez-i-Puigarnau for their suggestions. Finally, we are grateful for the extensive and helpful comments of the referees.

Role of the funding source

This research was supported by Transumo. Transumo had no involvement in the research.

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., 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.

Brownstone, D., Ghosh, A., Golob, T., Kazimi, C., Van Amelsfort, D., 2003. Drivers' willingness-to-pay to reduce travel time: evidence from the San Diego I-15 congestion pricing project. *Transportation Research Part A* 37(4), 373-387.

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

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

de Palma, A., Lindsey, R, 2002, Comparison of morning and evening commutes in the Vickrey bottleneck model, *Transportation Research Record* 1807, 26–33

Evans, A.W., 1992. Road congestion pricing: When is it a good policy? *Journal of Transport Economics and Policy* 26(3), 213-244.

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.

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.

Odeck, J., Bråthen, S., 1997. On public attitudes toward implementation of toll roads: the case of Oslo toll ring. *Transport Policy* 4(2), 73-83.

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.

van den Berg, V., Kroes, E., Verhoef, E.T., 2010. De effecten van reiskostencompensatie op treinreizigers. Tijdschrift Vervoerswetenschappen 45(3), 102-110.

van den Berg, V., Verhoef, E.T., 2010. 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. *Tinbergen Institute discussion paper* 2010-016/3.

Verhoef, E.T., Small K.A., 2004. Product differentiation on roads: constrained congestion pricing with heterogeneous users. *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.

Xiao, F., Qian, Z., Zhang, H. M., 2009. The morning commute problem with coarse toll and nonidentical commuters. *University of California Davis working paper*.

Appendix A: Description of some of the used symbols

Table A1: Description of some of the used symbols

Symbol	Description
α_i	Value of time per hour of type i
α^*	Critical value of time in the pay-lane models (all types with $\alpha_i<\alpha^*$ travel on the free-lane)
β	Value per hour of schedule delay early (i.e. while arriving before t^*); this parameter is the same for all drivers
γ	Value per hour of schedule delay late (i.e. while arriving after t^*); this parameter is the same for all drivers
$\tau[t]$	Toll for arrival moment t ($\tau[t]= \tau_t[t]+\overline{\tau}$)
$\tau_t[t]$	Time-variant part of the toll
$\overline{\tau}$	Time-invariant part of the toll
B	The common to all users element of the slope of the demand function
$b_i[\alpha_i]$	Value of time specific element of the demand slope, it ranges between zero and one and integrates to one
$H[\alpha_i]$	Price for type i, in the NT equilibrium, as a fraction of the highest price faced by the highest-α-drivers

$K[\alpha_i]$	Price for type *i,* using the free-lane in a pay-lane equilibrium, as a fraction of the price of the α^* type
N	Total number of users
n_i	Number of users of type *i* (i.e. with a value of time of α_i)
P_i	Generalised price of the car trip for type *i* users
t	Arrival time at the destination
t^*	Preferred arrival time, which is the same for all and normalised to zero

Appendix B: Discrete heterogeneity and the no-toll equilibrium price

In equilibrium, the bottleneck operates at capacity throughout the peak. If it operated below capacity at any point in time, drivers could lower their costs by moving to that moment. The duration of the peak equals the number of users divided by capacity (i.e. $t_e–t_s=N_{NT}/s$). At t_s (start peak) and t_e (end peak) group M users arrive and face a zero queue length. Prices at t_s are $–\beta\cdot t_s$; at t_e they are $\gamma\cdot t_e$. These two prices are equal in equilibrium, and thus $t_s=–\eta\cdot t_e$. Inserting this equation for t_s into the peak duration formula and rewriting results in equation (B.1) for t_s, and (B.2) for t_e. Multiplying (B.1) by β or (B.2) by γ gives the NT price of type M users, as given in formula (B.3).

$$t_s = -\frac{\eta}{1+\eta}\frac{N}{s} = -\frac{\delta}{\beta}\frac{N}{s} \tag{B.1}$$

$$te = \frac{1}{1+\eta}\frac{N}{s} = \frac{\delta}{\gamma}\frac{N}{s} \tag{B.2}$$

$$P_{M} = \delta N/s \tag{B.3}$$

The relation $t_s=–\eta t_e$ between the moments that the first and last group M driver arrive is the same for all other groups. As Arnott et al. (1988) infer, this implies that for each type a share $\eta/(\eta+1)$ arrives before t^* and the remainder after. Accordingly, scheduling costs at t_{si}, which are $–\beta\cdot t_{si}$, equal those at t_{ei}, which are $\gamma\cdot t_{ei}=–\gamma\cdot t_{si}/\eta$. For a group *i*, the total number of arrivals by users with larger values of time before t_{si} or after t_{ei} is $\sum_{j=i+1}^{j=M} n_j$. The remaining users arrive between t_{si} and t_{ei}. In equilibrium, these users can just arrive between $t_{ei}–t_{si}$. Using this, and that t_{si} equals $–\eta\, t_{ei}$, equations (B.4) and (B.5) can be found for t_{si} and t_{ei}.

