Journal of Transport Economics and Policy, Volume 36, Part 3, September 2002, pp. 407–445

A Stochastic Model of Congestion Caused by Speed Differences

Jan Rouwendal, Erik Verhoef, Piet Rietveld and Bert Zwart

Address for correspondence: Jan Rouwendal, Department of Social Sciences, Wageningen University, PO Box 8130, 6700 DA Wageningen, Netherlands. Erik Verhoef and Piet Rietveld are at the Department of Spatial Economics, Free University, Amsterdam; Bert Zwart is at the Department of Mathematics and Computer Science, Eindhoven University of Technology. The research of Erik Verhoef has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. Comments by Richard Arnott, Robin Lindsey, and Ken Small, and two anonymous referees, are gratefully acknowledged. The usual disclaimer applies

Abstract

The authors study interaction on a two-lane road between the trips of two types of drivers who differ by their desired speeds. The difference in desired speeds causes congestion, because slow vehicles force fast vehicles to reduce their speed. Results for this type of congestion with respect to tolling are very different from those of the classic Pigou–Knight model, where the marginal external costs are an increasing function of the number of road users. In our model we find the opposite result: the marginal external costs of slow vehicles are a decreasing function of the number of slow vehicles. This leads to rather different policy recommendations.

Date of receipt of final manuscript: September 2001

Introduction

In this paper we study what may be considered to be the most elementary form of traffic congestion: vehicles that are unable to maintain the desired speed because of slower traffic in front of them. This form of congestion is well known to each car driver. It is also relatively easy to model in a structural way, because the causal relations involved are simple. Many other models of congestion (the bottleneck model of Arnott *et al.* (1993) is an important exception) use a relationship between demand (or traffic flow or density) and travel time (or speed) of a reduced form character, usually the "fundamental diagram." We prefer to model the mechanism that leads to congestion explicitly. This implies that the impacts of slow vehicles on fast vehicles have to be modelled.

Tzedakis (1980) is, as far as we know, the first study of congestion caused by different vehicle speeds.[1] His study refers to a single lane where traffic enters with stochastic arrival times. In Verhoef, Rouwendal and Rietveld (1999) we studied congestion caused by speed differences without overtaking in a situation with deterministic arrival times.[2] Although deterministic arrival times have some advantages for elaborating the model, the realism of this assumption is, of course, questionable. Even for rail traffic with a fixed schedule of trains, there are random delays. Huisman and Boucherie (2001) have recently studied the implications of this for actual running times on railway sections with two types of trains (with different speeds for each type), although with different methods from those we adopt here. We have not been able to find something like an economic analysis of overtaking on two-lane roads. We have only one example of a study on platoon formation on two-lane roads, Barzily and Rubinovitch (1979) who develop a model in which road sections where overtaking is either completely unrestricted or impossible alternate. In fact they model the situation as a three-lane highway in which the lane in the middle is divided into sections that are alternately reserved for traffic in both directions. This study does not give an economic analysis of the congestion implied by platoon formation.

Newell (1998) has studied the effects of a "moving bottleneck"; that is, of a slow vehicle on a road used by vehicles that drive faster. His

[1] Wardrop's (1952) classic paper discusses speed differences and the associated necessity of frequent overtaking. It suggests using the ratio of actual to desired overtakings as a measure of congestion, but does not analyse this type of congestion itself.

[2] A comparison between the results of that paper and those obtained in the present is provided at the end of the next section.

analysis differs from ours in that he considers a situation in which two lanes are available for traffic moving in a single direction, so that overtaking is always possible. Moreover, he concentrates on the effects of the higher density of fast vehicles on the lane used for overtaking the slow vehicle on the average speed of the fast vehicles, whereas we will assume that cars will have constant speeds independent of traffic density, unless the distance to its immediate predecessor becomes smaller than a threshold value. We will return to Newell's analysis in the final section of this paper.

In this paper we study traffic on a road used by two types of vehicles that are identified by their desired speeds. Slow vehicles have a lower desired speed than fast vehicles and the presence of the former type of vehicle causes congestion for the latter type. The extent of congestion in this context, and various policy measures to deal with it, are the subject of this paper.

The situation studied in the paper is a two-lane road with cars driving in both directions. It is assumed that the same types of vehicle are present on both lanes, although possibly in different numbers.

We start our analysis by considering the situation in which overtaking is impossible or prohibited and introduce the possibility of overtaking later on. It will be argued in the sections that follow that tolling is hardly a useful policy for traffic control under these circumstances and that other measures (such as prohibiting slow vehicles to enter some roads) should be considered instead.

An obvious, but costly, possibility to relieve congestion problems of the type considered here would be to make a second lane available for traffic in each direction. Except for very high levels of traffic demand, all congestion will then disappear if there are only two types of vehicles. In reality the second lane is often the one used by the traffic proceeding in the opposite direction, and this implies that it is only available for overtaking if sufficiently large gaps occur in oncoming traffic. This situation is studied in the second part of the paper.

Throughout the paper we refer to vehicles as the actors that determine the traffic phenomena in which we are interested. The term vehicle should be interpreted as referring to the vehicle as well as to its driver, while the latter is usually not mentioned explicitly for the sake of brevity. By adopting this convention we attempt to bring out clearly that traffic is not only determined by drivers' characteristics, but also by vehicles' performances. For instance, a truck driver may have a very high desired speed, but be unable to reach it because of the characteristics of the vehicle. In such a case the truck may have a speed that is lower than that of other cars and will be referred to as a slow vehicle.

Overtaking Impossible or Prohibited

Preliminaries

We consider traffic on a road segment of finite, but arbitrary, length l. This road segment has two lanes and traffic moving in opposite directions. Overtaking may either be prohibited or impossible (for example, because the lanes are physically separated). The next section will consider the situation in which overtaking is possible if traffic on the other lane allows a driver to do so. In the present section attention may be limited to vehicles on a single lane.

There are two types of vehicles on the road: those belonging to the first group have a preferred speed s_1 and those belonging to the second group a preferred speed s_2. We assume that $s_1 > s_2$ and will often refer to vehicles in the first group as "fast" and to vehicles in the second as "slow." We assume that both types of vehicle choose their preferred speeds unless the circumstances on the road force them to reduce it. In general one would expect speeds to be chosen endogenously on the basis of travel conditions. In situations without congestion vehicles have a preferred speed that depends on the conditions of the road. Our analysis assumes homogeneous travel conditions, and hence a constant speed for both types of vehicle. Since there are only two types of vehicle, slow vehicles can always use their preferred speeds and it seems natural to assume that they will do so.[3] For fast vehicles the situation is somewhat different. If they anticipate the necessity to slow down because of a slow driver in front of them, they may decide to slow down gradually instead of maintaining their own preferred speed as long as possible. However, even though incorporating this behaviour into our model may result in a more realistic picture of actual traffic, it will not change the results about congestion as long as the fast vehicles are as close as possible behind the slow vehicles when leaving the road segment that we study.[4] The total travel time of fast vehicles does not depend on the way they reduce their speed when they are forced to reduce speed because of a slow vehicle in front of them.

[3] If the risk of accidents influences the preferred speed, one can imagine that the presence of fast vehicles will reduce their speed even more. However, this may increase the risk of accidents and a more rational reaction may be that slow vehicles *increase* their preferred speed. Since an analysis of this issue would take us too far from the scope of the present paper, we assume a preferred speed that is independent of traffic conditions.

