WINNING OR LOSING FROM DYNAMIC BOTTLENECK CONGESTION PRICING?[#,##]

The Distributional Effects of Road Pricing with Heterogeneity in Values of Time and Schedule Delay

Vincent A.C. van den Berg[*]
Department of Spatial Economics
VU University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
+31-20-598 6160
vberg@feweb.vu.nl

Erik T Verhoef[**]
Department of Spatial Economics
VU University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
+31-20-598 6090
everhoef@feweb.vu.nl

Key words: Traffic congestion; Road pricing; Heterogeneity; Distributional impacts; Bottleneck model
JEL codes: D62, H23, R41, R48,

Abstract

This paper analyses the efficiency and distributional impacts of congestion pricing in Vickrey's (1969) dynamic bottleneck model of congestion, allowing for continuous distributions of values of time and schedule delay. We find that congestion pricing can leave a majority of travellers better off even without returning the toll revenues to them. We also find that the consumer surplus losses or gains from tolling are not strictly monotonic in the value of time, because they also depend on the value of schedule delays. 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 for that value of schedule delays. For second-best pricing with an untolled alternative, the pattern of distributional effects is quite similar to that for first-best pricing. In contrast with results from prior static models, users who are indifferent between the two alternative routes are not the ones who gain least from this type of second-best pricing. Our results suggest that, in assessing the distributional impacts of road congestion pricing, it is important to take into account both the distribution of the value of time and of the value of schedule delays, as well as the dynamics of departure time choice.

[#] Financial support from TRANSUMO and ERC (AdG Grant #246969 OPTION) is gratefully acknowledged. We thank the editor and two anonymous reviewers for their extensive and helpful comments. We also thank Vincent van der Goes and Eva Gutiérrez-i-Puigarnau for their help. The usual disclaimer applies.

##This is a post-print of the article published in Journal of Public Economics, 95(7–8), 983–992: for the published version see http://dx.doi.org/10.1016/j.jpubeco.2010.12.003 , the journal website is http://www.journals.elsevier.com/journal-of-public-economics/

[*] Corresponding author.
[**] Affiliated to the Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam.

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) was first to recognise that traffic congestion entails an external cost, and that efficiency requires a toll equal to the marginal external congestion cost. A large literature on the economics of road pricing has emerged since then (see for example Small and Verhoef, 2007, for a recent review). A seminal contribution was made by Vickrey (1969). He developed what has become the workhorse dynamic economic model of traffic congestion, in which congestion takes the form of queuing at a bottleneck, and departure time decisions are endogenous.

Despite the strong economic case for road congestion pricing, practical applications remain scarce, although growing in number with schemes implemented for example in Singapore, London, and Stockholm. 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 the value of time and income, this is often taken as implying that the poor lose and the rich gain (*e.g.* Layard, 1977). Still, even if higher incomes have higher values of time on average, they also have higher car-ownership and longer commutes. Hence, they may lose less per kilometre, but their position may worsen as they drive more. The net balance of these opposing forces is an empirical matter. Foster (1974, 1975) for example suggested that road pricing may be progressive, while Richardson (1974) maintained that regressiveness is more likely.

Distributional impacts of transport policies have been studied not only in the context of road congestion pricing. To name a few recent examples, West (2004) studies pricing of vehicle pollution; Mayeres and Proost (2001) investigate how the attractiveness of various road and public transport policies vary with the degree of inequality aversion in a general equilibrium framework; Parry (2002) compares alternative congestion policies for heterogeneous users in a multi-modal (road *versus* rail) setting; and Cain and Jones (2008) focus on how differences in car ownership affect the distributional impacts of road pricing.

But the welfare effects of congestion pricing have received continuing interest, with network aspects being one of the emerging themes. Small and Yan (2001), for instance, develop a static model with two groups, distinguished by value of time, on a small network with only two parallel links. One of the policies they consider concerns so-called pay-lanes, or express lanes, where toll-free capacity is offered in parallel with priced but less congested lanes. They find that accounting for heterogeneity of values of time improves the efficiency of this second-best scheme, as the implied product differentiation better fits the preferences of the two groups. Verhoef and Small (2004) consider a comparable network, but use a continuous distribution of values of time. The impacts of first-best pricing are as expected: drivers' losses decrease, or gains increase, with the value of time. However, the distributive impacts of second-best pricing are not monotonous in the value of time. Instead, the largest losses are incurred by users with an intermediate value of time, namely those who are indifferent between the tolled and untolled link. Drivers with a lower value of time use the untolled link, and suffer less from the increased travel time as the value of time is lower. Drivers with a higher value of time use the tolled link, and gain more from the reduced travel time, the higher their value of time is. The implied non-monotonous pattern of gains and

losses, with its minimum at an intermediate value of time, is of great importance for understanding patterns of support for and opposition against second-best road pricing in practice. Clearly, such patterns are best detected with a continuous distribution of the value of time, and therefore imply a warning for the use of simplified models with only a few or just two classes of travellers.

Empirical evidence, for example as provided by Small, Winston and Yan (2005), confirms that values of time differ substantially over drivers, 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, starting with Vickrey's (1969) bottleneck model mentioned above, emphasise the importance of 'schedule delay costs' as another congestion cost. These are costs associated with an arrival at a less-than-most-desired moment. As is true for values of time, values of schedule delay are also likely to differ across individuals, and for an individual between trip purposes and days. The question that we are interested in is how such heterogeneity in values of schedule delay, combined with heterogeneity in the value of time, affects the distributional impacts of first-best and second-best dynamic pricing on a small network, consisting of two parallel bottlenecks.

We are not the first to study heterogeneity in the bottleneck model. For example, Vickrey (1973) considers a case to be discussed in more detail below, where the values of time and schedule delays vary proportionally: *i.e.* the ratios of the cost parameters are identical across drivers. With fixed demand, all drivers are better off with first-best tolling than without tolling—except the users with very lowest values, who are unaffected. De Palma and Lindsey (2002) study a different case, which is also considered further later on, where the heterogeneity is in the value of time, and the values of schedule delay are identical. With fixed demand, all users lose due to first-best tolling—except the users with the very highest value of time, who are unaffected. In both these papers, the distributional effects are monotonic. Van den Berg and Verhoef (2011) introduce price sensitive demand to the model of de Palma and Lindsey. They find that the share of users who lose due to first-best or pay-lane tolling increases with the amount of heterogeneity in the value of time. Arnott, de Palma and Lindsey (1994) consider two types of users with different values of time and schedule delay and fixed demand. First-best tolling raises prices for low values of time. With a greater dispersion in values of time, this price increase is larger. With sufficient heterogeneity, users with the lower value of time may remain worse off even after toll revenues are recycled in equal portions. Lindsey (2004) considers the existence and uniqueness of equilibrium in the bottleneck model for an arbitrary number of groups, with discrete distributions of values of time and schedule delay. Finally, Xiao, Qian, and Zhang (2010) consider step tolls in the model of Vickrey (1973).

This paper is, to the best of our knowledge, the first to consider bivariate continuous distributions of values of time and schedule delay, combined with price-sensitive demand. This setting allows for a rather rich analysis of the distributive impacts of congestion pricing.

