CONGESTION CAUSED BY SPEED DIFFERENCES

Erik Verhoef[*], Jan Rouwendal[**] and Piet Rietveld[*]

[*]Department of Spatial Economics, Free University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, Phone: +31-20-4446094, Fax: +31-20-4446004, Email: everhoef@econ.vu.nl, http://www.econ.vu.nl/vakgroep/re/members/everhoef/et.html

[**]Department of Economics and Management, Wageningen Agricultural University, PO Box 8060, 6700 DA Wageningen, The Netherlands

Key words: Congestion, Speed, Road Pricing
JEL codes: R41, R48, D62

Abstract

In this paper, we investigate congestion caused by differences in desired or possible speeds. Especially outside peak hours, speed differences are probably one of the most important reasons for congestion. Although the model setting, with one lane and no overtaking, may seem simple at first sight, the problem turns out to result easily in quite complicated mathematical expressions. Some main conclusions are that optimal tolls for slow vehicles are higher than those for fast drivers, that the marginal external costs and the optimal tolls for slow drivers are actually decreasing in the equilibrium number of slow drivers, and that 'platooning' may become an attractive option especially when the desire for a low speed is caused by a lower value of time.

Buβgeld für Langsamfahrer

AFP München - Wer mit seinem Auto auf einer unübersichtlichen Straße im "Schneckentempo" fährt und nachfolgende Wagen behindert, muß mit einem Bußgeld rechnen. Nach einem gestern vom ADAC in München veröffentlichten Urteil des Amtgerichts Gemünden verstoßen Langsamfahrer gegen die Verkehrsordnung, wenn ein Überholen wegen ihrer Fahrweise nicht möglich ist. (Az.: OWi 372 Js 59889/96, DAR97, 251).

From: Die Welt, July 12, 1997. [1]

1. Introduction

In economic models of traffic congestion attention is focused on the external effect that drivers impose upon each other. The externality involved is that the generalized travel costs of each driver increase by the presence of other drivers. An important component of these costs is the travel time, which is inversely related to the speed during the trip. In bottleneck models, a lower speed is imposed upon a driver when waiting in the queue at the bottleneck (Arnott, De Palma and Lindsey, 1993). In other models, a speed-flow relationship is used to motivate the decrease in speed that is associated with higher density of traffic. In many, if not all of these models, the traffic flow is assumed to be homogeneous in the sense that all drivers are assumed to have the same speed under identical circumstances on the road network. This approach is suggested by the concept of a 'representative consumer' which is an important, although controversial, analytical tool in economics. For instance, Rotemberg [1985] has analyzed the efficiency of equilibrium traffic flows on the basis of such a model.

However, the assumption of a representative driver need not always be realistic. An important part of actual congestion, especially outside peak hours, seems to be caused primarily by the fact that various drivers using the same road have different preferred and actual speeds, either because of the characteristics of their vehicle, or by 'pure' choice. An important example is the presence of trucks and private cars on the same roads. Typically, truck drivers have a lower preferred speed, or technically possible maximum speed, than private car drivers. If overtaking is impossible or prohibited, the drivers of private cars are forced to have the same speed as the trucks whenever the distance between them becomes smaller than a minimum determined by safety considerations. Another possibility is when different types of passenger traffic use the same road. Often, business travellers have a higher desired speed than touristic road users, or drivers who make a trip for social purposes (see also Rienstra and Rietveld, 1996). Other examples of different desired speeds for different types of road users can be thought of once it is recognized that people with different socio-economic backgrounds may often differ in terms of their desired speeds. Clearly, this type of congestion cannot be analysed with a model that assumes a population of identical optimizing drivers.

Most analytical models of road traffic congestion and congestion pricing focus on peak hour congestion, and consider homogeneous travellers. Even when heterogeneity of

[1] **Fines for slow drivers.** Those who drive their cars at snail-pace on a difficult to survey road, and hinder following cars, should count on a fine. According to a verdict by ADAC, Munchen, made public yesterday, slow drivers offend traffic laws when overtaking is not possible due to their driving behaviour.

drivers is allowed for, it is usually assumed that all drivers have the same speed when driving in congestion. The typical source of heterogeneity considered in such models concerns income differences (see, for instance, Arnott, De Palma and Lindsey, 1994). In this paper, we study a different type of congestion, and a different type of heterogeneity, namely congestion caused by differences in desired or possible speeds. This type of congestion is also important for traffic policy. In the Netherlands, for instance, overtaking by truck has recently been prohibited on some parts of the highway network in order to prevent this type of congestion. An economic analysis of such a measure would call for a model in which drivers are heterogeneous with respect to the speed they choose. To the best of our knowledge, the only economic analysis of this topic that comes close to our model is the one by Tzedakis (1980). Our paper differs from his in that we derive analytical expression for optimal tolls for both types of drivers, consider second-best tolling, and study 'platooning'. Other related papers include those studying speed differences in relation to safety (e.g. Rodriguez, 1990), and those dealing with speed limits (Lee, 1985; Lave, 1985; followed by Fowles and Loeb, 1989; Levy and Asch, 1989; Snyder, 1989; and Lave, 1989)

In this paper, we assume that there are two types of drivers: those with a high, and those with a low preferred speed. In the following section we consider traffic on a road segment used by these two types of drivers and derive their travel time function. Section 3 discusses optimal and second-best tolls, and presents a numerical illustration. Section 4 introduces an additional policy, namely 'platooning', where slow drivers have to wait until a certain number of them is present at the entrance of the road before a platoon of slow drivers is allowed to enter the road. Again a numerical illustration is provided. Finally, Section 5 concludes.

2. Travel time with two speeds

We consider traffic on a road segment of length m. It is assumed that this road segment has no junctions, is completely flat without bends, etc. There are two types of drivers: those who want to drive fast (type 1) and those who want to drive slow (type 2). Fast drivers want to maintain a constant speed s_1 during their trip, slow drivers a speed s_2 (of course: $s_2<s_1$). On the road segment overtaking is impossible. All drivers want to maintain a minimum distance d^* to each other under all circumstances (d^* is measured as the difference between two subsequent cars' fronts). Fast drivers drive at their preferred speed as long as this critical distance to a preceding car is not trespassed, and slow down instantly to the speed of this predecessor as soon as it is reached. Whether or not a fast driver will experience congestion is then completely determined by the number and location of vehicles on the road segment, and the type to which these drivers belong, at the moment a fast driver enters the road segment.

Under the assumptions made, we can be somewhat more specific: whether or not a fast driver will be able to maintain his preferred speed on the whole road segment depends only on the location y of the slowly driven vehicle that was the last to enter that road

segment, and on the number k of fast driven automobiles that have entered behind this slowly driven vehicle, but before this driver whose travel time we want to determine. This driver will experience congestion as soon as he reaches the back of the platoon of k automobiles that has in front the slowly driven automobile that was at y when the fast driver entered the road segment.

For the most elementary setting of the model, we assume that the arrival patterns of the two types of drivers are fully deterministic (a formulation with stochastic arrivals can be found in Rouwendal, Verhoef, Rietveld and Zwart, 1997). Both groups i have an arrival rate ρ_i, which is endogenized below by considering elastic demand. In the present deterministic model, it is assumed that the time span between the arrivals of two subsequent drivers of the same group at the entrance is always exactly equal to $z_i=1/\rho_i$. To avoid bottleneck queuing before the entrance, we assume throughout the paper that $1/(\rho_1+\rho_2)>d^*/s_2$. In other words, in this paper, we isolate congestion resulting from speed differences from 'ordinary' road traffic congestion, and do so by assuming that the latter type does not occur on our road segment – postponing the joint consideration of these two types of congestion to future research.