[4] Fast vehicles have an incentive to do so if the end of the road segment coincides with a junction, or a possibility for overtaking that would give them an opportunity to return to their higher preferred speed.

Demand for trips on the road segment, expressed in the number of cars per time unit, will be denoted as μ_i, $i = 1, 2$. The costs associated with a trip consist of expected travel time w_i and possibly a toll z_i. It seems natural to assume that for both types of vehicles demand is a decreasing function of the generalised travel cost, that is, the sum of the monetary value of expected travel time w_i and a possible toll z_i. The inverse demand function g_i, $i = 1, 2$ can then be written as:

$$g_i(\mu_i) = v_i w_i + z_i, \qquad i = 1, 2. \tag{1}$$

where v_i is group i's valuation of travel time, which is assumed to be a strictly positive number.

In this section we consider the situation in which traffic is restricted to a single lane. As soon as a fast driver is forced to slow down, he drives at the low speed of his leader until the end of the road segment.

The demand functions will be used to derive optimal tolls, but before we are able to do this, we first have to consider how travel times are related to demand.

Minimum distance

We assume that a minimum distance d^* must be maintained between the fronts of subsequent cars on the same road. If the distance between two subsequent cars becomes equal to d^* the follower will reduce its speed to that of the leader. In the situations to be considered in this paper, the close proximity of a slow driver in front of a fast driver is the only reason why fast vehicles reduce speed. Slow vehicles are always able to maintain their preferred speed.

Capacity of the road

If car A enters the road segment, a following car, car B, can enter only after car A has moved d^* metres on the road segment. If car A has a slow driver, this takes d^*/s_2 time units, if it has a fast driver d^*/s_1 units. In order to keep the presentation simple, we assume that always d^*/s_2 seconds pass before another car enters the road segment.[5] The

[5] Robin Lindsey informed us that the California Uniform Vehicle Code stipulates that vehicles maintain a two-second headway behind the vehicle ahead of them. This would justify our assumption.

maximum capacity c of the road segment, defined in this way, is therefore equal to:

$$c = 1/(d^*/s_2)$$
$$= s_2/d^*. \tag{2}$$

It is assumed throughout the paper that the demand for trips does not exceed the capacity of the road.

Arrivals

During the first d^*/s_2 seconds after entrance of a car, the arrival rate equals 0, but otherwise it takes on the constant value λ.[6] The density function of the additional time t until arrival of the next car is therefore $f(t) = \lambda \exp(-\lambda t)$, with expectation $1/\lambda$. λ is the sum of the arrival rates of fast and slow cars, denoted as λ_1 and λ_2.

The arrival rates λ_i are of course related to demand μ_i. In fact the arrival rates would be equal to the demand volumes if the requirement of a minimum time between the arrival of subsequent cars had not been made. Because cars cannot enter arbitrarily soon after each other, the expected number of cars entering the road segment is smaller than $\lambda_1 + \lambda_2$, implying that the arrival rate must be higher than $\mu_1 + \mu_2$.

In order to derive the relationship between arrival rates and demand, we set demand equal to the expected number of arrivals. The expected value of the time that elapses between the entrance of subsequent vehicles is $d^*/s_2 + 1/\lambda$, and the expected number of vehicles $E(n)$ that enter during one time unit is therefore equal to $E(n) = \lambda/(\lambda d^*/s_2 + 1)$.

In order to make this number equal to $\mu_1 + \mu_2$, λ should be equal to:

$$\lambda = \frac{\mu_1 + \mu_2}{1 - (\mu_1 + \mu_2)/c}. \tag{3}$$

If the capacity of the road is large, λ will be close to $\mu_1 + \mu_2$, as one should expect.

For each car that enters the road there is fixed probability $p = \mu_1/(\mu_1 + \mu_2)$ that it is fast, and a complementary probability $(1 - p)$ that it is slow. The arrival rates of both types of vehicles are therefore:

$$\lambda_i = \frac{\mu_i}{1 - (\mu_1 + \mu_2)/c} \qquad i = 1, 2. \tag{4}$$

[6] Alternatively, we could have assumed a constant arrival rate, in combination with the possible existence of a queue on the on ramp. This would complicate our derivations, but would probably not change the results. We have therefore adopted the simpler approach.

Travel times

The main interest of this paper is in congestion, and therefore in the time needed to travel across the road segment. For a slow driver this time is always equal to:

$$w_2 = \frac{l}{s_2}. \tag{5}$$

Clearly, slow vehicles do not experience congestion. Fast vehicles may experience congestion because of the presence of slow vehicles on the road segment at the time they travel it. Since a stochastic process determines the arrivals of both fast and slow vehicles, travel time for the fast vehicles is a random variable. It is shown in the Appendix that its expected value is equal to:

$$w_1 = \frac{l}{s_2} - \frac{1}{\lambda_2}\left(1 - \exp\left(-\lambda_2\left(\frac{l}{s_2} - \frac{l}{s_1}\right)\right)\right). \tag{6}$$

It is immediately clear from this formula that expected travel time for the fast vehicles is at most equal to l/s_2, the travel time for the slow vehicles. It will, of course, only reach this limit if every fast driver is forced to slow down to the lower speed immediately after entering the road. This happens when λ_2 becomes very large and the second term on the right hand side of the formula vanishes.

It follows also from the formula that expected travel time approaches its upper bound $l/s_2 - 1/\lambda_2$ if the length of the road becomes very large. The reason is that on a very long road a fast car will usually be forced to slow down by a slow car when a small part of the road has been travelled. (This small part may, of course, be a large number of kilometres.)

Although it is not obvious from the formula, it can be shown (see the Appendix) that for $\lambda_2 \to 0$ expected travel time for fast vehicles becomes equal to l/s_1, the travel time of fast vehicles who experience no congestion, as should be expected.

Expected travel time is an increasing function of the arrival rate λ_2 as should also be expected. However, as shown in the Appendix, travel time is a *concave* function of the arrival rate. This is worthy to be stressed, since it has important consequences for tolling. The economic analysis of traffic congestion in the tradition of Pigou and Knight assumes a *convex* relation between traffic demand and travel time.

It is also noteworthy that the arrival rate for fast vehicles, λ_1, does not appear in the formula for the travel times of these vehicles. Neither the travel time of the fast vehicles nor that of the slow vehicles depends on the arrival rate of fast vehicles. This suggests that optimal tolling would require a toll for slow vehicles only. Although it will be shown below that

this is actually not completely true, it is usually a good approximation to the truth.

Demand and travel time

Substitution of equation (4) (for $i = 2$) into (6) gives the desired relation between the travel time of fast vehicles and demands:

$$w_1 = \frac{l}{s_2} - \frac{1-(\mu_1+\mu_2)/c}{\mu_2}\left(1-\exp\left(-\frac{\mu_2}{1-(\mu_1+\mu_2)/c}\left(\frac{l}{s_2}-\frac{l}{s_1}\right)\right)\right). \quad (7)$$

It is clear from this formula that the travel time of fast cars on the road segment depends on the demand for trips of slow vehicles *as well as* on that of fast vehicles.

If total demand approaches the capacity c, both arrival rates λ_1 and λ_2 will become very large and the expected travel time close to its maximum value l/s_2.