In contrast with Verhoef and Small (2004) and de Palma and Lindsey (2002), we find that consumer surplus losses or gains from first-best tolling are not strictly monotonic in the value of time, since the value of schedule delays also determines such gains or losses. As a result, some drivers with a certain value of time may incur a greater loss than drivers with a

lower value of time. With second-best pricing with an untolled alternative, the gains or losses are also not monotonic in the value of time or in the value of schedule delays. But again unlike in the static model of Verhoef and Small (2004), it is not true that drivers who are indifferent between the two routes incur the greatest losses. These results underline that it is important to consider both types of heterogeneity in analysing the distributional effects of congestion pricing. We also find that congestion pricing can leave a majority of users better off even without returning the toll revenues to them. This is a striking contrast with analyses that only consider heterogeneity in the value of time, where (almost) everyone is worse off.

Although we focus on road congestion, the insights we obtain are likely to be relevant for a wider class of congestible facilities, in particular those where both the quality of service and the moment of use are important determinants of the consumers' utility. Obvious examples include transport infrastructures such as airports and railway systems, but also non-transport facilities such as telecommunication and computer networks, theatres, and recreational facilities.

The paper is organised as follows. Section 2 introduces the basic model, and discusses the main findings from the prior literature using examples with two discrete groups. The discussion emphasises the results that will help interpreting the results from our more sophisticated model with continuous bivariate heterogeneity. Section 3 presents the model with continuous heterogeneity and discusses its no-toll equilibrium. Next, Section 4 considers first-best pricing, while Section 5 analyses second-best pricing with an untolled alternative. Section 6 presents a sensitivity analysis in which we vary the shapes of the distributions. Section 7 concludes.

2. The model

2.1. The basics

There are of course various ways of modelling the dynamics of traffic congestion; see for example Small and Verhoef (2007) for a review. Most economic analyses follow the work of Vickrey (1969) and Arnott, de Palma and Lindsey (1990, 1993), and use some variant of Vickrey's bottleneck model. In that model, traffic congestion takes the form of queuing behind a bottleneck of finite capacity. Vickrey considers 'pure' bottleneck congestion, meaning that without a queue and as long as the arrival rate of users at the bottleneck is below its capacity, no travel delays are incurred. Otherwise, the queuing delay is given by the length (in vehicles) of the queue at the moment of joining it, divided by the capacity of the bottleneck. While thus capturing what is probably the most important aspect of dynamic traffic congestion in practice, namely queuing; the absence of an explicit spatial dimension in the model's description of traffic, and the disregarding of travel delays when flows are below capacity, greatly simplifies its analytics.

In this paper we use the bottleneck model as well. The bottleneck is located on a road connecting a single origin and destination. For convenience, we set the free-flow travel time at zero. Without a queue, the departure ('from home') occurs at the same moment as the passage of the bottleneck and the arrival ('at work'). The bottleneck has a finite capacity, s. When there is no queue when the driver reaches the bottleneck, vehicles freely pass the bottleneck and the arrival rate at work r_A equals r_D provided r_D does not exceed s. In all

other cases, a queue of length Q grows or shrinks at a rate $\dot{Q} \equiv r_D - s$. The travel time $T[t]$ for someone who arrives at time t then depends on the length of the queue at the moment of departure t_D in the following way (note that we use square brackets to denote arguments of functions):

$$T[t] = \frac{Q[t_D]}{s}, \quad \text{with} \quad t_D = t - T[t]. \tag{1}$$

As is customary in this literature, we follow Small's (1982) model of scheduling behaviour, and assume that an individual with characteristics i faces two congestion related costs. One is travel delay cost, defined as the product of travel time and the individual's value of time, α_i. The other is schedule delay cost, which is found by multiplying the absolute difference between the actual arrival time, t, and the desired arrival time, t_i^*, by the value of schedule delay early (β_i) or late (γ_i), depending on whether she arrives before or after t_i^*.[1] The sum of these is the 'generalised cost', c. A driver i's trip price for an arrival at t is then:

$$p_i[t] \equiv c_i[t] + \tau[t] = \alpha_i \cdot T[t] + \tau[t] + \begin{cases} \beta_i \cdot \left(t_i^* - t\right) & \text{if } t \le t_i^* \\ \gamma_i \cdot \left(t - t_i^*\right) & \text{if } t > t_i^* \end{cases}, \tag{2}$$

where $\tau[\cdot]$ denotes a toll that is defined in terms of the arrival time (*i.e.* the moment of passing the bottleneck).

2.2. Homogeneous users

To better appreciate our results for heterogeneous users and price-sensitive demand, we first discuss the basic bottleneck model, in which all utility parameters are equal across individuals, and demand is fixed and denoted by N. We will keep the discussion concise; the model is discussed in greater depth by, for example, Arnott, de Palma and Lindsey (1990, 1993). In the equilibrium without tolling, the queue grows linearly over time for early arrivals, and shrinks linearly for late ones, in such a way that the sum of travel delay and schedule delay costs is constant over time. This ensures that the dynamic equilibrium condition that the price be constant throughout the peak is fulfilled.

The left panel in Figure 1 depicts this equilibrium. The indicated equilibrium slopes of $T[t]$ can be verified by equating the time-derivative of equation (2) to zero. Because the drivers at the start (t_s) and end (t_e) of the peak only incur schedule delay cost while their prices are equal, and given that the duration of the peak is N/s, the equilibrium generalised price can be easily determined:[2]

---

[1] The value of time is the marginal utility of travel time savings divided by the marginal utility of income; values of schedule delay early and late are defined analogously.

[2] Schedule delay costs at t_s and t_e must be equal, so that $\beta \cdot (t^* - t_s) = \gamma \cdot (t_e - t^*)$. A fraction $\gamma/(\beta+\gamma)$ of the travellers will therefore arrive early, and a fraction $\beta/(\beta+\gamma)$ will arrive late. Multiplying these fractions with the total duration of the peak, N/s, gives the schedule delays for the first and last travellers, respectively. Multiplying these with the appropriate shadow prices β and γ gives for both travellers the generalised price in (3).

$$p = c = \delta \cdot \frac{N}{s} \quad \text{with} \quad \delta = \frac{\beta \cdot \gamma}{\beta + \gamma}, \tag{3}$$

The two intersections with the horizontal axis are indicative of the generalised price: *i.e.* $p = \beta \cdot (t^* - t_s) = \gamma \cdot (t_e - t^*)$. This will be useful when identifying changes in generalised prices in later diagrams. Also note that the equilibrium pattern $T[t]$ can be interpreted as an iso-price function; a higher position would correspond with a higher generalised price.

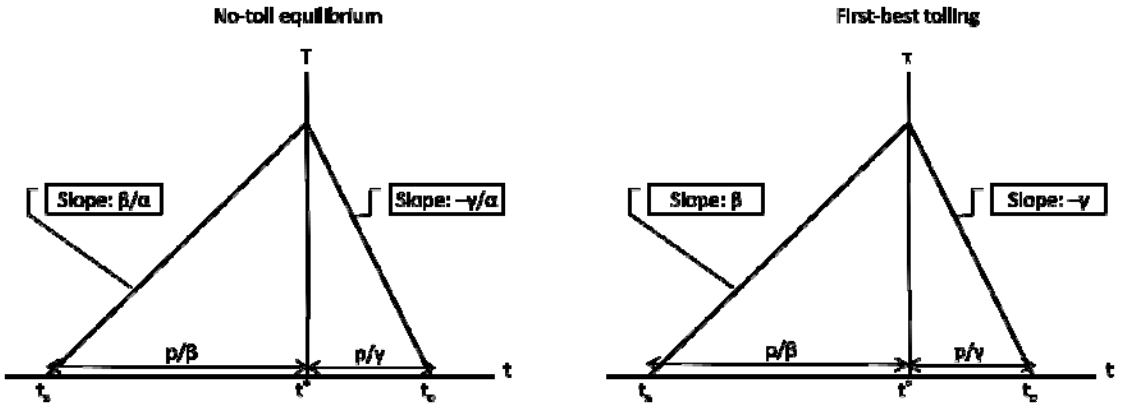

Figure 1. No-toll equilibrium (left panel) and optimum (right panel) with homogeneous drivers

This no-toll equilibrium requires a departure rate $r_D = s \cdot \alpha / (\alpha - \beta)$ for early arrivals, and $r_D = s \cdot \alpha / (\alpha + \gamma)$ for late ones. The former expression indicates that, for this equilibrium to exist with finite positive departure rates, $\alpha > \beta > 0$ is required. This requirement is consistent with the plausible assumption that, when arriving early, a driver prefers getting out of the car over extending the trip by making a detour. For heterogeneous users, similar requirements apply. Defining the ratio of value of time to value of schedule delay late as $\mu_i \equiv \alpha_i / \beta_i$, this means that we assume $\mu_i > 1$ for all i when turning to heterogeneous drivers.