$$t_{si} = -\frac{\eta}{(1+\eta)s}\sum_{j=1}^{j=i} n_j = -\frac{\eta}{(1+\eta)s} F[\alpha_i]\, N \tag{B.4}$$

$$t_{ei} = \frac{1}{(1+\eta)s}\sum_{j=1}^{j=i} n_j = \frac{1}{(1+\eta)s} F[\alpha_i]\, N \tag{B.5}$$

The *isocost* curves of groups M and M–1 intersect at $t_{s(M-1)}$ and $t_{e(M-1)}$. For group M the generalised prices at $t_{s(M-1)}$ and t_s are the same. Therefore, group M's queuing costs at $t_{s(M-1)}$ equal the difference between the price at t_s and schedule delay costs at $t_{s(M-1)}$. From this, the queue length, as given in (B.6), can be calculated. Inserting (B.4) and (B.6) into the generalised price function gives the equilibrium price for group M–1 in (B.7).

$$T_D[t_{s(M-1)}] = q[t_{s(M-1)}]/s = \frac{\delta}{\alpha_M}\frac{n_M}{s} \tag{B.6}$$

$$P_{(M-1)} = -\beta\, t_{s(M-1)} + \alpha_{(M-1)}\, T_D[t_{s(M-1)}] = \frac{\delta}{s}\left((N - n_M) + \frac{\alpha_{M-1}}{\alpha_M} n_M \right) \tag{B.7}$$

The $(N–n_M)\cdot\delta/s$ in (B.7) gives the scheduling costs at $t_{s(M-1)}$ or $t_{e(M-1)}$. Queuing costs are $(n_m\cdot\alpha_{M-1}/\alpha_M)\cdot\delta/s$. At moments that group M–1 users also arrive, but that are closer to t^*, the price is the same. Yet, then scheduling costs are lower and queuing costs higher. Using the same procedure to derive prices for the other types is straightforward and hence not shown.

The equilibrium price for group M–1 is generalised for any i in (B.8). In the last part of (B.8), we insert the cumulative distribution function $F[\alpha_j]$ and density function $f[\alpha_j]$.

$$P_i = \frac{\delta}{s}\left(\sum_{j=1}^{j=i} n_j + \alpha i \sum_{j=i+1}^{j=M} \frac{1}{\alpha_j} n_j \right) = \frac{\delta N}{s}\left(F[\alpha_i] + \alpha_i \sum_{j=i+1}^{j=M} \frac{1}{\alpha_j} f[\alpha_j] \right) \tag{B.8}$$

Appendix C: Derivation of the pay-lane equilibrium

This appendix shows the derivation of the pay-lane equilibrium. Equating the difference between the pay-lane and NT prices to the change in inverse demand, and rewriting the result gives (C.1). This equation gives the number of type i users on each lane. n_{i1} is the number of type i pay-lane users, and n_{i2} the free-lane users. Persons with an α_i above the critical value of time (α^*) use the pay-lane, and those with an α_i above α^* use the free-lane. The $K[\alpha_i]$ function in (C.2) gives what fraction i's free-lane price is of the α^* drivers' price.

$$n_{ip} = \begin{cases} n_{i1} = n_{iNT} + \dfrac{\delta}{s\,B} b_i[\alpha_i]\, H[\alpha_i]\, N_{NT} - \dfrac{b_i[\alpha_i]}{B}\left(\dfrac{\delta}{s\,\rho} N_1 + \overline{\tau}_p \right) & \overline{\alpha} \ge \alpha_i \ge \alpha^* \\ n_{i2} = n_{iNT} + \dfrac{\delta}{s\,B} b_i[\alpha_i]\, H[\alpha_i]\, N_{NT} - \dfrac{b_i[\alpha_i]}{B}\left(\dfrac{\delta}{s(1-\rho)} K[\alpha_i]\; N_2 \right) & \underline{\alpha} \le \alpha_i \le \alpha^* \end{cases} \tag{C.1}$$

$$K[\alpha_i] = \left(\int_{\underline{\alpha}}^{\alpha_i} n_{j2}\, d\alpha_j + \alpha_i \int_{\alpha_i}^{\alpha^*} (n_{j2}/\alpha_j) d\alpha_j \right) / N_2 \tag{C.2}$$

Section 6 integrated the n_{iFB} over all values of α_i to find the total number of users. Now we integrate n_{i2} from $\underline{\alpha}$ to α^* to find N_2, and n_{i1} from α^* to $\overline{\alpha}$ for N_1. Still, there seems to be no closed-form solution. For some α^* and time-invariant toll, we integrate the upper part of (C.1) and rewrite the result to find N_1 in (C.3). Inserting (C.3) into the upper formula for n_{i1} of equation (C.1) results in (C.4), which gives the number of type i users on the pay-lane.