Optimal tolls

Optimal tolls can be derived by maximising the social surplus, that is, the sum of the consumer surpluses of both types of vehicle and the toll revenues under appropriate side conditions (see the Appendix, where the possibility of boundary solutions is also discussed). If an interior solution is relevant, the optimal tolls satisfy the following equations:

$$\begin{aligned} z_1 &= \mu_1 v_1 \frac{\partial w_1}{\partial \mu_1}, \\ z_2 &= \mu_1 v_1 \frac{\partial w_1}{\partial \mu_2}. \end{aligned} \quad (8)$$

These formulas are not too surprising, but their elaboration is more interesting. For the purpose of computing the optimal tolls it is more convenient to use the alternative, but equivalent formulas that start from equation (6) instead of (7) and use the chain rule to complete the derivation:

$$\begin{aligned} z_1 &= \mu_1 v_1 \frac{\partial w_1}{\partial \lambda_2}\frac{\partial \lambda_2}{\partial \mu_1}, \\ z_2 &= \mu_1 v_1 \frac{\partial w_1}{\partial \lambda_2}\frac{\partial \lambda_2}{\partial \mu_2}. \end{aligned} \quad (9)$$

For positive demands of both types of vehicles, all partial derivatives appearing in equation (9) are positive, and so are the tolls.

Further elaboration allows one to write the optimal tolls as:

$$z_1 = v_1 \frac{\mu_1}{\mu_2} \frac{1}{c} \left(1 - \left(1 + \frac{\mu_2}{1 - \frac{\mu_1 + \mu_2}{c}} \left(\frac{l}{s_2} - \frac{l}{s_1} \right) \right) \right.$$
$$\left. \times \exp\left(- \frac{\mu_2}{1 - \frac{\mu_1 + \mu_2}{c}} \left(\frac{l}{s_2} - \frac{l}{s_1} \right) \right) \right) \tag{10}$$
$$z_2 = z_1 \left(1 + \frac{c - (\mu_1 + \mu_2)}{\mu_2} \right).$$

From the way these equations are written down, it is clear that the toll for slow vehicles is always higher than that for fast vehicles. The difference will be small if demand for trips by slow vehicles is large, and the difference will be large if demand for trips by slow vehicles is small. In particular, it can be shown that:

$$\begin{aligned} &\lim_{\mu_2 \to 0} z_1 = 0 \quad \text{if } \mu_1 > 0, \\ &\lim_{\mu_2 \to 0} z_2 = \infty \quad \text{if } \mu_1 > 0. \end{aligned} \tag{11}$$

This implies an *inverse* relationship between the level of the congestion toll for slow vehicles and the level of congestion itself, which is at variance with the results from the Pigou–Knight analysis. The reason for the inverse relationship is the fact that a fast driver who is forced to slow down by a slow driver is, from that moment on, insensitive to the presence of more slow vehicles on the road. Additional slow vehicles can only hinder fast vehicles who are not already forced to slow down by the presence of the slow vehicles already using the road. Since the number of "free flowing" fast vehicles is a decreasing function of the number of slow vehicles, the additional congestion caused by an extra slow driver will decrease. An example will clarify this.

Illustration

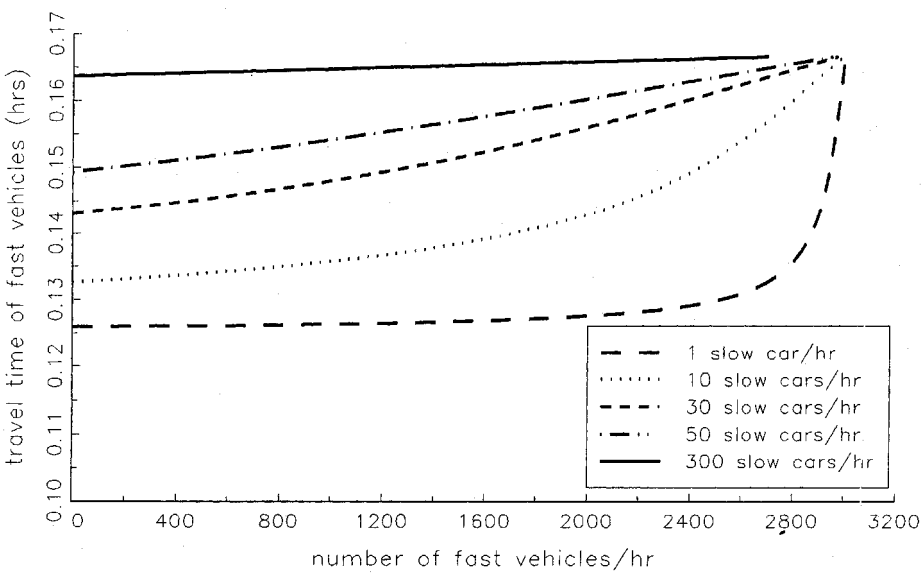
Consider a road with a length of 10 kilometres where fast vehicles want to drive at 80 kilometres per hour, and slow vehicles at 60 kilometres per hour. The minimum required distance between (the noses of) subsequent vehicles is 20 metres, which implies that capacity equals 3,000 cars per hour. Such a road may be thought of as one connecting two villages in a rural area. If overtaking is impossible or prohibited on that road, expected

travel time for fast vehicles is given by equation (6). The travel time for slow cars on this road is equal to 10 minutes (0.166 hours).

Figure 1 shows the relations between the travel time of fast vehicles and the volume of their demand for trips on the road. The lowest line in that figure refers to a situation in which the demand for trips by slow vehicles is very low: 1 car per hour. The figure shows that in that case the curve is similar to the ones usually presented in the Pigou–Knight analysis of congestion. Travel time is close to its minimum value unless the number of fast vehicles approaches the capacity sufficiently closely.

Figure 1

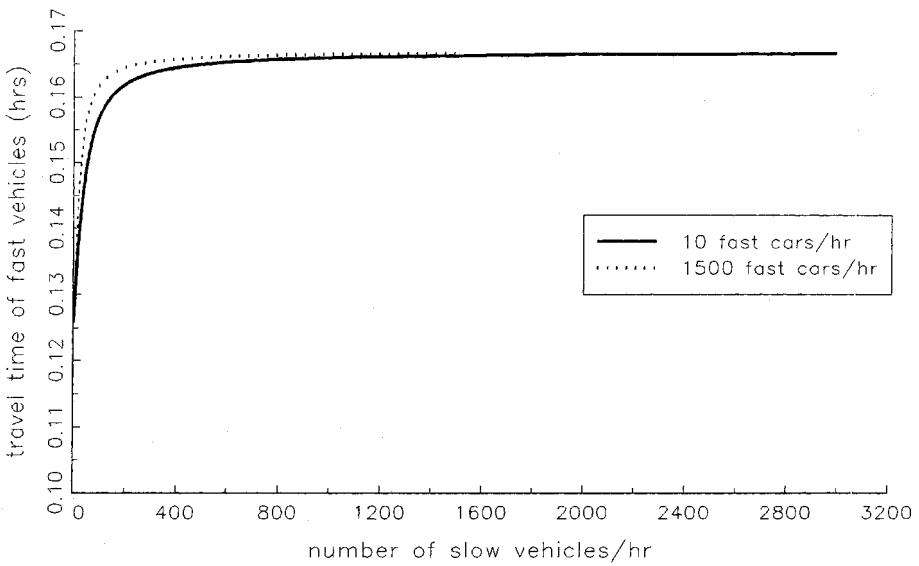
The relationship between the number of fast vehicles and their travel time for various numbers of slow drivers



The figure also shows that this relationship changes remarkably for even moderately high levels of demand for trips by slow vehicles. Travel time of fast vehicles becomes an almost linear function of their demand for trips that flattens out for higher levels of demand by slow vehicles. This suggests that it will be interesting to look also to the relationship between the travel time of fast vehicles and the demand for trips by slow vehicles.