The right panel of Figure 1 depicts the optimum of the basic homogeneous-user model. Queuing delay is a pure loss in the model, as it can be reduced without increasing schedule delay costs as long as the queue has a positive length. It is therefore optimal to fully eliminate all queuing, while keeping outflow at capacity to avoid unnecessary schedule delay costs. This requires $r_D = s$ throughout the peak, which can be achieved by setting the toll at each instant equal to what the value of queuing delay was in the no-toll equilibrium. Multiplying $T[t]$ in the left panel by α thus gives the optimal toll $\tau[t]$ depicted in the right panel. The sum of tolls and schedule delay costs is then constant over time as long as the bottleneck is used, so that in dynamic equilibrium also the travel delay cost will be constant. Given that the initial queue length is zero, this means that it will remain zero throughout the peak, as required for the optimum.

Because the bottleneck still operates at full capacity throughout the peak, t_s and t_e will not shift due to tolling, and the generalised price will not change, remaining equal to the value

in (3). But because travel delays are fully replaced by toll payments, the total, average, and marginal costs are half as high as in the no-toll equilibrium. The potential importance of considering the schedule delay costs in analysing the welfare effects of congestion pricing, as we do in this paper, is therefore highlighted by the fact that, in the no-toll equilibrium of the basic model, aggregate schedule delay costs are (exactly) 50% of total congestion costs—and a full 100% in the first-best optimum.

2.3. Introducing heterogeneity: two examples with discrete groups

Vickrey (1973) was also the first to introduce heterogeneity in the bottleneck model. He considered the case where α, β, and γ vary proportionally over travellers, so that they all have the same ratios $\eta \equiv \gamma / \beta$, $\mu \equiv \alpha / \beta$, and α / γ. Because each of the values α, β, and γ is the ratio of some marginal utility over the marginal utility of income, this type of heterogeneity could result from income differences between otherwise identical individuals.

Figure 2 shows the resulting equilibrium and optimum if we restrict the number of groups to two. In the no-toll equilibrium, the travel time evolves in the same triangular fashion as in Figure 1, because the ratios α / β and α / γ are identical across groups. The two groups therefore travel jointly, not separately, in time.

Also for heterogeneous drivers, queuing delay is a pure loss, and it remains optimal to fully eliminate it. The first-best toll schedule therefore has, at a particular moment, a slope equal to β or $-\gamma$ for the group that is arriving at that time. This means that drivers will be separated temporally: travellers with high values of $\{\alpha,\beta,\gamma\}$, or group H in short, will arrive close to t^*. The earlier interpretation of the toll schedule as an iso-price line now helps explaining why separation occurs. For drivers from both groups L and H, an arrival in the other group's interval would imply ending up on an iso-price line that corresponds with a price above the equilibrium level.

The toll schedule thus effectively separates the travellers. A consequence is that, besides eliminating travel delays, the toll reduces aggregate schedule delay cost. This benefit is reflected in the price reduction $\Delta p/\beta$ in the right panel for group H, while group L has an unchanged generalised price.

It is not hard to verify geometrically that the following prices apply, where subscripts denote groups and superscripts the regimes no-toll (NT) versus first-best (FB), and $\delta_i \equiv \beta_i \cdot \gamma_i / (\beta_i + \gamma_i)$:

$$p_L^{NT} = \delta_L \cdot \frac{N_L + N_H}{s}, \tag{4a}$$

$$p_H^{NT} = \delta_H \cdot \frac{N_L + N_H}{s}, \tag{4b}$$

$$p_L^{FB} = \delta_L \cdot \frac{N_L + N_H}{s}, \tag{4c}$$

$$p_H^{FB} = \delta_H \cdot \frac{\frac{\beta_L}{\beta_H} \cdot N_L + N_H}{s}. \tag{4d}$$

The high-value-of-time group *H* thus benefits from the imposition of first-best tolling, and does so because it benefits from the existence of heterogeneity in the first-best optimum. This happens because group *L* has a lower slope of the optimal toll schedule, so that an additional type-*L* traveller adds less to groups *H*'s generalised price than an additional type-*H* traveller, the ratio of marginal price effects being $(\partial p_H^{FB} / \partial N_L) / (\partial p_H^{FB} / \partial N_H) = \beta_L / \beta_H$. Vickrey showed that, for larger numbers of groups, all groups except those with the lowest values of $\{\alpha,\beta,\gamma\}$ benefit from the imposition of optimal tolling, while the gains are highest for the users with the highest values. The explanation is the same as for the two-groups example of Figure 2.

Figure 2. No-toll equilibrium (left panel) and optimum (right panel) when α, β, and γ vary but in fixed proportions

A second example of two-groups heterogeneity concerns the heterogeneity of de Palma and Lindsey (2002) and Van den Berg and Verhoef (2011). Now the scheduling coefficients β and γ are fixed and equal across users, so that also $\eta \equiv \gamma / \beta$ is again the same for the two groups, while α and hence $\mu \equiv \alpha / \beta$ vary across users. This ratio μ has a behavioural interpretation: it reflects the willingness to accept greater schedule delays in order to reduce travel time. Heterogeneity in μ could therefore result from differences in comfort level or possibilities to use in-vehicle time productively, but also from differences in the tightness of scheduling constraints.

Figure 3 shows the equilibrium and optimum, again for two groups. Because the ratios β/α and γ/α differ, drivers are now separated temporally in the no-toll equilibrium. Drivers who have lower α and μ are more "willing to queue", and thus arrive closer to t^*. These drivers, group *L* in this example, are now the ones who benefit from heterogeneity: group *H* users build up the queue length more slowly than group *L* users, and this brings group *L* on a lower iso-price line. Therefore, group H imposes lower marginal external costs. Group *L*, however, loses this benefit when tolling is introduced. This loss amounts to $\Delta p/\beta$, as identified in the left panel. The other group *H* has an unchanged generalised price. The loss from tolling is therefore greater for drivers with a lower value of time. The generalised prices in this example can again be verified geometrically in the diagram, and amount to:

$$p_L^{NT} = \delta_L \cdot \frac{N_L + \frac{\mu_L}{\mu_H} \cdot N_H}{s} = \delta_L \cdot \frac{N_L + \frac{\alpha_L}{\alpha_H} \cdot N_H}{s}, \tag{5a}$$

$$p_H^{NT} = \delta_H \cdot \frac{N_L + N_H}{s}, \tag{5b}$$

$$p_L^{FB} = \delta_L \cdot \frac{N_L + N_H}{s}, \tag{5c}$$

$$p_H^{FB} = \delta_H \cdot \frac{N_L + N_H}{s}. \tag{5d}$$

Note that, in fact, $\delta_L = \delta_H$ under the present assumptions; we maintain the subscripts for ease of later reference. With more groups than two, drivers with the highest α are still equally well off due to tolling. All other groups lose, and more so if α is smaller.