$$N_1\left[\alpha^*, \overline{\tau}_p\right] = \frac{\left(N_{NT}\left(\int_{\alpha^*}^{\overline{\alpha}} b_i[\alpha_i]\, d\alpha_i + \dfrac{\delta}{s\,B} \int_{\alpha^*}^{\overline{\alpha}} H[\alpha_i]\, b_i[\alpha_i]\, d\alpha_i \right) - \int_{\alpha^*}^{\overline{\alpha}} b_i[\alpha_i]\, d\alpha_i \dfrac{\overline{\tau}_p}{B} \right)}{1 + \int_{\alpha^*}^{\overline{\alpha}} b_i[\alpha_i]\, d\alpha_i \cdot \delta \Big/ (s \cdot \rho \cdot B)} \tag{C.3}$$

$$n_{i1}\left[\alpha^*, \overline{\tau}_p\right] = N_{NT} f[\alpha_i]\left(1 + \frac{\delta}{s\,B} H[\alpha_i] \right) - \frac{f[\alpha_i]}{B}\left(\frac{\delta}{s\,\rho} N_1\left[\alpha^*, \overline{\tau}_p\right] + \overline{\tau}_p \right) \quad \overline{\alpha} \ge \alpha_i \ge \alpha^* \tag{C.4}$$

For (C.5) we integrated the formula for n_{i2} from $\overline{\alpha}$ to α^*, and rewrote the result, to get the formula for N_2. Then, we inserted (C.5) back into the lower formula of (C.1) for n_{i2} to find the solution to the number of type i free-lane users conditional on the α^* and value of time distribution of the free-lane users in the pay-lane equilibrium.

$$N_2[N_{NT}, \alpha^*] = N_{NT} \frac{\left(\int_{\underline{\alpha}}^{\alpha^*} b_i[\alpha_i]\, d\alpha_i + \dfrac{\delta}{sB}\left(\int_{\underline{\alpha}}^{\alpha^*} H[\alpha_i]\, b_i[\alpha_i]\, d\alpha_i \right) \right)}{1 + \dfrac{\delta}{s\,(1-\rho)\,B} \int_{\underline{\alpha}}^{\alpha^*} K[\alpha_i]\, b_i[\alpha_i]\, d\alpha_i} \tag{C.5}$$

Appendix D: Numerical solutions for pay-lane and time-invariant toll

The numerical solution for the pay-lane starts with an initial α^* and time-invariant toll. Then, we calculate the prices and value of time density function on the pay-lane. Using this α^* and

some starting distribution of values of time on the free-lane,[8] we calculate the resulting free-lane prices. These are then equated to the inverse demands, and from this we derive the total number of free-lane users and new distribution of n_{i2}. This new distribution is, for now, not equal to the starting distribution, since prices and inverse demands are not in equilibrium.

This new distribution is then approximated by a cubic spline. This means that we split up the distribution into 71 sections and in each section fit a cubic polynomial. This spline distribution is used as the next iteration's starting distribution. This procedure is repeated until convergence, which is defined as a maximum absolute percentage difference of 10^{-6}% between the spline and the resulting demand.[9]

This outcome is conditional on the assumed α^*. In practise, α^* is not exact, implying that prices for α^* users are not the same on the pay-lane as on the free-lane. Therefore, a new α^* is sought for which the two prices are closer. Then, the above procedure is repeated, using the earlier iteration's values of time distribution as the starting distribution. This second procedure is repeated until convergence, which is a maximum absolute percentage difference of 10^{-5}% between the pay-lane and free-lane price for α^*. To find the optimal time-invariant toll, we used a grid search style method. This search was helped by the fact that welfare and profits seem globally concave for all values of the time-invariant toll that were tried. The minimum accuracy of the toll was a tenth of a Euro cent.

For the time-invariant toll we basically use the same numerical procedure as for the pay-lane. In the TI case there is not the difficulty of finding α^*. Accordingly, the TI case procedure is easier than the pay-lane's procedure. For this model, we use a cubic spline with 82 sections. We use more sections for the TI policy than for the pay-lane's because the distribution is approximated over a larger area: from $\underline{\alpha}$ to $\overline{\alpha}$ instead of from $\underline{\alpha}$ to α^*. The optimal toll is again found using a grid search style method.

[8] We used the NT equilibrium distribution. We also tried other starting distributions, and the results do not seem to depend on the starting distribution.

[9] The approximation by a spline is, in principle, not necessary, since it is possible to use the n_{i2} from iteration j as the starting distribution in iteration $j+1$. This results, however, in an increasingly complex starting distribution (e.g. for iteration number j the starting distribution has a $\log(\alpha_i)^{j-1}$ in it). This in turn causes the computation time per iteration to exponentially increase. Moreover, without the approximation the numerical integrations do not converge for a high number iteration, causing the program to crash.