Figure 2 depicts that relation. One line in the picture refers to a situation in which demand of fast vehicles is very low (10 drivers per hour), and demand by slow vehicles can vary over almost the full range from 0 to 3,000. The figure shows that travel time of fast vehicles is, in this case, a steeply increasing function of the demand by slow vehicles as long as this

Figure 2
The relationship between the number of slow vehicles and the travel time of fast vehicles for various numbers of fast vehicles



demand is small. For higher volumes of demand by slow vehicles, travel time of fast vehicles is almost equal to that of slow vehicles (one sixth of an hour) and insensitive to further increases in the demand by slow vehicles.

Similar lines can be drawn for other volumes of the demand for trips by fast vehicles. These other lines are all remarkably similar to the one shown in the figure. The most important difference is that they end at an earlier level of demand by slow vehicles, because of the capacity constraint. In Figure 2 a second line is drawn that is based on demand by fast vehicles that equals 1,500. It appears, therefore, from the figure that the travel time of fast vehicles is hardly dependent on the demand for trips by either fast or slow vehicles, unless the latter demand is small, that is, in the range of 0 to 400 cars per hour. Note the concavity of the lines drawn in the figure.

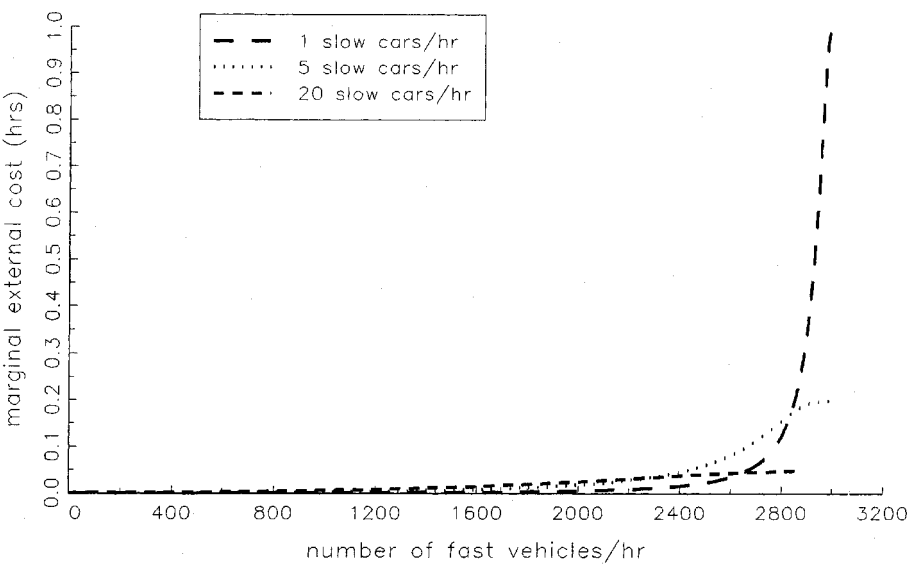
The well known diagram on which the conventional Pigou–Knight analysis is based is sometimes interpreted as having *average* travel time on its vertical axis and the number of vehicles per hour that use the road on its horizontal axis. This could give the impression that the comparison between the Pigou–Knight model and the one presented in the present paper would be facilitated by drawing a diagram that gives the average travel time of fast and slow vehicles that use the road as a function of their total number. This is easily done when keeping the fractions of both types of drivers constant, and the resulting concave lines have shapes that are

very similar to those drawn in figure 2,[7] and have therefore not been drawn separately. However, it should be noted that the fractions of fast and slow drivers should in reality be expected to depend on the travel time itself. In particular, one would expect the fraction of fast drivers to be smaller if travel time is relatively large, unless travel demand by fast vehicles would be completely inelastic.[8] The construction of such a picture would therefore require, in the context of the present paper, the development of an equilibrium analysis.

Figures 3 and 4 show the marginal external costs, expressed in time units, for fast and slow vehicles as a function of the volume of their demand, for various levels of the demand by the alternative type of vehicles. The marginal external costs are equal to the optimal tolls if the corresponding levels of demand would be the equilibrium levels. Figure 3 shows the pattern that is well known from the Pigou–Knight analysis when the number of slow vehicles is very low (1 driver per hour). For higher volumes of demand by slow vehicles, optimal tolls for fast vehicles become very small.

Figure 3

The relationship between the number of fast vehicles and their marginal external cost for various numbers of slow vehicles



[7] This is explained by the similarity of the curves shown in Figure 5 for different numbers of fast cars and by the fact that the travel time of slow vehicles is independent of traffic conditions.

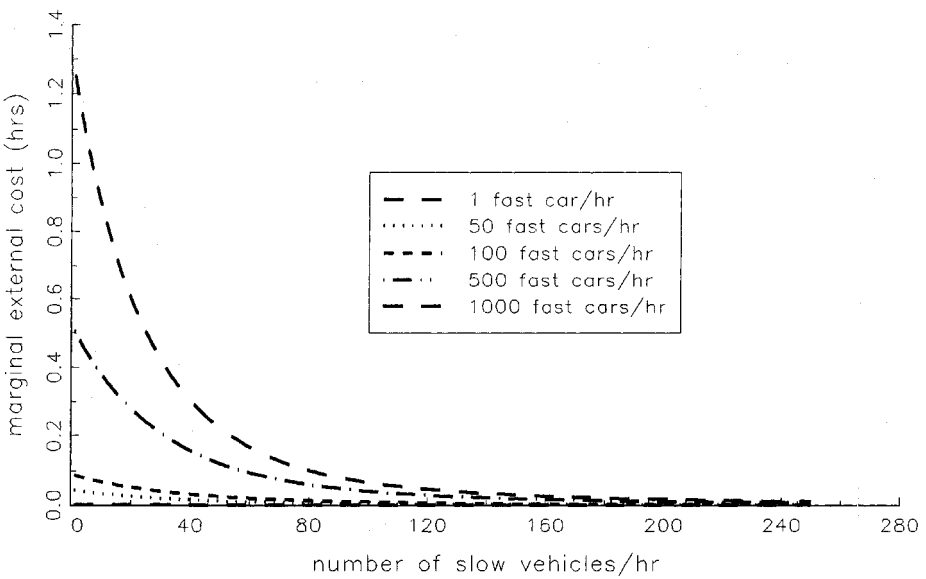
[8] Note that the travel time of slow drivers is assumed to be independent of traffic circumstances.

The curves drawn in Figure 4 are completely different. They show that the optimal toll for slow vehicles are a *decreasing* function of their demand for trips. The reason for this phenomenon is the concavity of the curves shown in Figure 2. An additional slow driver increases the travel time of fast vehicles especially when the number of slow vehicles is still small. When their number is large, travel time of the fast vehicles is already close to its maximum, and cannot increase further.

The optimal toll for the first slow vehicles that enter the road is highly dependent on the volume of the demand for trips by fast vehicles. If this is

Figure 4

The relationship between the number of slow vehicles and their marginal external cost for various numbers of fast vehicles.



very small (1 driver per hour) the optimal toll is negligible (the curve is indistinguishable from the horizontal axis). However, when demand is moderately large (50 cars per hour or more) optimal tolls for slow vehicles can be much larger than those for fast vehicles. Figure 4 shows that it can be equal to the monetary value of 1 hour if demand by fast vehicles equals 1000 cars per hour, and it can become much higher (more than 10 hours) for higher levels of demand by fast vehicles.