The expected travel time for slow drivers, $E(T_2)$, is simply equal to m/s_2. In order to determine the expected travel time for the fast drivers, $E(T_1)$, we first derive a fast driver's travel time for a given y and k. At the instant that a fast driver starts her trip, her foreseen position at the back of the platoon is at a distance of $y-(k+1)\cdot d^*$ meters. After a time span of $\tau=(y-(k+1)\cdot d^*)/(s_1-s_2)$ seconds, the fast driver has taken her position at the back of the platoon. Note that, no matter where the k fast drivers were at the instant the fast driver under consideration entered the road, they will all have taken their respective positions in the platoon as soon as this fast driver does. In the meantime, the slow driver has moved another $s_2\cdot\tau$ meters, and still has to drive another $m-y-s_2\cdot\tau$ meters. Since we assume that all vehicles speed up as soon as the slow driver has reached the end of the road, the fast driver's travel time for a given y and k, $T_1(y,k)$, can therefore be calculated as:

$$T_1(y,k)=\frac{m-y-s_2\cdot\left(\frac{y-(k+1)\cdot d^*}{s_1-s_2}\right)}{s_2}+\frac{y+s_2\cdot\left(\frac{y-(k+1)\cdot d^*}{s_1-s_2}\right)}{s_1}$$

$$=\frac{m}{s_2}-\frac{y}{s_2}+\frac{(k+1)\cdot d^*}{s_1} \qquad (1a)$$

$$with:\ \frac{m}{s_1}\le T_1(y,k)<\frac{m}{s_2}+\frac{d^*}{s_1} \qquad (1b)$$

The upper bound of $T_1(y,k)$, which may be larger than expected at first sight (one would expect m/s_2), can be understood after realizing that ρ_1 and ρ_2 need not be exact multiples, so that it is possible that a fast driver starts her trip so soon after a slow one that the safety distance requirement is violated. When y approaches zero, the fast driver has to wait (almost) d^*/s_2 seconds before starting, after which she drives $m-d^*$ meters at speed s_2, and

the last d^* meters at speed s_1; hence the upper bound for $T_1(y,k)$ (we assume that, if a fast and a slow driver arrive at exactly the same instant, the fast one takes advantage).

For the determination of $E(T_1)$, we define p_c as the probability that a fast driver experiences congestion (is hindered) at all. The relevant inequality defining combinations of y and k for which the platoon speeds up earlier than or exactly when the fast driver takes her position, so that she is not hindered, is $y+s_2\cdot(y-(k+1)\cdot d^*)/(s_1-s_2)\geq m$.

Since we assume a fully deterministic process, we can write k as a function of y. Although k(y) will be a step function in reality[2], we will use a continuous expression for k for reasons of mathematical convenience. Because, for a given y, y/s_2 seconds have passed since a slow driver at y has entered the road, this function can be found to be of the form $k(y)=\rho_1\cdot y/s_2$. First of all, this means that we can rewrite (1a) as:

$$T_1(y) = \frac{m}{s_2} - \frac{y}{s_2}\cdot\left(1-\frac{d^*\cdot\ _1}{s_1}\right)+\frac{d^*}{s_1} \tag{2}$$

Moreover, we can now write the inequality defining whether a fast driver is hindered as:

$$y < \frac{m\cdot(s_1-s_2)+s_2\cdot d^*}{s_1-d^*\cdot\ _1}$$

Finally, since y can take any value in the interval $<0,s_2/\rho_2]$ with equal probability, we can now express p_c as a function of ρ_1, ρ_2 and the relevant constants:

$$p_c(\ _1,\ _2)=1 \quad if \quad \frac{s_2}{\ _2} < \frac{m\cdot(s_1-s_2)+s_2\cdot d^*}{s_1-d^*\cdot\ _1}$$

$$p_c(\ _1,\ _2)=\frac{m\cdot(s_1-s_2)+s_2\cdot d^*}{\left(s_1-d^*\cdot\ _1\right)\cdot\frac{s_2}{\ _2}} \quad otherwise \tag{3}$$

The expected travel time for group 1 can therefore be written as:

$$E(T_1)=(1-p_c(\ _1,\ _2))\cdot\frac{m}{s_1}+ \tag{4a}$$
$$p_c(\ _1,\ _2)\cdot\int_0^{\frac{m\cdot(s_1-s_2)+s_2\cdot d^*}{s_1-d^*\cdot\ _1}} T_1(\ ,\ _1)\cdot\frac{s_1-d^*\cdot\ _1}{m\cdot(s_1-s_2)+s_2\cdot d^*}\,d \qquad if\ \ p_c<1$$

$$E(T_1)=\int_0^{\frac{s_2}{\ _2}} T_1(\ ,\ _1)\cdot\frac{\ _2}{s_2}\,d \qquad if\ \ p_c=1 \tag{4b}$$

[2] Observing that y/s_2 seconds have passed since a slow driver at y has entered the road, the step function for k(y) can be written as: k=0 if $0<y/s_2\leq(z_1=)1/\rho_1$; k=1 if $1/\rho_1<y/s_2\leq2/\rho_1$; k=2 if $2/\rho_1<y/s_2\leq3/\rho_1$, etc. The function in the main body is the continuous version of this function. Note that it is assumed for this step function that if a fast and a slow driver arrive at the entrance at exactly the same instant, the fast driver 'wins'.

where $T_1(\psi,\rho_1)$ is as given in (2). Note that the upper limits in the integrals give the maximum value that y can take under the relevant regime. Since both integrands are linear in y, we can substitute the expected value of y (half the upper limit of integration in (4a) and (4b)) into the integrands in (4a) and (4b) to determine $E(T_1)$. After some manipulations, one can then obtain:

$$E(T_1) = (1 - p_c(\ _1,\ _2)) \cdot \frac{m}{s_1} + p_c(\ _1,\ _2) \cdot \left(\tfrac{1}{2} \cdot \left(\frac{m}{s_1} \right) + \tfrac{1}{2} \cdot \left(\frac{m}{s_2} + \frac{d^*}{s_1} \right) \right) \qquad if\ \ p_c < 1 \tag{5a}$$

$$E(T_1) = \frac{m}{s_2} - \frac{1}{2 \cdot\ _2} \cdot \left(1 - \frac{d^* \cdot\ _1}{s_1} \right) + \frac{d^*}{s_1} \qquad if\ \ p_c = 1 \tag{5b}$$

Equation (5a) shows that if $p_c<1$, the impact of the two arrival rates on the expected travel time for group 1 runs completely via their impact on the probability of being hindered. The overall expected travel time for group 1 is $(1–p_c)$ times the free-flow travel time for a fast driver, plus p_c times the expected travel time conditional on congestion – which is, as expected, the average of the minimum travel time and the maximum travel time. Note that by (3), $p_c>0$ if $\rho_1>0$ and $\rho_2>0$, so that the expected travel time for group 1 is always greater than the minimum possible travel time for that group for all non-trivial cases. If $p_c=1$, the expected travel time is the maximum possible travel time minus a term that is decreasing in ρ_1 and ρ_2. (Note that $s_1–d^*\cdot\rho_1>0$ because of the assumption that queuing before the entrance of the road does not occur, so that the term between the large brackets in (5b) is always positive).