Figure 3. No-toll equilibrium (left panel) and optimum (right panel) when β and γ are fixed and α (and hence μ) varies

The similarity between these two examples[3] is that tolling is less harmful, or more beneficial, to the drivers with the higher α. The main difference is that in the first example, one group is equally well off and all other groups gain from the imposition of tolling, while in the second example, one group is equally well off and all other groups lose. Because our model has both types of heterogeneity—both in μ and in β—we may anticipate that some users in our model will be better off due to tolling, and some users worse off. More precisely, we may expect that the proportion of individuals who gain depends on the relative degrees of heterogeneity in β (a positive effect) and μ (a negative effect). The latter expectation is consistent with the

[3] A referee suggested mentioning a third example of heterogeneity with two groups, namely the case where drivers differ with respect to β but have equal α (again keeping η equal). It is not hard to show geometrically in diagrams such as Figures 2 and 3 that the two groups will be separated temporally in both the no-toll equilibrium and the optimum, with the drivers with higher β arriving closer to t^*. Neither group gains or loses from tolling. We thank the reviewer for this suggestion.

results of Van den Berg and Verhoef (2011), who find that most users are worse off due to tolling when there is only heterogeneity in μ.

2.4. *Generalising the model: simultaneous heterogeneity in β and μ for an arbitrary number of discrete groups*

It is now relatively straightforward to derive the generalised price for the more general case with *I* groups. For convenience, we assume that each group has a different ratio $\mu_i \equiv \alpha_i / \beta_i$, and different values of β_i and γ_i, so that all groups travel separated in time in the no-toll equilibrium as well as in the optimum. We impose a common ratio $\eta_i \equiv \gamma_i / \beta_i = \eta$ for all groups. We maintain this latter assumption throughout this paper. This assumption is not essential for our results, but it helps in restricting heterogeneity to two dimensions, namely μ and β. The assumption means that the timing of the overall arrival period is independent of the degree of heterogeneity. The desired arrival times are equal for all users, and are normalised such that $t_i^* = t^* = 0$.

To characterise the no-toll equilibrium, we define indices *j* such that μ_j increases with *j*; group *J* (with *J*=*I*) has the highest value. As explained in Figure 3, drivers arrive ordered by μ, with the lowest values arriving closest to t^*. Consistent with equations (4a,b) and (5a,b), the following generalised price can be derived for group *i*:

$$p_i^{NT} = \frac{\delta_i}{s} \cdot \left(\sum_{j=1}^{i} N_j + \sum_{j=i+1}^{J} N_j \cdot \frac{\mu_i}{\mu_j} \right) \quad \forall i = 1,...,J . \tag{6}$$

This price increases with δ_i and with μ_i. A higher μ_i implies (for a given β_i and total demand) an arrival further from t^* and hence a smaller advantage from flatter iso-cost curves for other groups. For a single group, equation (6) naturally simplifies to equation (3). For group *J*, with the highest μ_i, the price would be the same if all other users had the same characteristics as group *J* itself. All other groups benefit from heterogeneity. In Figure 3, the argument is that these groups benefit from the flatter iso-cost curves for groups who arrive closer to the edges of the peak. Equation (6), accordingly, implies that users with a higher μ_i impose lower marginal external costs (total cost are $\sum_i N_i \cdot p_i^{NT}$).

In the first-best optimum, the arrival order changes, unless there would be perfect rank correlation between μ_i and β_i. It is therefore convenient to introduce a new index *k* such that β_k increases with *k*. The highest value is *K* (with *K*=*I*). Generalising the logic of Figures 2 and 3, we find:

$$p_i^{FB} = \frac{\delta_i}{s} \cdot \left(\sum_{k=1}^{i-1} N_k \cdot \frac{\beta_k}{\beta_i} + \sum_{k=i}^{K} N_k \right) \quad \forall i = 1,...,K . \tag{7}$$

With homogeneous users, equation (7) again simplifies to $p = \delta \cdot N / s$, confirming the equality of NT and FB generalised prices in the basic bottleneck model (Arnott, de Palma and Lindsey, 1993). In the general case, it is group 1, with the lowest β_i, for whom the

generalised price is the same with and without heterogeneity. All other groups benefit from heterogeneity, even though the FB generalised price increases with β_i.

Whether, with *I* groups, a certain group benefits or loses due to optimal tolling depends on its position in both orderings and the distributions of μ and β. We study this question in more detail below, but already note that there is no *a priori* reason to believe on the basis of equations (6) and (7) that the differences in prices $p_i^{NT} - p_i^{FB}$ are perfectly correlated with α_i. This further justifies using a model with continuous distributions of time value coefficients, to gain deeper insight into the distributional effects of congestion pricing.

3. The full (numerical) model and its no-toll equilibrium

Our numerical model generalises the above examples in a number of respects. This section presents that model and its base-case parameterisation. We note beforehand that the sensitivity analysis in Section 6 shows that the results derived from the base-case model are fairly robust. A first difference with the examples above is that demand is now price-sensitive. Second, rather than considering discrete groups, the model uses continuous distributions of β and μ. Because of the bivariate distribution, we sometimes need to use two indices to characterise groups *i*, in which case the composite *yz* denotes drivers with $\mu = \mu_y$ and $\beta = \beta_z$. When a single index does not cause ambiguity, we stick to the index *i* to denote a type of drivers with their specific combination of μ and β.

Figure 4 shows the density function of μ and β for the no-toll equilibrium. We use a symmetric triangular distribution, that is defined by its minima $\underline{\beta} = 2$ and $\underline{\mu} = 1.01$, and maxima $\overline{\beta} = 8$ and $\overline{\mu} = 3.01$. When units are dollars or Euros, the mean value of time of 10.05 indicates that our values are reasonable compared to empirical estimates (*e.g.* Small and Verhoef, 2007). We set $\eta \equiv \gamma_i / \beta_i = 3.9$, as in Arnott, de Palma and Lindsey (1990). The triangular shape, although perhaps intuitive, is otherwise not based on empirical evidence.

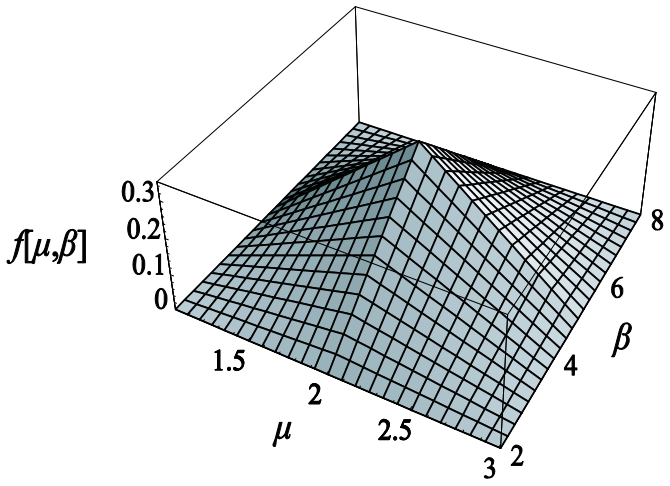

Figure 4. The numerical distribution of β and $\mu \equiv \alpha / \beta$ in the no-toll (NT) equilibrium

For every i, a linear inverse demand applies. We impose regularity on the distributions of intercepts and slopes of the inverse demands by using the following specification:

$$D_i[n_i] = d_i^0 - d_i^1 \cdot n_i \quad \text{with} \begin{cases} d_i^0 = A + A_i[\mu_i, \beta_i] \\ d_i^1 = \dfrac{B}{b_i[\mu_i, \beta_i]} \end{cases}, \tag{8}$$

where n_i denotes the density of users of type i. The specifications of d_i^0 and d_i^1 are driven by the calibration objectives of obtaining an equilibrium density function as shown in Figure 4; an equilibrium demand of 9000 given a capacity of 3600 vehicles per hour (so that the peak lasts 2.5 hours); and an average demand elasticity of –0.4.[4,5] In other words, the dependence of the intercepts and slopes on μ and β was introduced only for calibration purposes.