Discussion

We have just seen that the optimal toll for slow vehicles is a decreasing function of the volume of their demand for trips (at least for the parameter

values that have been used in constructing the figures). The toll approaches zero even for moderate values of demand by slow vehicles (say 200 cars per hour).

In the Pigou–Knight analysis of traffic congestion it is conventional to regard the curve that gives the price for the trip (including a possible toll) as similar to the supply curve in the analysis of market equilibrium. The analogous picture under the present circumstances is shown in Figure 5. The figure shows a linear inverse demand curve that crosses the vertical axis at 2, implying that the maximum value attached to a trip on the road segment is the (monetary) equivalent of two hours travel time. The maximum number of trips per unit of time demanded by slow vehicles is 250 if travel costs would be equal to zero. The figure also shows the travel time on the road segment. Without policy measures, the market equilibrium would be at the point where the demand curve crosses the horizontal line corresponding to the travel time.

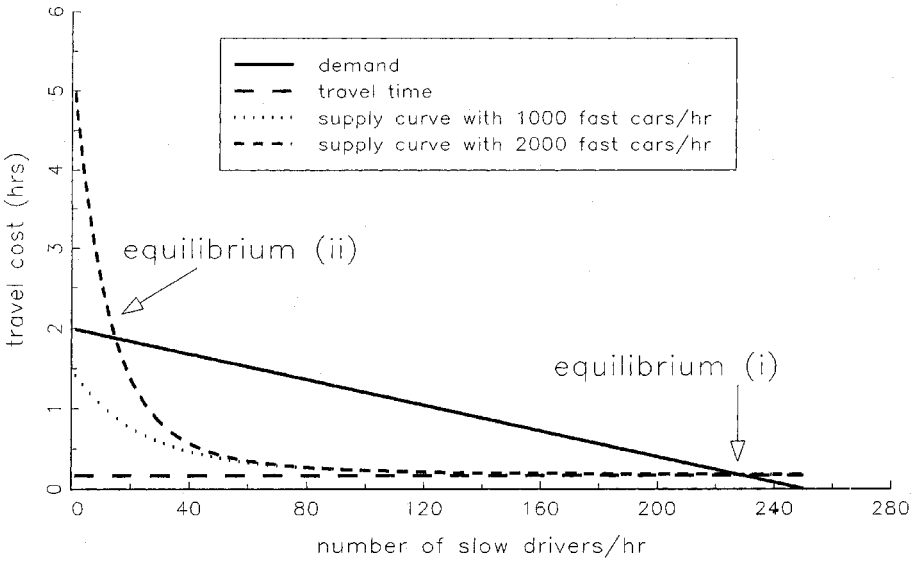
Finally, the picture shows two "supply" curves that give social travel cost as a function of demand for trips by slow vehicles. The "supply" curve is simply the sum of the travel time of slow cars and the marginal external cost. It is assumed that demand for trips of fast vehicles is inelastic and the figure shows the supply curve for various levels of this fixed demand. In both cases shown the supply curve crosses the demand curve at a point where the toll is relatively low and demand of slow vehicles is close to what it would be in a situation without tolling. This situation, which is indicated in the figure as equilibrium (i), is the first candidate for an optimum. However, more such points can be distinguished.

If the demand for trips of fast vehicles is high, there may be a second point where the two curves cross. At this point the optimal toll is so high that demand for trips by slow vehicles is reduced to a very small value. This situation is indicated in the figure as equilibrium (ii). However, equilibrium (ii) can never be an optimum since the supply curve crosses the demand curve from above. Situation (ii) therefore refers to a minimum.[9]

[9] To see this, consider the effects of a small change in the number of slow vehicles from situation (ii) on the social surplus. If the number of slow vehicles increases, their part of the social surplus, $\int g_2(m_2)dm_2 - v_2 w_2 \mu_2$, certainly increases by $g_2(\mu_2) - w_2$, the difference between the value of the trip of the marginal slow driver and her travel time. On the other hand, the social surplus of the fast vehicles decreases and this decrease is equal to the marginal external cost of slow vehicles. This marginal external cost is equal to z_2 (see (8)) and equals the difference between the value of the supply curve and the travel time. Both effects therefore cancel out in situation (ii) where supply and demand cross. However, if the number of slow vehicles increases, the positive effect on the surplus of slow vehicles becomes larger than the negative effect on the surplus of the fast vehicles and total social surplus increases. Similar reasoning shows that a decrease in the number of slow vehicles from situation (ii) onwards will also increase total social surplus.

Figure 5

"Market equilibria" for slow drivers at various levels of inelastic demands of fast drivers



The third possibility we have to consider is one in which the number of slow vehicles is equal to 0, situation (iii). In this situation the marginal external cost of slow vehicles is infinitely high (see (11)), so increasing the number of slow vehicles will at first always decrease the social surplus as long as $g_2(0)$ is finite (as we assume). Hence, situation (iii) is a local corner maximum. The only way to show whether it is a global maximum as well is by comparison with situation (i). If we denote the supply curve as $s(\mu_2)$ and the number of slow vehicles in situation (i) as μ^*, then we can write the difference in social surplus between situations (iii) and (i) as:

$$SS(iii) - SS(i) = \int_0^{\mu^*} [g_2(\mu_2) - s(\mu_2)]\, d\mu_2. \tag{12}$$

The situation depicted is not the only possible one. In the case where demand is linear, different situations occur if the demand curve becomes steeper. If the demand curve crosses the horizontal axis at low numbers of slow drivers, the difference between equilibrium (i) and the situation without tolling becomes more substantial. Moreover, equilibria (i) and (ii) become closer to each other. However, it must be noted that, given the characteristics of the "congestion technology" discussed above, such a situation can only occur if demand for trips by slow vehicles is small even at travel time zero. If the demand curve is very steep, no equilibrium with

tolling may exist, because the demand curve is everywhere below the "supply" curve. In this case a policy maker is only able to choose between the user equilibrium and situation (iii).[10]

If the demand curve is non-linear, there may of course be more points at which supply and demand curves cross and more candidates for a global maximum. However, the supply curve is always very steep at extremely low levels of demand of slow vehicles and almost horizontal for even moderate levels of that demand. This suggests that in many cases a situation such as (i) will correspond with a maximal social surplus. In such a situation the optimum is very close to a fourth possibly optimal situation, namely one in which no policy measures are taken. Even small costs associated with tolling will make this situation (iv) of *laissez faire* the real optimum.

In conclusion, under the circumstances studied here, tolling does not seem to be a very effective policy if one wants to maximise social surplus. If situation (i) is the optimum, even small costs associated with tolling make it preferable to do nothing. If situation (iii) is the optimum, a complete prohibition on slow vehicles from entering the road is the optimal instrument.

A numerical example

Figure 5 was drawn on the assumption of a completely inelastic demand for trips by fast vehicles, and we maintained that assumption throughout the discussion of the previous subsection. If demand is elastic, the supply curve for the slow vehicles can only be drawn on the basis of some assumption on the number of fast vehicles. However, except for very low levels of demand for trips by slow vehicles, these supply curves can hardly be distinguished from each other and we do not expect much harm from this tacit assumption. It may, nevertheless, be useful to provide a numerical illustration for the more general case in which demand for trips by fast vehicles is also elastic. We will not derive supply curves, but only compare the four situations listed above.