3. Optimal and second-best tolls

The standard economic prescription for dealing with road traffic congestion is to impose tolls that should reflect the marginal external congestion costs. Usually, such optimal tolls are increasing in road usage, because in most models, the marginal external congestion costs are. In this section, we will analyse tolls for the specific type of congestion described below, and we will conclude that this common wisdom does not hold for congestion caused by differences in desired speed – as long, of course, as this type of congestion occurs in isolation.

In order to be able to derive tolls, we first have to introduce values of time for the purpose of 'translating' expected travel times into expected travel costs. We use V_i to denote group i's value of travel time. In case slow drivers are truck drivers, $V_2>V_1$ is the most likely case; however, $V_2<V_1$ could correspond to the situation where the slow drivers are those who use the road for touristic or social purposes, and hinder business travellers. Using these values of time, we can write the expected travel costs for both groups, k_i, as follows:

$$k_1 = (\ \ -1)\cdot c_1^{\,0} + \ \ \cdot c_1^{\,1} \tag{6a}$$

with:

$$= 0 \ if \ \ p_c < 1 \quad and \quad = 1 \ if \ \ p_c = 1$$

$$c_1^{\,0} = V_1 \cdot \left[(1 - p_c(\ _1,\ _2)) \cdot \frac{m}{s_1} + p_c(\ _1,\ _2) \cdot \left(\tfrac{1}{2} \cdot \left(\frac{m}{s_1} \right) + \tfrac{1}{2} \cdot \left(\frac{m}{s_2} + \frac{d^*}{s_1} \right) \right) \right]$$

$$c_1^{\,1} = V_1 \cdot \left[\frac{m}{s_2} - \frac{1}{2 \cdot \ _2} \cdot \left(1 - \frac{d^* \cdot \ _1}{s_1} \right) + \frac{d^*}{s_1} \right]$$

$$k_2 = V_2 \cdot \frac{m}{s_2} \tag{6b}$$

Next, we specify two inverse demand curves $D_i(\rho_i)$ for both groups, which gives the marginal willingness to pay (in generalized costs). The stage is then set to derive congestion tolls. Two types of tolls will be considered: optimal tolls, which are group specific, and second-best tolls, where the regulator sets the same toll for both groups. This is relevant when either the regulator is not capable of distinguishing between fast and slow drivers, which would occur in case the slow travellers also have passenger cars, or when the tolling technology simply does not allow any fee differentiation.

Optimal tolls

The optimal tolls f_1 and f_2 can be derived by maximizing social welfare per unit of time subject to the behavioural relations under tolling:

$$\underset{f_1,\ f_2}{MAX}\ W = \int_0^{\ _1} D_1(x_1)\, dx_1 + \int_0^{\ _2} D_2(x_2)\, dx_2 - \ _1 \cdot k_1(\ _1,\ _2) - \ _2 \cdot k_2 \tag{7a}$$

s.t.:

$$D_1(\ _1) - k_1 - f_1 = 0 \tag{7b}$$

$$D_2(\ _2) - k_2 - f_2 = 0 \tag{7c}$$

The first-order conditions for (7a) can be combined with (7b) and (7c) to obtain familiar expressions for the optimal tolls:

$$f_1 = \ _1 \cdot \frac{k_1}{\ _1} \tag{8a}$$

$$f_2 = \ _1 \cdot \frac{k_1}{\ _2} \tag{8b}$$

Obviously, it is not (8ab) that we are primarily interested in, but rather the explicit expressions that can be obtained by using (5a) and (6a). For that purpose, observe that:

$$\frac{c_1^{\,0}}{1} = \frac{p_c}{1}\cdot\left[V_1\cdot\tfrac{1}{2}\cdot\left(\frac{m}{s_2}+\frac{d^*}{s_2}-\frac{m}{s_1}\right)\right]$$
$$= \frac{d^*\cdot\left(m\cdot(s_1-s_2)+s_2\cdot d^*\right)}{\left(s_1-d^*\cdot\ {}_1\right)^2\cdot\frac{s_2}{2}}\cdot\left[V_1\cdot\tfrac{1}{2}\cdot\left(\frac{m}{s_2}+\frac{d^*}{s_2}-\frac{m}{s_1}\right)\right] \quad (9a)$$

$$\frac{c_1^{\,0}}{2} = \frac{p_c}{2}\cdot\left[V_1\cdot\tfrac{1}{2}\cdot\left(\frac{m}{s_2}+\frac{d^*}{s_2}-\frac{m}{s_1}\right)\right]$$
$$= \frac{m\cdot(s_1-s_2)+s_2\cdot d^*}{\left(s_1-d^*\cdot\ {}_1\right)\cdot s_2}\cdot\left[V_1\cdot\tfrac{1}{2}\cdot\left(\frac{m}{s_2}+\frac{d^*}{s_2}-\frac{m}{s_1}\right)\right] \quad (9b)$$

$$\frac{c_1^{\,1}}{1} = V_1\cdot\frac{d^*}{2\cdot\ {}_2\cdot s_1} \quad (10a)$$

$$\frac{c_1^{\,1}}{2} = V_1\cdot\frac{s_1-d^*\cdot\ {}_1}{2\cdot\ {}_2^{\,2}\cdot s_1} \quad (10b)$$

The explicit expressions for the optimal fees can be found by substituting (9ab) when $p_c<1$ and (10ab) when $p_c=1$ into (8ab). It is easily verified that all marginal external congestion costs are larger than zero in non-trivial cases where $\rho_1>0$ and $\rho_2>0$. However, it is noteworthy that for $p_c=1$ in (10b), f_2 is *decreasing* in ρ_2. For $p_c<1$, (9b) is independent of ρ_2 (because p_c is then linear in ρ_2 by (3)), but since ρ_1 will be decreasing in ρ_2, we can expect that when comparing equilibrium values of the marginal external congestion costs of group 2, also here they will be decreasing in ρ_2. In other words, the more slow vehicles use the road in the optimum, the lower their optimal fee and their marginal external congestion costs.[3] This is caused by the fact that, when more slow vehicles are present, their marginal effect on expected travel times for fast drivers decreases. There are two reasons for this. First, the expected travel time for fast drivers has an upper limit (see (1b)). The more slow vehicles are present, the more closely this limit is reached and hence the smaller the impact of a marginal reduction in ρ_2 on $E(T_1)$. Secondly, when ρ_2 is larger, the number of fast drivers affected by an individual slow driver decreases. Because $\partial k_1/\partial\rho_2$ and f_2 are decreasing in ρ_2, multiple solutions can be found when using (9b) and (10b) for determining the optimal f_2. A single solution will only be found when D_2 is sufficiently inelastic. In other cases, one has to compare social welfare for each of the candidates to determine the 'true' optimum. Note that even second-order conditions have to be used with care, since these will only indicate the 'global' optimum for either the limited range where $p_c<1$ (for (9b)), or for the limited range where $p_c=1$ (for (10b)).