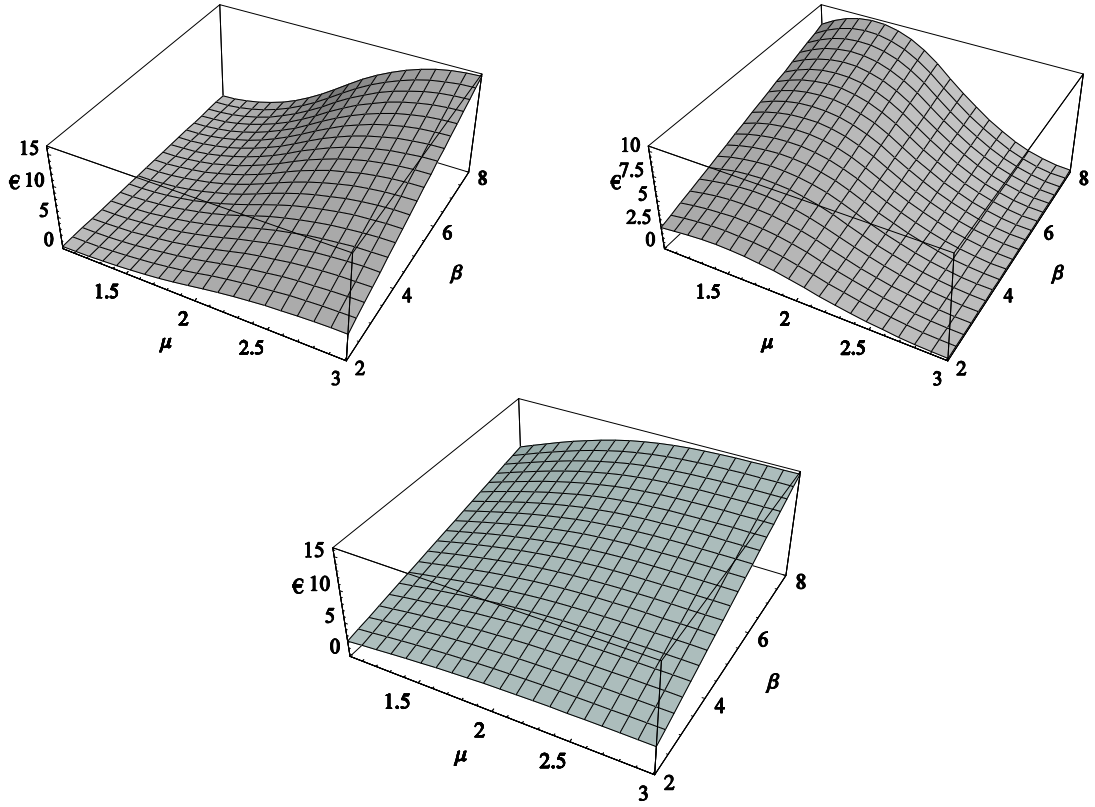

Figure 5. The distributions of schedule delay cost (top left), travel delay cost (top right), and their sum or generalised price (bottom) in the no-toll (NT) equilibrium

[4] This is the elasticity with respect to a total generalised price $P_i^{NT} \equiv p_i^{NT} + c_i^0$ that includes 30 minutes of free-flow travel time and operating costs of 7.30. These two costs components were considered only for calibration purposes, and are ignored in what follows.

[5] This was achieved with $b_i[\mu_i,\beta_i]$ set equal to the density function $f[\mu_i,\beta_i]$ shown in Figure 4, and $A_i[\mu_i,\beta_i]$ equal to the NT-equilibrium generalised price including operating and free-flow travel time costs, and $A \approx 53.1841$ and $B \approx 0.0059094$.

In the no-toll equilibrium, the average generalized price is 8.94. Travel delays vary from 0 to 47 minutes, with an average of 24 minutes. The average queuing cost of 3.97 is lower than the average schedule delay cost of 4.97.[6]

The generalised price components vary by type of user. Figure 5 shows the relevant patterns. The left-top panel shows that for a given μ (implying identical arrival times), schedule delay cost rises linearly with β. For a given β, schedule delay cost rises nonlinearly with μ: the higher μ_i is, the further one arrives from t^* and thus the higher the schedule delay. The upper-right panel shows the travel delay cost. For a given μ, this cost rises linearly with β, since for a given μ, α rises linearly with β. For a given β, it generally falls with μ, as high-α drivers avoid long queues by so much that their travel delays are often lower than those for low-α users. The bottom panel shows the sum of these two cost components (*i.e.* the generalised price); a higher μ or β increases the price.

The generalised prices in the bottom panel of Figure 5 follow from the analytical price equation that generalises (6) for the bivariate continuous distribution:

$$p_i^{NT} = \frac{\delta_i}{s} \cdot \left(\int_{\underline{\mu}}^{\mu_i} m_j^{NT}[\mu_j] \mathrm{d}\mu_j + \int_{\mu_i}^{\overline{\mu}} m_j^{NT}[\mu_j] \cdot \frac{\mu_i}{\mu_j} \mathrm{d}\mu_j \right), \tag{9}$$

with:

$$m_j^{NT}[\mu_j] = \int_{\underline{\beta}}^{\overline{\beta}} n_{jz}^{NT} \mathrm{d}\beta_z \ . \tag{10}$$

meaning that m_j^{NT} is the density of drivers with $\mu = \mu_j$ aggregated over β.

As was true for the discrete *I*-groups example in equation (6), users with a higher μ again impose lower marginal external cost. The intuition is the same: these users cause a less steep increase in queue length. Van den Berg and Verhoef (2011) show, for the case of a uniform β, that the average congestion externality and no-toll generalised price decrease with a mean-preserving increase in the heterogeneity of μ. Intuitively, this is closely related to the increasing convexity of the equilibrium travel time function. As equation (9) shows, a driver with a given μ benefits from replacing some drivers with a higher μ by drivers with an even higher μ, but does not suffer from replacing some drivers with a lower μ by drivers with a even lower μ. It is this asymmetry that causes aggregate costs to go down with a mean-preserving increase in the heterogeneity of μ.

4. Optimal pricing: the first-best equilibrium

Vickrey (1973) and Newell (1987) already argued that the optimum in this type of model requires an elimination of queuing, as is true for the model with homogeneous users. Schedule

[6] With homogeneous users these two averages are equal, due to the linearity of the schedule delay cost functions. Although the examples in Section 2.3 may lead one to expect average schedule delay costs exceeding average travel delay costs with heterogeneity in both μ and β, this is not necessarily the case. A counterexample can be constructed with, again, two groups. If the group with the higher μ has the lower value of α, then the average travel delay costs actually exceeds the average schedule delay cost.

delay costs are minimised in the optimum by having users arriving in order of increasing β before t^*, and decreasing $\gamma \equiv \eta \cdot \beta$ after t^*.