We consider again the same road segment but now we suppose that the inverse demand curves of both types of vehicles are linear. For simplicity, the value of time is taken to be equal to 1 for both types of vehicles. The intercept of the inverse demand curve of the fast and slow vehicles is assumed to be equal to two hours. At the minimum travel time of l/s_1 demand for trips by fast vehicles is taken to be equal to 1,000, demand for

[10] Between the cases of two equilibria with tolling and no equilibrium there is the special case in which the two tolling equilibria coincide.

trips by slow vehicles equals 500 if travel time is equal to l/s_2. The numerical results are summarised in Table 1.

Table 1
Four Situations

	Situation	Number of Fast Vehicles Per Hour μ_1	Number of Slow Vehicles Per Hour μ_2	Travel Time of Fast Vehicles w_1	Social Surplus SS	Toll on Fast Vehicles (hrs) z_1	Toll on Slow Vehicles (hrs) z_2
i	Optimal tolling	977.97	499.27	0.1657	1355.6229	0.0006	0.0026
ii	Suboptimal tolling	968.55	15.08	0.1396	919.4798	0.0132	1.7782
iii	No slow vehicles	1000.00	0.00	0.1250	937.5000	-	-
iv	*Laissez faire*	978.32	500.00	0.1657	1355.6218	-	-

Legend. "Optimal tolling 1" refers to optimal tolling at a high level of demand for trips by slow vehicles, "Optimal tolling 2" to a low level .

Optimal tolling at a high level of demand for trips (i) by slow vehicles implies hardly any difference with the situation of *laissez faire* (iv) in which there are no measures for traffic control. In situation (i) the tolls are negligible for both types of vehicles and they have hardly any influence on demand or social surplus.

In the second equilibrium with tolling (ii) the number of slow vehicles is much smaller. Travel time for the fast vehicles decreases as a consequence, but consumer surplus also falls considerably when compared to (i) and (iv). The value of the toll for slow vehicles is high: the monetary equivalent of 1.78 hours of time. The demand for trips by these vehicles falls to a level as low as 15 per hour. Nevertheless, the influence of these 15 vehicles on the travel time of fast vehicles is substantial, as a comparison with the situation without slow vehicles on the road makes clear.

The table shows that in the present situation complete abandoning of slow vehicles (iii), which implies that toll revenues are zero, leads to a higher social surplus than (ii) even though demand of fast vehicles is now elastic.[11] Situations (ii) and (iii) have a substantially lower social surplus than (i) and (iv). Clearly, in the circumstances considered here, even the slightest costs associated with tolling make *laissez faire* the preferred alternative.

For other combinations of parameters the relative attractiveness of the alternatives may be rather different. For instance, when the demand for

[11] If demand becomes more elastic examples can be constructed in which (ii) has a slightly higher social surplus than (iii).

slow traffic is much lower than that for fast traffic, the "no slow vehicles" alternative (iii) may easily perform better than the *laissez faire* alternative. The reason is that the loss of consumer surplus by the slow group—when it is excluded from the road—is much smaller than in the case of Table 1. Note also that the marginal external costs of slow vehicles are very high when there are only a few of them.

The discussion above proceeded on the assumption that vehicle types can be easily identified for the purpose of tolling. This is obviously the case when special kinds of vehicle (agricultural vehicles, trucks) are the slow ones, while private cars are the other. The situation becomes more difficult if speed differences stem from characteristics of the vehicles or the drivers that are not (or not easily) observable. For instance, if only heavily loaded trucks have a low speed, whereas the others have a high speed, it may for all practical purposes be impossible to select the group to be tolled. In such a situation the effectiveness of tolling is less than under the first best conditions assumed above.

Relevance

In reality we see in the Netherlands that in the countryside there is a dense road network connecting all houses and farms where all kinds of traffic are allowed. There is also a coarser network of better roads, connecting the villages, where slow traffic (tractors, bicycles, and so on) is sometimes not admitted.

The main motivation to separate slow and fast traffic in this situation relates to safety (SWOV, 1990): speed differences are known to be an important determinant of accidents. Our analysis reveals that this policy is also supported by considerations of time losses and social welfare. As the analysis given above indicates, routes with a very high demand of fast traffic can probably best be used exclusively for fast traffic, while the resulting welfare losses to slow traffic are moderate if there is a sufficiently dense network of other roads on which it is allowed.

Comparison with earlier work

The version of the model discussed in the present section—with a single lane, without overtaking—is among the scenarios considered in this paper closest to the model we presented in Verhoef *et al.* (1999). The main qualitative difference involves the consideration of stochastic arrival rates here, as opposed to deterministic arrival rates in Verhoef *et al.* (1999). This means that the model discussed in the present section allows us to assess the impacts of the introduction of stochasticity in arrival rates alone upon the properties of the model.

In a qualitative sense, the main results of the present model are consistent with those obtained with deterministic arrival rates. Also for that model, the marginal external costs of slow vehicles are decreasing in their equilibrium use level, and the marginal external costs of fast vehicles are typically a small fraction of those caused by slow vehicles. Also in that model fast vehicles' equilibrium speeds are bounded between their own preferred speed and that of the slow vehicles.

The main qualitative difference between the two formulations stems from the fact that with deterministic arrival rates, a distinction can be made between situations where fast vehicles will experience congestion (that is, have to slow down) with a probability (p_c) smaller than unity, and cases where this will occur with certainty ($p_c = 1$). The resulting 'regime shifts' that may occur when varying equilibrium numbers of users for either groups, as in Figures 1–5 just discussed (that is, when p_c changes between unity and below-unity), will cause the relevant curves to show sharp kinks. Figure 1 in Verhoef *et al.* (1999) gives a clear example. The most important consequence of this is that in a formulation with deterministic arrival rates, marginal external costs for slow vehicles exhibit a relatively flat segment for $p_c < 1$, and start falling sharply—as in Figure 4 in the present paper—only from $p_c = 1$ onwards. For fast cars, the pattern is different: there, the marginal external costs rise—as in Figure 3 in the present paper—only for $p_c < 1$, and fall from $p_c = 1$ onwards, as the equilibrium speed for fast cars approaches its minimum (the slow vehicles' speed), and extra fast cars no longer affect their own speed. Apparently, in the present model, this latter effect does not yet occur for flow levels (of fast cars) even near the maximum of 3,000 vehicles per hour (see Figure 3).

The formulation in the present paper has the practical advantage that such regime shifts and kinks are avoided, and it probably better reflects congestion in reality, for which of course it is typically not possible to predict with certainty whether or not a slow vehicle will be encountered during a trip.

Finally, as far as quantitative aspects are concerned, a quick check revealed that the curves depicted in Figures 2 and 4 are very similar to those that would be found with an equivalent model based on deterministic arrivals (apart from the kinks just discussed). These are obviously the most significant curves from a policy perspective. For travel times and marginal external costs as a function of the use level of fast cars, as in Figures 1 and 3, the differences are relatively larger. Given the small size of the externality involved, we will not investigate these differences further here, but instead move on to consider an extension to the model in which a second lane and overtaking possibilities are introduced.

Two Lane Road, Overtaking Possible

In this section we analyse the changes that occur in the model of the preceding section when we introduce the possibility of overtaking. We consider a road with two lanes and traffic moving in opposite directions on both lanes. Overtaking is possible if a sufficiently large gap occurs in the traffic moving in the opposite direction.

We assume that similar stochastic processes describe traffic on both lanes, although some of the parameters may have different values.