Finally, it can be observed that both for $p_c<1$ and for $p_c=1$, we find:

[3] Tzedakis (1980) obtained a similar result with a simulation model.

$$\frac{\frac{k_1}{\quad 2}}{\frac{k_1}{1}} = \frac{s_1 - d^* \cdot \quad 1}{d^* \cdot \quad 2} \tag{11}$$

According to (11), the marginal external costs of slow drivers are always greater than those of fast drivers, as long as no bottleneck congestion before the entrance occurs. This can be seen by rewriting the inequality that (11) be greater than 1 to $s_1/d^*>\rho_1+\rho_2$, and next to $d^*/s_1<1/(\rho_1+\rho_2)$. Since the no-queuing condition is $d^*/s_2<1/(\rho_1+\rho_2)$, and $s_2<s_1$ by definition, this condition is always satisfied. Therefore, (11) is always greater than 1, so that we can conclude that optimal tolls for slow drivers are always higher than for fast drivers.

Second-best tolls

The optimal common fee f_c for both groups can be derived according to equation (A7) in the appendix of Verhoef, Nijkamp and Rietveld (1995). The procedure is to set up a Lagrangean as follows:

$$\begin{aligned}\Lambda = &\int_0^{1} D_1(x_1)\,dx_1 + \int_0^{2} D_2(x_2)\,dx_2 - \quad 1 \cdot k_1(\quad 1,\quad 2) - \quad 2 \cdot k_2 \\ &+ \quad 1 \cdot \left(D_1(\quad 1) - k_1 - f_c\right) + \quad 2 \cdot \left(D_2(\quad 2) - k_2 - f_c\right)\end{aligned} \tag{12}$$

The set of first-order conditions for Λ with respect to ρ_1, ρ_2, λ_1, λ_2 and f_c can be solved to yield the following expression for the second-best optimal value for f_c:

$$f_c = \frac{\frac{1 \cdot \frac{k_1}{1}}{\frac{D_1}{1} - \frac{k_1}{1}} + \frac{1 \cdot \frac{k_1}{2}}{\frac{D_2}{2} + \frac{k_1}{2}}}{\frac{1}{\frac{D_1}{1} - \frac{k_1}{1}} + \frac{1}{\frac{D_2}{2} + \frac{k_1}{2}}} \tag{13}$$

An important characteristic of this second-best solution is that the Lagrangean multipliers λ_1 and λ_2 are generally different from zero, which they would be for first-best tolls in case these were derived in a way similar to (12). According to (13), f_c is a weighted average of the marginal external congestion costs for the two types of drivers in the second-best optimum as long as the two weights have the same sign. However, whereas the weight for group 1 is always negative, the weight for group 2 may actually turn out to be positive in sign. In that case, the weighted average property no longer holds. The second-best common fee may then exceed the marginal external congestion costs of group 2 (which are greater than those for group 1 by (11)), and may even approach plus infinity when the weights w_i of the two groups have the same absolute value, but opposite signs in the second-best optimum.

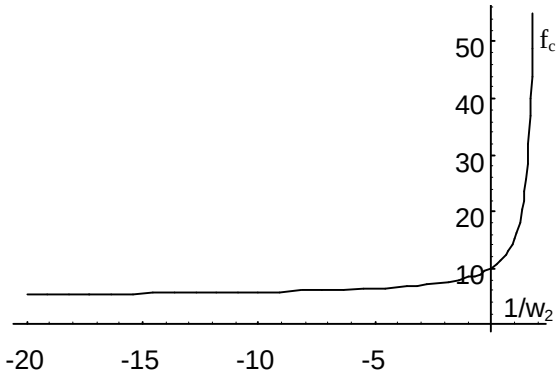

Figure 1. The optimal common fee as a function of w_2 for fixed marginal external costs and $1/w_1=-2$

Figure 1 illustrates these impacts of the weights on the second-best optimal value of the common fee. In order to concentrate on the impacts of the weights only, the 'marginal external costs' for both groups were fixed at a level of 5 for group 1 and 10 for group 2, whereas $1/w_1$ was fixed at –2 (note that the expression (13) implies that it is more natural to work with $1/w_i$ than with w_i). The diagram then shows f_c according to (13) for the value of $1/w_2$ depicted along the horizontal axis. Clearly, when $w_1 \approx -w_2$, a common fee is not a very attractive instrument, because of its socially unacceptable high level. It is, however, important to bear in mind that a second-best optimum near $w_1 \approx -w_2$ is very unlikely to occur. Both the marginal external costs and the weights are endogenous on road usage by both groups. Hence, an outcome where (13) takes on, for instance, a value approaching plus infinity with ρ_1 and ρ_2 having the values that actually produce this near-infinite second-best optimal tax will not occur as long as the demand for both groups is only slightly elastic. Moreover, if the demand for group 2 would approach complete inelasticity, w_2 would presumably be negative anyway, so that an outcome near $w_1 \approx -w_2$ is highly unlikely in that case. Therefore, it should be realized that (13) gives the expression for the second-best optimal common toll when evaluated in that second-best optimum.

A numerical example

It is instructive to use a numerical example to illustrate the comparative static properties of the model outlined above. For that purpose, we consider a road segment of 5 kilometres which is used by fast drivers who wish to drive at a speed of 100 km/h (27.8 m/s), and slow drivers who prefer 80 km/h (22.2 m/s). This could correspond to the type of road that are called 'autowegen' in The Netherlands. These roads usually have one lane for traffic in each direction, while overtaking is sometimes forbidden and, if not, often impossible due to traffic on the other lane. We assume that d^* has a value of 15 meters. The minimum travel time for slow drivers is then 225 seconds, and for fast drivers 180 seconds (the maximum travel time for fast drivers is 225.54 seconds; see (1b)). We set values of time of V_1=37 and V_2=65 (DFl/hr), which are in line with the values of time for the Netherlands

for passenger traffic (weighted over business travellers and commuters) and freight traffic, respectively (HCG; 1990, 1992).

Finally, we postulate two affine demand curves of the type $D_i(\rho_i)=d_i-a_i\cdot\rho_i$. For the base-case of the model, we set $d_1=5$, $a_1=24$, $d_2=10$ and $a_2=475$ (note that ρ_i is defined in terms of vehicles per second), which yields a non-intervention equilibrium in which 451 fast and 45 slow vehicles use the road per hour, $p_c=0.61$, $E(T_1)=194$, $k_1=1.99$ and $k_2=4.06$. In the optimum, $f_1=0.01$ and $f_2=1.45$; 455 fast and 34 slow vehicles use the road per hour; and generalized costs net of the fee are $k_1=1.96$ ($E(T_1)=190.5$) and $k_2=4.06$. Note that, in this case, in the optimum the expected travel time for fast drivers has only moderately decreased.

By varying the parameter a_2 (the slope of group 2's demand curve), the non-intervention and optimal equilibrium values of ρ_2 can be affected, so that the comparative static impacts of the equilibrium number of slow drivers can be traced. Figure 2 shows this. Along the horizontal axis, a_2 decreases each step by 20%, which causes the equilibrium number of slow vehicles per hour in the non-intervention case to increase from 4.8 up to 419.