Because queuing is eliminated, prices depend on the values of schedule delay β and γ only, and no longer on the value of time α. Hence, the distribution of price components can now be depicted along one dimension, as a function of β alone. Figure 6 shows the resulting diagram. A given schedule delay is more costly with a higher value of schedule delays. But higher-β users arrive at more central moments, causing schedule delay costs to fall with β over a substantial range of Figure 6. However, these users have to pay relatively high tolls to obtain these arrival times. The result is that the generalised price rises monotonously with β. This is a general result, at least in this model. A user with a lower β could always travel at the same moment as a driver with a higher β, paying the same toll, but incurring lower scheduling cost. Because a low-β driver chooses to drive at a different moment, this implies that her price is even lower.

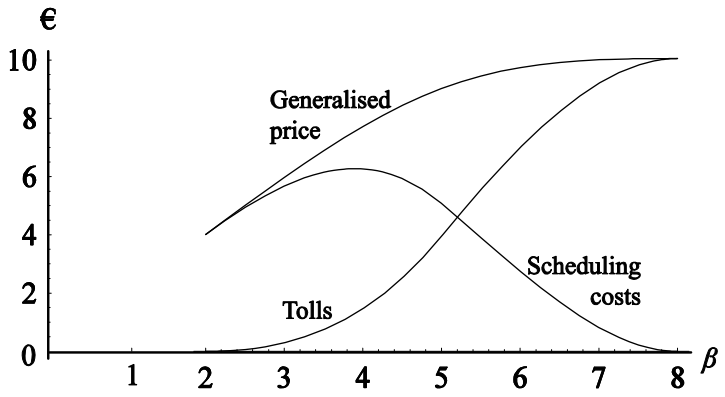

Figure 6. The distributions of schedule delay cost, toll, and their sum or generalised price in the first-best (FB) equilibrium

The generalised prices in Figure 6 are consistent with the analytical solution that we obtain as a generalisation of (7) for a continuous distribution:

$$p_i^{FB} = \frac{\delta_i}{s} \cdot \left(\int_{\underline{\beta}}^{\beta_i} q_j^{FB}[\beta_j] \cdot \frac{\beta_j}{\beta_i} \mathrm{d}\beta_j + \int_{\beta_i}^{\overline{\beta}} q_j^{FB}[\beta_j] \mathrm{d}\beta_j \right), \tag{11}$$

with:

$$q_j^{FB}[\beta_j] = \int_{\underline{\mu}}^{\overline{\mu}} n_{yj}^{FB} \mathrm{d}\mu_y \,, \tag{12}$$

meaning that q_j^{FB} is the density of drivers with $\beta = \beta_j$, aggregated over μ. As in the discrete example, the highest-β users benefit most from heterogeneity in the optimum. These same users also cause the highest increase of the generalised prices for other users (because part of this price is a toll, we are cautious not to call it the marginal external cost).

We can now compare the prices of equations (11) and (9), and use the difference $\Delta P_i = p_i^{FB} - p_i^{NT}$ 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 a contour plot in the right one. Note that, for ease of reference, we use α, not μ, along one of the axes. In the contour plot, the straight line on the left represents the locus with $\alpha=\beta$.

A number of patterns are worth emphasising. One is that it is striking that so many user types win from the imposition of first-best pricing. 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 generalised prices. In the present model, besides eliminating travel delays, the toll reduces aggregate schedule delays. This is an additional benefit that is enjoyed by the travellers, causing the change to be more favourable to them.

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. Social surplus (including toll revenues) rises 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 total 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 due to the more efficient order of arrivals. In short, the average user benefits from first-best pricing, even before toll revenues are recycled. Moreover, a majority of drivers, of 55%, benefits.

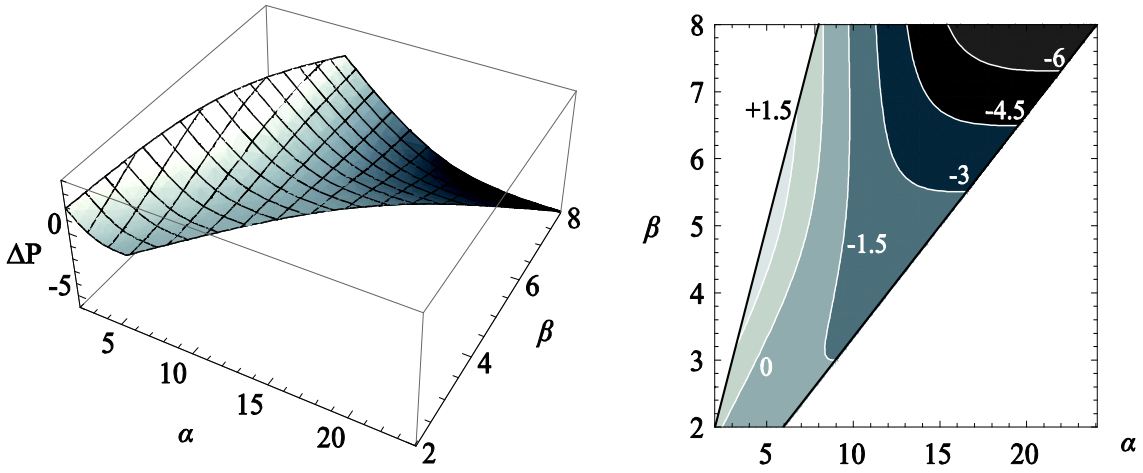

Figure 7. ΔP as a function of α and β for first-best pricing

The contour plot in Figure 7 shows that, for a given β, ΔP_i decreases with α, and therefore tolling is more beneficial or less harmful for drivers with a larger value of time. This confirms the common wisdom on the distributional impacts of road pricing. However, the dependence of the gains on β causes an interesting twist. The greatest losses are not incurred by the drivers with the smallest α. Instead, users with an intermediate β and the lowest possible α for that β have the highest ΔP_i and therefore lose most. Their low α allowed these 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, these are the users who 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 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 *and* a higher β. In other words, with heterogeneity in the value of schedule delays, the gains and losses of first-best congestion pricing need not be perfectly correlated with the value of time.

Finally, recall that Vickrey (1973) considers the case where the values of time and schedule delay vary proportionally over users. He finds that first-best pricing produces a strict Pareto improvement. Conversely, in our model some users lose. A first reason is that 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), possibly making them worse off. Secondly, we extend Vickrey's heterogeneity with heterogeneity in μ (the relative size of the value of time). As we shall see in Section 5, the number of users that gains from an FB toll decreases with the heterogeneity of μ, which is consistent with Vickrey's result for a homogeneous μ.

5. Second-best pricing: an untolled alternative

It is not uncommon for road congestion pricing schemes in practice to offer an untolled alternative. For example, on so-called 'express lanes' or 'pay-lanes', some of the lanes on a highway are tolled and some are free of charge. Another example is a tolled highway with an untolled secondary road running parallel. From a distributional viewpoint, this is an interesting case. Such schemes allow drivers to choose their preferred mix of travel time and toll, which presumably changes the distributional impacts of pricing.