The minimum required length of a gap

A fast driver on the right lane who overtakes a slow driver has to be able to use the left lane for the time needed to complete the overtaking. Overtaking starts when the nose of the fast car is at the minimum distance d^* behind the car in front. The fast driver switches to the left lane and overtakes the slow car at the high speed s_1.[12,13] The fast driver moves back to the right lane as soon as the nose of his car is at the minimum distance to the nose of the slow car that is overtaken. We remove any problems caused by the possibility that two slow vehicles are driving behind each other so closely as to make it impossible to overtake them one by one.[14] The time needed for overtaking is therefore equal to that of proceeding $2d^*$ metres at a speed $s_1 - s_2$: $2d^*/(s_1 - s_2)$. During this time the fast driver travels $s_1 2d^*/(s_1 - s_2)$ metres. If the first oncoming car on the left lane has speed s_3 it travels $s_3 2d^*/(s_1 - s_2)$ metres during the same time. The required length of the gap in traffic on the left lane is therefore equal to:

$$g = 2\,\frac{s_3 + s_1}{s_1 - s_2}\,d^*. \tag{13}$$

[12] If the fast car was forced to slow down before overtaking started, speed switches to the high value at the start of overtaking.

[13] In reality overtaking often takes place at a higher speed in order to limit the duration of the overtaking. In order to take this into account, we should introduce a third speed, the one preferred during overtaking, into our model. However, in reality it seems also the case that drivers do not switch back to the right lane as soon as possible as is assumed in our model, but remain on the left lane somewhat longer. The two simplifying assumptions used here may therefore be expected to compensate for each other to some extent. Relaxing the assumption that drivers use the preferred speed during overtaking would complicate the model and require us to deal with the extraordinary possibility that travel time of fast drivers decreases as a result of the presence of some slow vehicles because of the higher speed they use during overtaking, which is probably of no practical interest.

[14] This can be motivated by assuming that the last of these two slow vehicles will throttle back in order to enable the overtaking car to move to the right, if this is necessary because of oncoming traffic on the other side of the road.

The speed s_3 may be equal to s_1 or s_2, but may also be different.

Some assumptions

In order to keep the model tractable, we introduce a number of assumptions that differ from those made in the previous section.

- We assume that if the relationship between demand and arrival rates is taken into account, as was done in equations (2) and (3), there is no need to deal explicitly with interactions between fast vehicles that are the result from platooning and so on. Previously this was a result established in the Appendix. Here it is an assumption. It may be possible to prove its validity also in the present situation, but we have not done so.
- All cars on the left lane have the same speed s_3. This allows us to derive expressions for the probability that the length of a gap exceeds a certain value, for the expected value of the time one has to wait until a sufficiently large gap occurs, and so on. Demand for trips on the left lane is denoted as μ_3 and it is assumed that the same minimum distance d^* has to be maintained between cars on that side of the road. The arrival rate λ_3 is therefore equal to $\mu_3/(1-\mu_3/c)$.[15]
- We consider a steady state of traffic moving on a long road in which a fast driver alternately drives at a high or a low speed. We therefore adopt a special feature of the model of the previous section: all fast cars that entered the road could proceed at the high speed when entering the road.[16]

Average speed and average travel time

We now consider the average speed and travel time of a fast driver on the right hand side of the road under the circumstances described above. His experience is a sequence of events consisting of two parts:

- a period during which the high speed can be maintained that starts when a slow car begins to be overtaken and ends when the distance to another slow car is minimal, so that he has to slow down to s_2;

[15] Although it would have been more satisfactory to treat traffic in both directions symmetrically, this would also complicate our derivations considerably. The assumption of homogeneous traffic is easier to use, and probably does not change the qualitative results. The even simpler assumption of deterministic arrival rates on the left lane would imply that it is either always possible or always impossible to overtake between two subsequent cars, which is clearly not realistic.

[16] Note that the steady state in the model of the previous section is a situation in which all fast cars are forced to slow down.

- a period of driving at the low speed, until a sufficiently large gap occurs in traffic moving in the opposite direction.

The average speed $A(s)$ of the fast driver is therefore equal to a weighted average of the high and the low speed, with the weights equal to the expected length of the periods during which both speeds are used. Denoting these periods as φ_1 and φ_2, we can write:

$$A(s) = \frac{\varphi_1 s_1 + \varphi_2 s_2}{\varphi_1 + \varphi_2}. \tag{14}$$

For the average travel time per unit of distance, $A(w)$ we find:

$$A(w) = \left(\frac{1}{A(s)} =\right) \frac{\varphi_1 + \varphi_2}{\varphi_1 s_1 + \varphi_2 s_2} \tag{15}$$

In the Appendix the following expressions for φ_1 and φ_2 are derived:

$$\varphi_1 = \tau_1/(1-\pi_1), \text{ with } \tau_1 = \frac{d^* + s_2/\lambda_2}{s_1 - s_2} \text{ and } \pi_1 = e^{-\lambda_3 g/s_3}/(d^*\lambda_3/s_3 + 1).$$

$$\varphi_2 = \tau_2/\pi_2 \text{ with } \tau_2 = \frac{1}{s_2 + s_3}\left(d^* + \frac{s_3}{\lambda_3} - g\frac{e^{-\lambda_3 g/s_3}}{1 - e^{-\lambda_3 g/s_3}}\right) \text{ and } \pi_2 = e^{-\lambda_3 g/s_3}.$$

Optimal tolls

For the marginal external costs the following equations can be derived easily from the Lagrangean given in the appendix:

$$\begin{aligned} z_1 &= \mu_1 v_1 \frac{l\,\partial A(w)}{\partial \mu_1}, \\ z_2 &= \mu_1 v_1 \frac{l\,\partial A(w)}{\partial \mu_2}, \\ z_3 &= \mu_1 v_1 \frac{l\,\partial A(w)}{\partial \mu_3}. \end{aligned} \tag{16}$$

We focus attention on the first and second group of vehicles. Observe that $A(w)$ depends on μ_1 and μ_2 only through λ_2, while λ_2 appears only in φ_1.

After some elaboration we find:

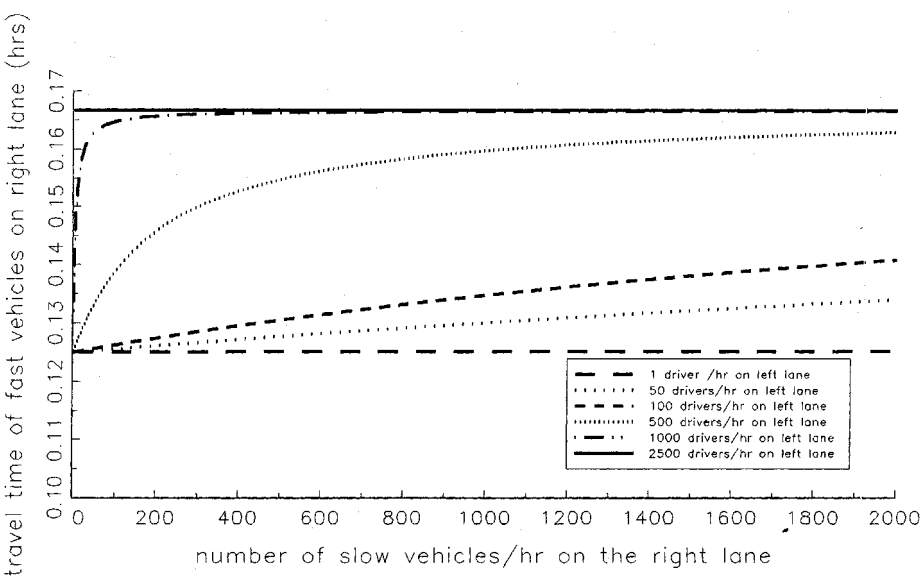
$$\frac{\partial A(w)}{\partial \mu_1} = \frac{\varphi_2 s_2}{(\varphi_1 s_1 + \varphi_2 s_2)^2} \frac{1}{(1-\pi_1)} \frac{1}{c} \frac{1}{\mu_2},$$
$$\frac{\partial A(w)}{\partial \mu_2} = \frac{\varphi_2 s_2}{(\varphi_1 s_1 + \varphi_2 s_2)^2} \frac{1}{(1-\pi_1)} \left(\frac{c - (\mu_1 + \mu_2)}{\mu_2} + 1 \right) \frac{1}{c} \frac{1}{\mu_2}. \tag{17}$$