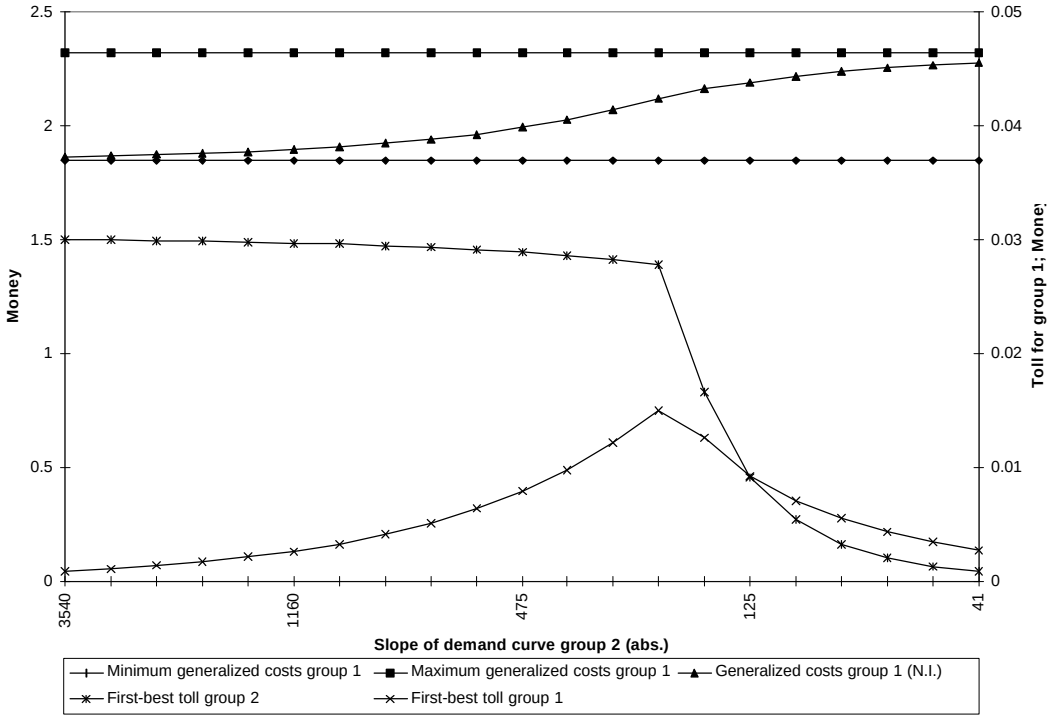

Figure 2. Expected travel costs for group 1 and optimal fees

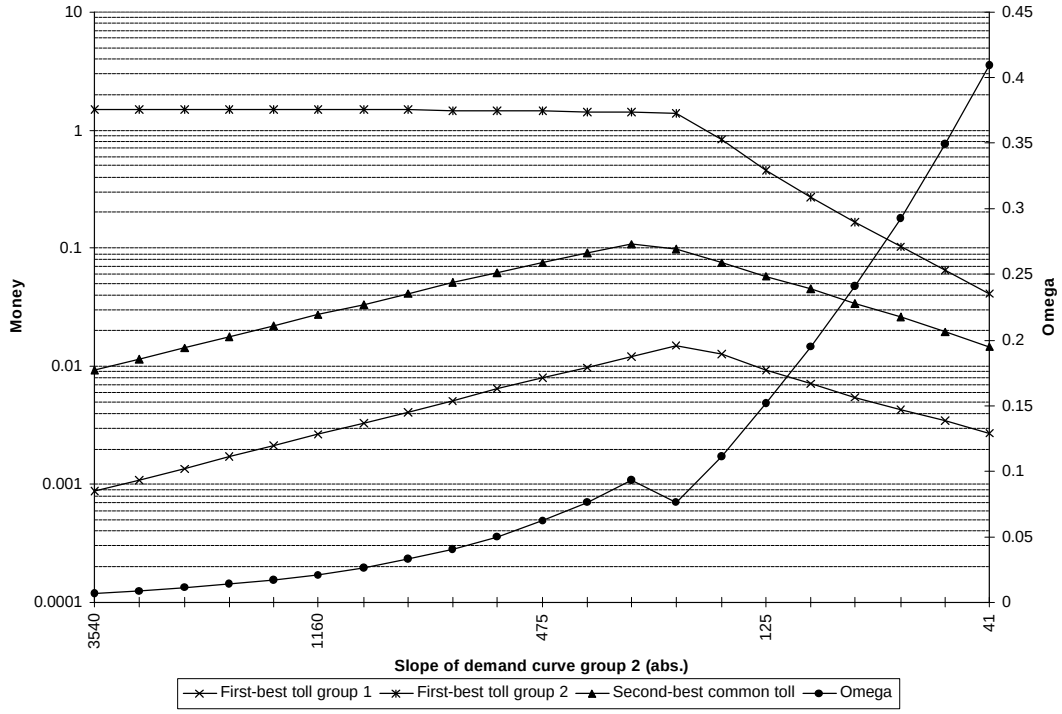

Figure 3. First-best and second-best fees, and the index of relative welfare improvement

The three upper curves show how, as a result, the non-intervention expected travel costs for group 1 k_1 increase from a level slightly above the minimum possible level (1.85) up to a level almost equal to the maximum possible level (2.31). The point of inflection is at the level of a_2 implying a level of ρ_2 beyond which p_c has switched from values below 1 to a value of 1 (this cannot be seen from the diagram, but was deduced from the simulation results).

Next, it can be seen that the optimal fee for the slow drivers f_2 is decreasing in the optimal ρ_2, as was claimed above. Moreover, the curve is sharply kinked at the point where p_c has switched from values below 1 to a value of 1. In principle, one could therefore expect situations where the demand curve $D_2(\rho_2)$ intersects $k_2+f_2(\rho_2)$ twice, in which case one has to check which of the two candidates for optima is the best. For the present simulation, these turned out to be the optima indicated. Finally, the course of f_1 can be understood by means of inspection of (11). Note that the value of f_1 is depicted along the axis on the right-hand side; f_1 is only a small fraction of f_2.

Figure 3 shows the second-best common fee f_c as derived above in (13). First of all, the figure repeats the patterns of f_1 and f_2 (note that the left axis is logarithmically scaled), and it can be seen that f_c is always between the values of f_1 and f_2. As soon as f_2 starts to decline, so does f_c. The ascending line shows the so-called 'index of relative welfare improvement' ω, which is defined as the ratio of the welfare gain that can be achieved with a second-best policy and the welfare gain that first-best regulation brings. Because the two first-best fees f_1 and f_2 approach each other when moving rightwards, ω increases – especially in the range where $p_c=1$, and f_2 drops sharply. The sudden drop in the ω-curve near this turning point is caused by the fact that over a certain limited range, first-best taxes still bring us in a regime where $p_c<1$, where second-best regulation – because of the lower fee for group 2 – already has $p_c=1$.

4. Platooning

Apart from tolling under the acceptance of the 'fact of life' that slow drivers appear at the entrance of the road and start their trips one-by-one, a quite different type of policy could be to impose what we will call 'platooning'. With this, we mean that slow drivers have to wait at the entrance of the road until a sufficiently large number of them has gathered, after which they are allowed to start their trips together, in a 'platoon'. Since this may drastically reduce the probability on congestion for the fast drivers, platooning may offer an interesting option to deal with congestion caused by speed differences.

Thinking about this possibility, one could actually envisage two such schemes. The first one involves 'unpredictable platooning' where it is not known beforehand exactly when a platoon would start, but it is only known that a platoon of a certain size π is required before the slow vehicles can start. The alternative is 'predictable platooning', in which case both slow drivers and fast drivers can anticipate the clock times at which platoons of slow drivers are allowed to start their trip. In this section, we only consider

'unpredictable platooning'. In that case, we can use V_2 as the value of waiting time for slow vehicles, and we can maintain the property that fast drivers behave the way they did in the previous section. For predictable platooning, it is likely that the value of waiting time for group 2 (now the time span between the moment a slow driver prefers to drive, and the moment his preferred platoon starts) becomes lower than V_2, since slow drivers will anticipate so that it is unlikely that a postponement of the trip would cost the same as when one would have to wait at the entrance of the road (note that this latter option is always possible). Furthermore, with predictable platooning, it is also unlikely that fast drivers would start their trip just after a platoon of slow drivers has started. This option certainly deserves attention in future work.