Verhoef and Small (2004) study the distributional impacts of this policy in a static model of traffic congestion with a continuous distribution of the value of time. In their model, a second-best toll causes a shift of traffic from the pay-lane to the free-lane. Whereas their results for first-best pricing confirm the expectation that the private loss falls with the value of time, this second-best policy has a non-monotonous impact on drivers' generalised prices. The largest loss is incurred by users with the 'critical value of time', who are indifferent between the 'pay-lane' and the 'free-lane'. People with a lower value of time choose the free-lane. They incur a smaller loss, because the longer travel time on the 'free-lane' is less costly for them. People with higher values of time choose the pay-lane, and incur a smaller loss, or even a gain, because the shorter travel time is more valuable with a higher value of time. On average, changes of generalised prices are smaller for second-best pricing than for first-best pricing, since the second-best toll on the pay-lane is lower than the first-best toll, and there is no toll on the free-lane.

Braid (1996) and de Palma and Lindsey (2000) demonstrate how with two bottlenecks in parallel, it remains optimal to eliminate all queuing at the priced bottleneck (we will refer to this priced alternative as the pay-lane). But there is still queuing at the untolled bottleneck (the free-lane). Because this implies that for equally long lasting peaks, the marginal social cost would be lower on the pay-lane, it is second-best optimal to set a negative time-independent tax $\bar{\tau}$ (*i.e.* a subsidy) on this pay-lane. Consequently, the pay-lane has a longer peak and larger average schedule delays, but no travel delays. The relative efficiency of this

second-best policy is generally higher in the bottleneck model than for static congestion, where there are no benefits from departure time adjustments.

The second-best (SB) equilibrium in our model is characterised by a mixture of generalised prices as in (11) for the pay-lane, with of course $\bar{\tau}$ added, and as in (9) for the free-lane. We were unable to derive an analytical solution for this model, even when we reduced the complexity by assuming fixed demand. One may expect that for a given β, it is the users with a relatively low α who use the free-lane, as the travel delays are less harmful to them. Reversely, for a given α, if there are drivers on the free-lane, it is more likely that these are drivers with a higher β, as the negative time-invariant component of the toll makes the duration of the peak on the free-lane shorter, and this makes an arrival close to t^* more likely.

All this is confirmed in the right panel of Figure 8, which shows changes in generalised prices, $\Delta P_i = p_i^{SB} - p_i^{NT}$, and depicts the critical combinations of α and β for which drivers are indifferent between the two alternatives as the $\alpha^*[\beta]$-contour. In this simulation, the pay-lane has one-third, and the free-lane two-thirds, of the joint capacity of the single bottleneck in the previous sections.

Figure 8 shows that all drivers benefit from second-best pricing. An important reason is that the combination of a time-invariant subsidy and a time-variant toll means that the net toll receipts for the regulator are relatively modest. Again, for a given β, the benefits increase with α. Further, it is again drivers with an intermediate β, and the lowest α given this β, who gain least from SB pricing. The pattern is therefore not too dissimilar from that in Figure 7 for first-best pricing. A difference is that there are now increasing gains towards the lower right corner of the bivariate distribution, near α=5 and β=2: these are drivers with a low value of schedule delay and a relatively high value of time. They choose to drive early or late on the pay-lane and benefit from the net subsidy. Hence, surprisingly, not only users with high values of time and schedule delay use the pay-lane, but also drivers with a relatively low value of time, as long as their value of schedule delay is sufficiently low. These drivers' low β and γ make it worthwhile to go and collect the subsidy for very early or late arrivals on the pay-lane.

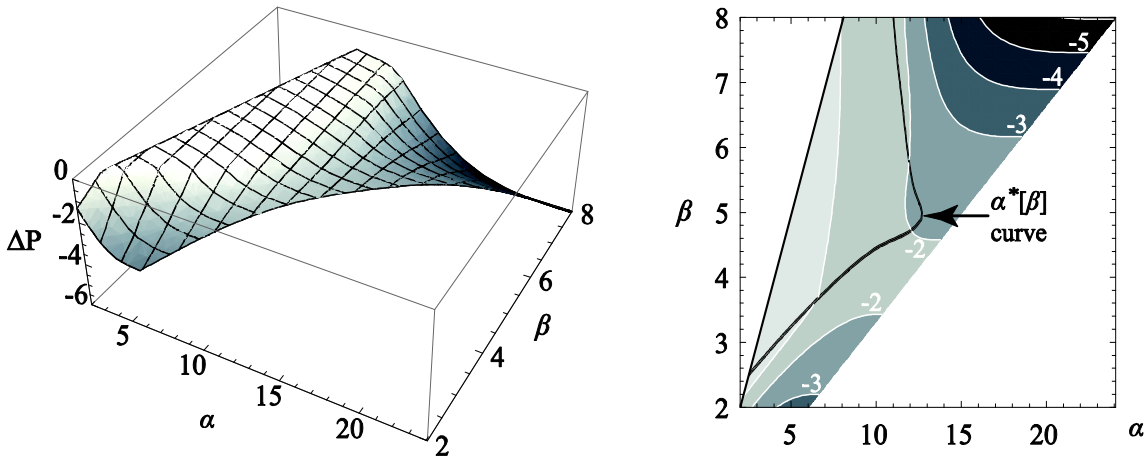

Figure 8. ΔP as a function of α and β for second-best pricing

Also in Braid's (1996) homogeneous user model, all users benefit from second-best pricing. Still, with heterogeneity in the value of schedule delay, users gain more. This follows from the self-ordering by value of schedule delay. Low-β drivers arrive on the outside of the pay-lane peak when schedule delays are the highest. High-β users arrive close to t^*.

In the static model, drivers with the critical values of time lose most due to second-best pricing. In our model, there seems 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 α. An interesting question is why ΔP_i is falling with α for the free-lane users (who are to the left of $\alpha^*[\beta]$), causing the discrepancy with the static model. The answer is that the subsidy on the pay-lane reduces the duration of the peak on the free-lane. This reduces travel delays, from which the higher-α drivers benefit more than the lower-α users.[7]

The second-best time-invariant toll is –5.36. The average time-variant toll is 6.25, and the net tax revenue is only 4095. Consumer surplus increases by 7.0% from the NT case to 256 142 (1.4% in FB), and social surplus by 8.7% to 260 237 (17.7% in FB). The number of users increases by 3.4% to 9309.8 (0.6% in FB). The gain in social surplus relative to the gain from first-best pricing is 49%, even though the pay-lane only has one-third of capacity. This contrasts strongly with results from static models, where much lower relative gains are reported, even when the pay-lane makes up half the total capacity. For example, in a model with heterogeneity in the value of time, Small and Verhoef (2007) find values between 23% and 35% for pay-lanes having 25% to 50% of total capacity.

6. Sensitivity analysis

It is important to test the sensitivity of our results with respect to some key parameters. We focus on the effects of different assumptions on the bivariate distribution of β and μ, to reflect our emphasis on distributional effects. We consider five distributions: *homogeneity* with identical preferences; the *base case* considered above with the triangular distribution of Figure 3; a case with *less heterogeneity in* μ; a case with *less heterogeneity in* β; and a *uniform distribution*. In all cases, the mean of β is 5, and of μ it is 2.01. The base case spread in the distribution for μ is 2, and for β it is 6. With *less heterogeneity in* μ, the former is reduced to 1; with *less heterogeneity in* β, the latter is reduced to 2. With the *uniform distribution*, the variances are the same as in the *base case*. Otherwise, we recalibrated such that the aggregate demand elasticity remains –0.4 and the number of NT users 9000, using the same calibration method as for the base case.[8]

Arguably, the most striking result in Table 1 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 and thus raises the gain from pricing. Reducing the heterogeneity in β lowers the gains

[7] Consistent with this explanation, we found for the case of a profit-maximising pay-lane, not reported in this paper for reasons of brevity, a positive time-invariant tax and local maxima of ΔP_i at the contour $\alpha^*[\beta]$.