A first ingredient we need to study unpredictable platooning is the waiting time for slow vehicles. Denoting the platoon size as π, and assuming that the platoon starts as soon as the last of the π slow drivers has arrived, we know that one slow driver has waited 0 seconds, one $z_2=1/\rho_2$ seconds, one $2/\rho_2$ seconds, and so forth; and the most unlucky one $(\pi-1)/\rho_2$ seconds. From this series, an expected waiting time $(\pi-1)/(2\cdot\rho_2)$, and a total waiting time per platoon of $\pi\cdot(\pi-1)/(2\cdot\rho_2)$ can be inferred – where the latter implies total waiting time per unit of time equal to $\pi\cdot(\pi-1)/2$.

It is assumed that the slow drivers, when waiting, form a queue next to the road's entrance. Defining y now as the position of the platoon's first slow driver's front, the maximum travel time for a fast driver when y marginally exceeds 0 is now somewhat larger than in the case without platooning. The fast driver now first has to wait until all slow drivers in the platoon have passed the entrance, and the last one is at d^* meters from the entrance. This will take $\pi\cdot d^*/s_2$ seconds. Next, this fast driver has to drive $m-d^*$ seconds at speed s_2, and drives the last d^* meters at s_1. Hence, the maximum travel time for a fast driver is now $(m+(\pi-1)\cdot d^*)/s_2+d^*/s_1$.

Observing that $k(y)=\rho_1\cdot y/s_2$ still holds, and assuming that cumulative queuing at the entrance still does not occur – apart from queuing by slow vehicles waiting until π is realized, and the type of queuing by fast drivers described above (when the platoon of slow drivers starts) – we can now completely rework the analysis in the previous section allowing for a variable platoon size π. The relevant equations are given below. The travel time for a fast driver, and for a given y and k, becomes:

$$T_1(y) = \frac{m}{s_2} - \frac{y}{s_2}\cdot\left(1-\frac{d^*\cdot\ _1}{s_1}\right)+\frac{d^*}{s_1}+\frac{(\ \ -1)\cdot d^*}{s_2}$$
$$with: \frac{m}{s_1} \le T_1(y) < \frac{m+(\ \ -1)\cdot d^*}{s_2}+\frac{d^*}{s_1} \qquad (14)$$

The inequality defining whether a fast driver is hindered now is:

$$y < \frac{m\cdot(s_1-s_2)+s_2\cdot d^*+(\ \ -1)\cdot d^*}{s_1-d^*\cdot\ _1}$$

Since y can now take any value in the interval $<0,\pi\cdot s_2/\rho_2]$ with equal probability, we find the following expression for $p_c(\rho_1,\rho_2)$:

$$p_c(\ _1,\ _2)=1 \quad if \quad \frac{\cdot s_2}{\ _2} < \frac{m\cdot(s_1-s_2)+s_2\cdot d^*+(\ -1)\cdot s_1\cdot d^*}{s_1-d^*\cdot\ _1} \tag{15a}$$

$$p_c(\ _1,\ _2)=\frac{m\cdot(s_1-s_2)+s_2\cdot d^*+(\ -1)\cdot s_1\cdot d^*}{\left(s_1-d^*\cdot\ _1\right)\cdot\frac{\cdot s_2}{\ _2}} \quad otherwise \tag{15b}$$

The expected travel time for group 1 can then be derived in the same manner as before, and can be expressed as:

$$E(T_1)=(1-p_c(\ _1,\ _2))\cdot\frac{m}{s_1}+p_c(\ _1,\ _2)\cdot\tfrac{1}{2}\cdot\left(\frac{m}{s_1}+\frac{m+(\ -1)\cdot d^*}{s_2}+\frac{d^*}{s_1}\right) \quad if\ \ p_c<1 \tag{16a}$$

$$E(T_1)=\frac{m}{s_2}-\frac{}{2\cdot\ _2}\cdot\left(1-\frac{d^*\cdot\ _1}{s_1}\right)+\frac{d^*}{s_1}+\frac{(\ -1)\cdot d^*}{s_2} \quad if\ \ p_c=1 \tag{16b}$$

According to (15a), the set of parameter combinations for which $p_c=1$ is reduced roughly with a factor π (if d^* is sufficiently small compared to m), and according to (15b), p_c is reduced roughly by a factor π when $p_c<1$. Therefore, the expected travel costs for group 1 as given by (16a) and (16b) can be expected to be lower when $\pi>1$. In order to be more precise, we will treat π as a continuous variable and derive a first-order condition also for π for the maximization of social welfare. Since π will be discrete in reality, however, this particular first-order condition will of course have only limited practical relevance: when it implies a non-natural value for π, it suggests two possible optimal discrete values of π, and when it implies a value $\pi<1$, we know that the optimal π is 1. In the numerical example presented below, we will make sure that π only takes on natural values.

We will derive only a first-best solution, where optimal tolls and an optimal value of π are derived simultaneously. In order to decentralize the selection of π, the regulator can make the toll f_2 dependent on the platoon size. We then know that a platoon will depart as soon as for each platoon member, the benefits of waiting until the platoon size has increased by 1 in terms of a reduced toll $f_2(\pi)$ do not outweigh the cost of waiting $V_2\cdot z_2=V_2/\rho_2$. Recalling that the per unit of time total waiting costs for slow drivers queuing up in the platoon is $V_2\cdot(\pi-1)/2$, we can write the social welfare maximization problem as:

(17a)

(17b)

$$\underset{\pi_1,\pi_2,}{MAX}\ W = \int_0^{\pi_1} D_1(x_1)\,dx_1 + \int_0^{\pi_2} D_2(x_2)\,dx_2 - \pi_1 \cdot k_1(\pi_1, \pi_2, \ ) - \pi_2 \cdot k_2 - V_2 \cdot \frac{-1}{2}$$

s.t.:

$$D_1(\pi_1) - k_1 - f_1 = 0$$

$$D_2(\pi_2) - k_2 - f_2 - V_2 \cdot \frac{-1}{2 \cdot \pi_2} = 0 \qquad (17c)$$

$$\frac{f_2(\ )}{\ } + \frac{V_2}{\pi_2} = 0 \qquad (17d)$$

We maximize (17a) w.r.t. ρ_1, ρ_2, and π, and substitute the results in (17b-d) respectively to find the following optimal tax rules:

$$f_1 = \pi_1 \cdot \frac{k_1}{\pi_1} \qquad (18a)$$

$$f_2 = \pi_1 \cdot \frac{k_1}{\pi_2} - V_2 \cdot \frac{-1}{2 \cdot \pi_2} \qquad (18b)$$

$$\frac{f_2}{\ } = \frac{2 \cdot \pi_1 \cdot \frac{k_1}{\ }}{\pi_2} \qquad (18c)$$

For group 1, the expression for the optimal toll in (18a) is the same as (8a): fast drivers should face a toll equal to their marginal external congestion costs (note that, of course, the equilibrium values for (8a) and (18a) will not be the same when π>1). For the slow drivers, things have changed somewhat more fundamentally. First of all, the optimal fee f_2 is reduced by the value of the expected waiting time at the entrance according to (18b). This reflects that these waiting costs in a way form a part of the fee that the slow drivers face. As far as (18c) is concerned, if we would multiply both sides by ρ_2, the condition would show that the fee *per platoon* should decrease twice as quickly in π as the marginal effect of π on congestion costs for the fast drivers. The only thing that may seem puzzling about this is the factor 2. This factor can intuitively be explained by the fact that the growing platoon of slow drivers, when waiting to leave, themselves count *at each instant* in total $\pi \cdot V_2$ as the relevant waiting costs per unit of time, and are therefore inclined to leave earlier than when the total waiting time per unit of time of (π-1)/2 were considered.

The conditions (18abc) can again be made explicit by using:

$$\frac{c_1^{\,0}}{\pi_1} = \frac{p_c}{\pi_1} \cdot \left[V_1 \cdot \tfrac{1}{2} \cdot \left(\frac{m + (\ -1) \cdot d^*}{s_2} + \frac{d^*}{s_2} - \frac{m}{s_1} \right) \right] \qquad (19a)$$

$$= \frac{d^* \cdot \left(m \cdot (s_1 - s_2) + s_2 \cdot d^* + (\ -1) \cdot d^* \cdot s_1 \right)}{\cdot \left(s_1 - d^* \cdot \pi_1 \right)^2 \cdot \frac{s_2}{\pi_2}} \cdot \left[V_1 \cdot \tfrac{1}{2} \cdot \left(\frac{m + (\ -1) \cdot d^*}{s_2} + \frac{d^*}{s_2} - \frac{m}{s_1} \right) \right]$$

(19b)

$$\frac{c_1^{\ 0}}{\ _2} = \frac{p_c}{\ _2} \cdot \left[V_1 \cdot \tfrac{1}{2} \cdot \left(\frac{m + (\ \ -1) \cdot d^*}{s_2} + \frac{d^*}{s_2} - \frac{m}{s_1} \right) \right]$$

$$= \frac{m \cdot (s_1 - s_2) + s_2 \cdot d^* + (\ \ -1) \cdot d^*}{\cdot \left(s_1 - d^* \cdot \ _1 \right) \cdot s_2} \cdot \left[V_1 \cdot \tfrac{1}{2} \cdot \left(\frac{m + (\ \ -1) \cdot d^*}{s_2} + \frac{d^*}{s_2} - \frac{m}{s_1} \right) \right]$$

$$\frac{c_1^{\ 0}}{\ _2} = \frac{p_c}{\ _2} \cdot \left[V_1 \cdot \tfrac{1}{2} \cdot \left(\frac{m + (\ \ -1) \cdot d^*}{s_2} + \frac{d^*}{s_2} - \frac{m}{s_1} \right) \right] + \frac{V_1 \cdot p_c \cdot d^*}{2 \cdot s_2} \tag{19c}$$

$$= \frac{-\left((m - d^*) \cdot (s_1 - s_2) \right) \cdot \frac{s_2}{\ _2} \cdot \left(s_1 - \ _1 \cdot d^* \right)}{\left(\left(s_1 - d^* \cdot \ _1 \right) \cdot \ \cdot s_2 / \ _2 \right)^2} \cdot \left[V_1 \cdot \tfrac{1}{2} \cdot \left(\frac{m + (\ \ -1) \cdot d^*}{s_2} + \frac{d^*}{s_2} - \frac{m}{s_1} \right) \right]$$

$$+ \frac{V_1 \cdot p_c \cdot d^*}{2 \cdot s_2}$$

$$\frac{c_1^{\ 1}}{\ _1} = V_1 \cdot \frac{\cdot d^*}{2 \cdot \ _2 \cdot s_1} \tag{20a}$$

$$\frac{c_1^{\ 1}}{\ _2} = V_1 \cdot \frac{\cdot \left(s_1 - d^* \cdot \ _1 \right)}{2 \cdot \ _2^{\ 2} \cdot s_1} \tag{20b}$$

$$\frac{c_1^{\ 1}}{\ } = \frac{s_1 - d^* \cdot \ _1}{2 \cdot \ _2 \cdot s_1} + \frac{d^*}{s_2} \tag{20c}$$

For cases where $p_c<1$, the marginal external costs for fast and slow drivers have roughly decreased by a factor π (ignoring the impact on the expected travel costs in case a fast driver does experience congestion in this situation). For $\pi=1$, the opposite has occurred, showing that if $p_c=1$ and every fast driver is going to experience congestion anyway, the marginal external costs increase linearly with π. This seems worrying at first sight. However, note that (20c) is always positive. This implies that an optimal platoon size larger than 1 can never be found in a region with $p_c=1$, since the first-order condition for (17a) w.r.t. π dictates that $\partial k_1/\partial\pi$ be negative. Hence, we can expect an optimal value $\pi>1$ if this produces an optimum with $p_c<1$ with substitution of (19c) into (18c) producing an expression $\partial f_2/\partial\pi<0$. Any optimum where $p_c=1$ can only be consistent with $\pi=1$, since the objective (17a) in that case is decreasing in π.

A numerical example

Based on the numerical example in the previous section, we now present some simulation results involving optimal platooning. On intuitive grounds, it is evident that especially the

value of time for slow drivers will be an important determinant for the desirability of platooning, since their time, in a way, is 'offered' in favour of improved traffic conditions for fast drivers. Therefore, we chose to vary V_2 to give an illustration of the impacts of platooning. All other variables were kept at their base-case values as described in the previous section.

Figure 4 shows the optimal platoon size as a function of V_2. For lower values of time for slow drivers, on the left-hand size of the figure, the optimal platoon size is relatively large. This is caused, in the present simulation, by two effects. First, the non-intervention usage by slow drivers increases because their travel costs decrease due their lower value of time. This makes platooning more attractive, because waiting times decrease. Secondly, the cost of waiting of course also decreases. This latter effect also plays an important role on the left-hand side of Figure 4, since the non intervention level of ρ_2 at V_2=0.0003 is less than 3 times as high as at V_2=102, whereas the optimal platoon size is 13 times as high. The index of relative welfare improvement ω shows the effect of 'first-best' tolls without platooning, as derived in Section 3. Clearly, π=1 indeed turns out to be optimal for the base-case where V_2=65. Platooning becomes increasingly desirable for low values of time for the slow drivers, which will generally not be the case when these drivers are truckers, but especially when slow drivers make their trips for touristic or social purposes.