[8] A by-product of this is that 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 the cases.

from the reordering of arrival times, lowering the gain from pricing. A larger share of the NT drivers benefits from first-best pricing with less heterogeneity in μ, and a smaller share gains with less heterogeneity in β. This is consistent with the predictions we made on the basis of the two two-groups examples in Section 2.

The distributional effects of first-best and second-best tolling, reflected by the patterns shown in Figures 7 and 8, are also rather robust. The main difference is that with more heterogeneity in β or μ, losses and gains are larger. This is, at least for first-best pricing, consistent with the observation that for homogeneous users, everybody is equally well off when pricing is introduced.

	Homogeneity	Base case	Less hetero-geneity in μ	Less hetero-geneity in β	Uniform
Spread of μ (NT)	–	2	1	2	1.141
Spread of β (NT)	–	6	6	2	4.243
		NT Equilibrium			
Number of users	9000	9000	9000	9000	9000
Social surplus= Consumer surplus	239 332	239 332	239 332	239 332	239 332
		FB Equilibrium			
Number of users	9000	9054.6	9122.3	8922.8	9055.3
Toll revenues	44 770	39 137	39 746	41 970	39 094
Consumer surplus	239 332	242 571	246 180	235 352	242 606
Social surplus	284 102	281 708	285 926	277323	281 701
%Δ Social surplus from NT	18.7%	17.7%	19.5%	15.9%	17.7%
% NT users with decrease in *price*	*p* unchanged	55%	66%	39%	53%
		SB Equilibrium			
Number of users	9302.8	9309.8	9356.4	9205.5	9293.9
Toll revenues	3917	4095	4153	5680	4439
Consumer surplus	255 706	256 142	258 692	250 408	255 260
Social surplus	259 623	260 237	262 845	256 089	259 699
%Δ Social surplus from NT	8.48%	8.73%	9.82%	7.00%	8.51%
% NT users with decrease in *p*	100%	100%	100%	100%	100%
Relative efficiency	0.453	0.493	0.505	0.441	0.481

Table 1. Results of sensitivity analysis: NT (top), FB (middle) and SB (bottom) equilibria

7. Conclusion

This paper reconsidered the distributional effects of road congestion pricing, taking into account heterogeneity in the values of time and schedule delays. To gain a more complete picture of the patterns of these distributional impacts, we used continuous distributions. Especially because the model also allowed for price-sensitive demand, this made analytical solutions hard to obtain. But numerical analysis provided interesting insights.

We found that congestion pricing can leave a majority of travellers better off even without returning the toll revenues to these drivers. This contrasts quite sharply with popular beliefs that a majority of travellers, or even all of them, will lose from the imposition of first-best congestion pricing—as is predicted also by the conventional static textbook model of traffic congestion with homogeneous users. It also contrasts with the findings in Van den

Berg and Verhoef (2011), who only consider heterogeneity in the value of time and keep the value of schedule delays constant across users, and find smaller gains from congestion pricing. The existence of these gains may eventually help in overcoming public resistance against congestion pricing, although it will presumably not be easy to communicate these more optimistic conclusions from this more complicated analysis.

Furthermore, we found that when there is heterogeneity in the value of schedule delays, it is no longer true that the gains and losses from first-best congestion pricing are perfectly correlated with the value of travel time. For a given value of schedule delay, they are. Yet, overall it is not users with the lowest value of time that incur the greatest losses, or enjoy the smallest gains; but rather drivers with an intermediate value of schedule delays, and the lowest value of time for that value of schedule delay. In our model, some users with a relatively low value of schedule delay and a low value of travel time lose less than some users with higher values of time and schedule delay. Likewise, there are users with a relatively high value of schedule delay that enjoy a greater gain than some users with a higher value of time but with a lower value of schedule delay.

For second-best pricing with an untolled alternative, the pattern of distributional effects is similar as for first-best pricing. Still, a difference arises for users with a low value of schedule delay. These drivers may be attracted to the earliest and latest arrival times on the tolled bottleneck, and substantially benefit from the net subsidy that then applies.

Earlier research with a static model of traffic congestion suggested that users with the 'critical' value of time, who are indifferent between using the tolled or untolled alternative, lose most (or gain least) from this type of pricing. In our model, however, it is not the indifferent users who gain least from second-best pricing, but users with an intermediate value of schedule delay and the lowest value of time for that value of schedule delay.

Our results suggest that it is important to take into account the distributions of both values of travel time and schedule delays in assessing the distributional impacts of congestion pricing. This reinforces the need of including schedule delay costs as a standard component in the assessment of the cost of congestion and the social benefits of combating it—which in itself is already motivated by the fact that schedule delay costs make up half the congestion cost in the no-toll equilibrium of Vickrey's (1969) bottleneck model, and all of it in optimum. With the increasing popularity of mixed logit modelling in the analysis of scheduling decision of road users, insight into the relevant distributions is growing rapidly. The sensitivity analysis in this paper suggests that the exact distributions need not be known with the greatest precision to assess the potential impacts of heterogeneity on the effects of road pricing, as the results appear to be quite robust for the distribution.

References

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., 1993. A structural model of peak-period congestion: a traffic bottleneck with elastic demand. American Economic Review 83(1), 161–179.

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., 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. Congestion pricing in the morning and evening peaks: A comparison using the Bottleneck Model. In: Proceedings of the 39th Annual Conference of the Canadian Transportation Research Forum: 2002 Transportation Visioning - 2002 and Beyond, Vancouver, Canada, 9–12 May 2004, 179–193.
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.
Newell, G.F., 1987. The morning commute for nonidentical travellers. Transportation Science 21(2), 74–88.
Parry, I., 2002. Comparing the efficiency of alternative policies for reducing traffic congestion. Journal of Public Economics 85(3), 333–362.
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., 1982. The scheduling of consumer activities: work trips. American Economic Review 72(3), 467–479.
Small, K.A., Verhoef, E.T., 2007. The Economics of Urban Transportation. London: Routledge.
Small, K.A., Winston, C., Yan, J., 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. Econometrica 73(4), 1367–1382.
Small, K.A., Yan, J., 2001. The value of "value pricing" of roads: second-best pricing and product differentiation. Journal of Urban Economics 49(1), 310–336.
van den Berg, V. and Verhoef, E.T., 2011. Congestion tolling in the bottleneck model with heterogeneous values of time. Transportation Research Part B 45(1), 60–70.
Verhoef, E.T., Small, K.A., 2004. Product differentiation on roads: constrained congestion pricing with heterogeneous users. Journal of Transport Economics Policy 38(1), 127–156.
Vickrey, W.S., 1969. Congestion theory and transport investment. American Economic Review (Papers and Proceedings) 59(2), 251–260.
Vickrey, W.S., 1973. Pricing, metering, and efficiently using urban transportation facilities. Highway Research Record 476, 36–48.
West, S.E., 2004. Distributional effects of alternative vehicle pollution control policies, Journal of Public Economics 88(3–4), 735–757.
Xiao, F.(E.), Qian, Z.(S.), Zhang, H.M., 2010. The morning commute problem with coarse toll and nonidentical commuters. Networks and Spatial Economics. DOI 10.1007/s11067-010-9141-8.