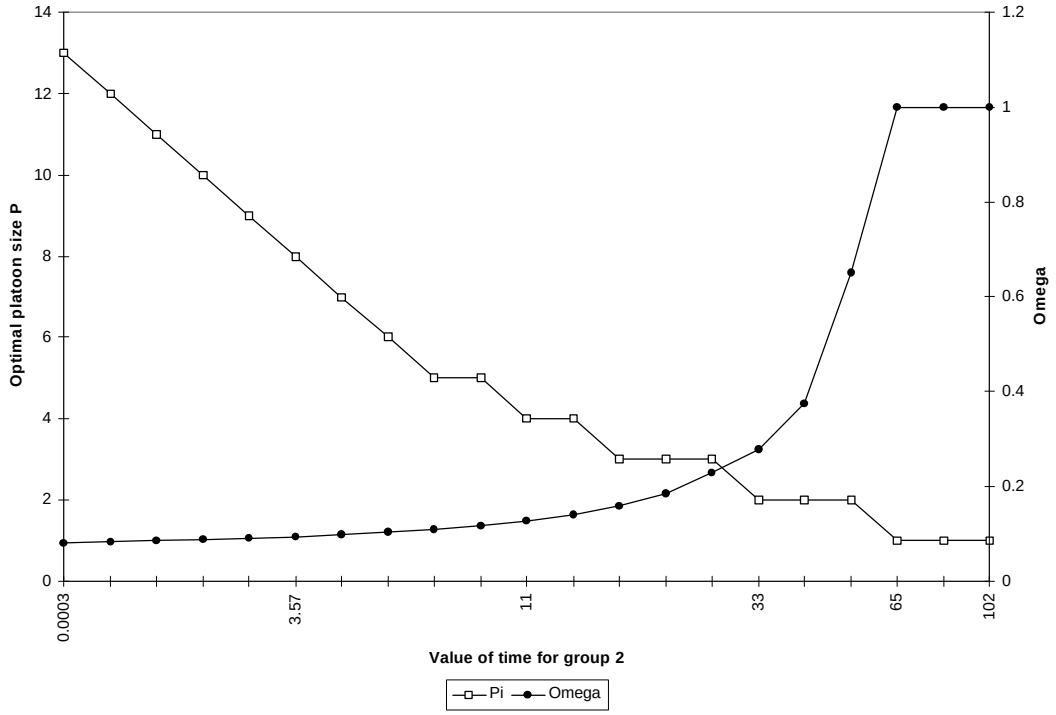

Figure 4. Optimal platoon size and index of relative welfare improvement for non-platooning

Finally, Figure 5 shows that actually both groups (slow drivers and fast drivers) will benefit from platooning when this is socially optimal. The figure shows the (expected) generalized travel costs for both groups, including tolls and waiting time for slow drivers, with and without platooning. In the latter case, these generalized costs are lower for both groups when $\pi>1$. Note that the generalized costs for both groups, in particular for fast drivers, do not follow smooth paths in Figure 5. This is caused by regime shifts when π takes on a different discrete value.

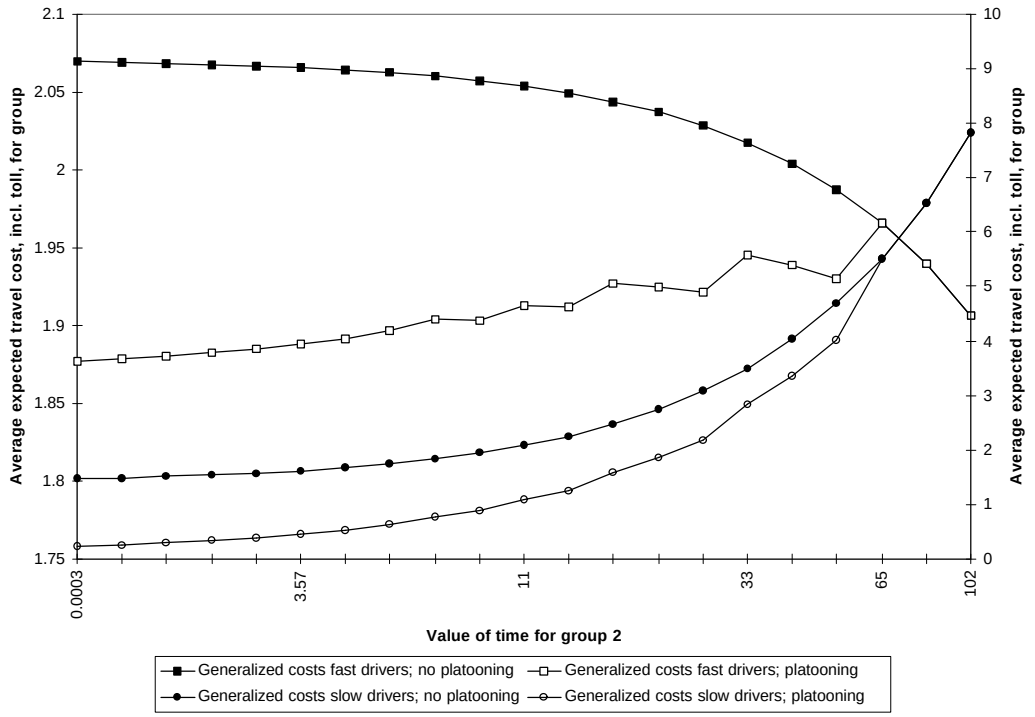

Figure 5. Generalized costs, including tolls, for fast and slow drivers with and without platooning

5. Conclusion

In this paper, we investigated congestion caused by differences in desired or possible speeds. Especially outside peak hours, speed differences are probably one of the most important reasons for congestion. Although the model setting, with one lane and no overtaking, seems simple at first sight, the problem turned out to result easily in quite complicated mathematical expressions. Some main conclusions are that optimal tolls for slow vehicles are higher than those for fast drivers, that the marginal external costs and the optimal tolls for slow drivers are *decreasing* in the equilibrium number of slow drivers, and that 'platooning' may become an attractive option especially when the desire for a low speed is caused by a lower value of time.

Since the purpose of the model was to outline the fundamentals of congestion caused by two speeds, a number of simplifying assumptions were made that should be relaxed in future work. Two important ones are the no-overtaking condition and the one-lane assumption. As far as the latter assumption is concerned, an interesting question to address would be the conditions under which it becomes attractive to use designated lanes for specific types of traffic. Other applications and extensions we have in mind is the question of optimal (maximum and minimum) speed limits, and applications of the model to rail transport.

References

Arnott, R., A. de Palma and R. Lindsey (1993) "A structural model of peak-period congestion: a traffic bottleneck with elastic demand" *American Economic Review* **83** (1) 161-179.

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

Fowles, R. and P.D. Loeb (1989) "Speeding, coordination and the 55 MPH limit: comment" *American Economic Review* **79** 916-921.

Lave, C.A. (1985) "Speeding, coordination and the 55 MPH limit" *American Economic Review* **75** 1159-1164.

Lave, C.A. (1989) "Speeding, coordination and the 55 MPH limit: reply" *American Economic Review* **79** 926-931.

Lee, D.F. (1985) "Policing cost, evasion cost, and the optimal speed limit" *Southern Economic Journal* **52** 34-45.

Levy, D.T. and P. Asch (1989) "Speeding, coordination and the 55 MPH limit: comment" *American Economic Review* **79** 913-915.

Rodriguez, R.J. (1990) "Speed, speed diversion, and the highway fatality rate" *Southern Economic Journal* **57** 349-356.

Rotemberg, J. (1985) "The efficiency of equilibrium traffic flows" *Journal of Public Economics* **26** 191-206.

Rienstra, S.A. and P. Rietveld (1996) "Speed behaviour of car drivers: a statistical analysis of acceptance of changes in speed policies in The Netherlands" *Transportation Research* **1D** 97-110.

Rouwendal, J., E.T. Verhoef, P. Rietveld and B. Zwart (1997) "Speed differences and congestion". Mimeo, University of Wageningen.

Snyder, D. (1989) "Speeding, coordination and the 55 MPH limit: comment" *American Economic Review* **79** 922-925.

Tzedakis, A. (1980) "Different vehicle speeds and congestion costs" *Journal of Transport Economics and Policy* **14** 81-103.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1995) "Second-best regulation of road transport externalities" *Journal of Transport Economics and Policy* **29** 147-167.