An example

As an example we take the same road segment as in the previous section and apply our new assumptions. An alternative interpretation, that is unrealistic but may nevertheless be helpful in interpreting the results, is to imagine a circular road with a circumference of 10 kilometres where traffic moves continuously. It is assumed in all figures that traffic on the left lane has a speed of 60 km/hr. Figure 6 shows the travel time of fast vehicles as a function of the demand of slow vehicles for various values of the demand for trips on the left lane. If that demand is very low, there is no congestion and travel time equals $1/s_1$. If the number of cars on the left lane increases, the travel time function takes on the concave shape discussed in the previous section. For very high traffic intensities on the left lane the travel time curve becomes flat again, but now at the maximum value $1/s_2$. In

Figure 6

The relationship between travel time of fast vehicles and the number of slow vehicles on the right lane



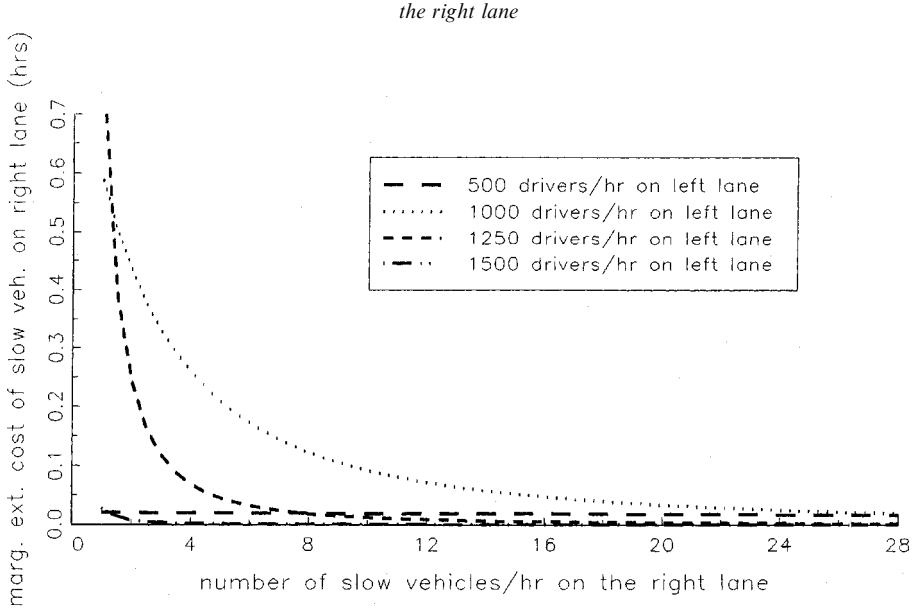
contrast with the model of the previous section, travel time of fast vehicles will ultimately be equal to $1/s_2$ for every positive value of the demand for trips by slow vehicles. This is a consequence of our steady-state assumption: if overtaking is impossible and there are slow vehicles on the road, every fast driver will ultimately be forced to slow down.

Although the assumptions that underlie both figures are somewhat different, we can get some idea of the effect of introducing the possibility of overtaking when we compare Figures 6 and 2. Figure 2 showed that the presence of even a small number of slow vehicles increased travel time of the fast vehicles considerably. In Figure 6 we find that travel time of fast vehicles remains close to its minimum value, even in the presence of a large number of slow vehicles, as long as traffic intensity on the left lane is low. If there is much traffic on the left lane the possibilities for overtaking disappear and the resulting situation is similar to that studied in the previous section.

Note that most of the lines drawn in Figure 6 are strictly concave and the others are straight lines. This suggests of course that optimal tolls will in the present circumstances also be decreasing for small volumes of demand for trips by slow vehicles and almost equal to zero for higher levels of demand. Figure 7 confirms this conjecture (note that the scale of the horizontal axis differs from that of Figure 6).

Figure 7

The relationship between the marginal external cost of slow vehicles and the number of slow vehicles on the right lane



As noted above, a special feature of the model of this section is that the travel time function becomes flat if there are many cars on the left lane. This makes one expect that optimal tolls become flat for high traffic intensities on the left lane. Figure 7 confirms this: if there are 1,500 cars per hour on the left lane, the marginal external cost is (much) lower than if there are 1,250 cars per hour. If the number of cars on the left lane increases to still higher values, optimal tolls are negligible even for extremely low numbers of slow vehicles.

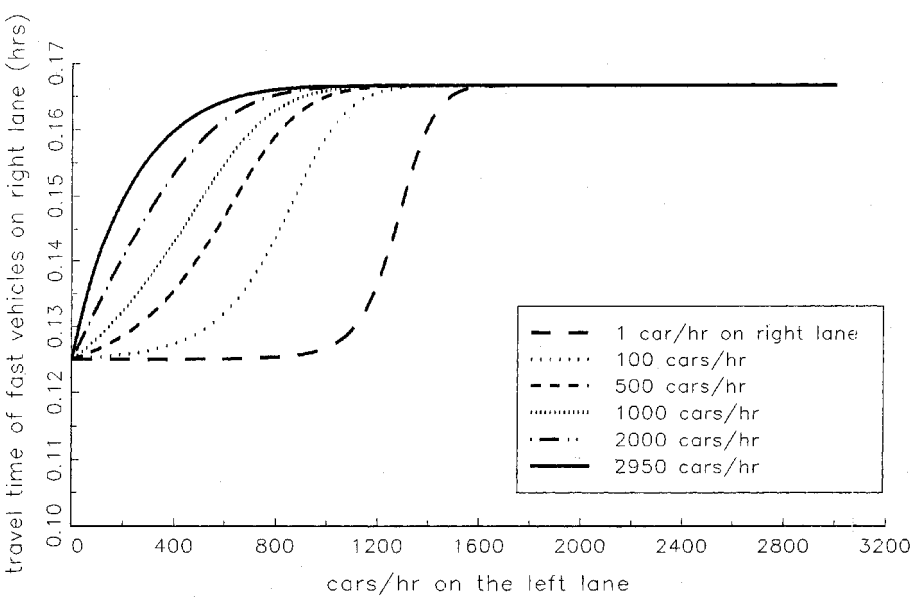
External effects of traffic moving in the opposite direction

In Figure 8 we look at the relation between the number of vehicles per hour on the left lane and the travel time of fast vehicles on the right lane. The curves drawn refer to various numbers of vehicles per hour on the right line. In all cases the number of slow vehicles is one-third of this total, the others are fast vehicles. As before, we assume that all vehicles on the left lane drive at 60 km/hr.

For very small traffic intensities on the right lane the travel time of fast vehicles on the right lane is almost equal to its minimum until traffic on the

Figure 8

The relationship between travel time of fast vehicles on the right lane and the number of vehicles on the left lane



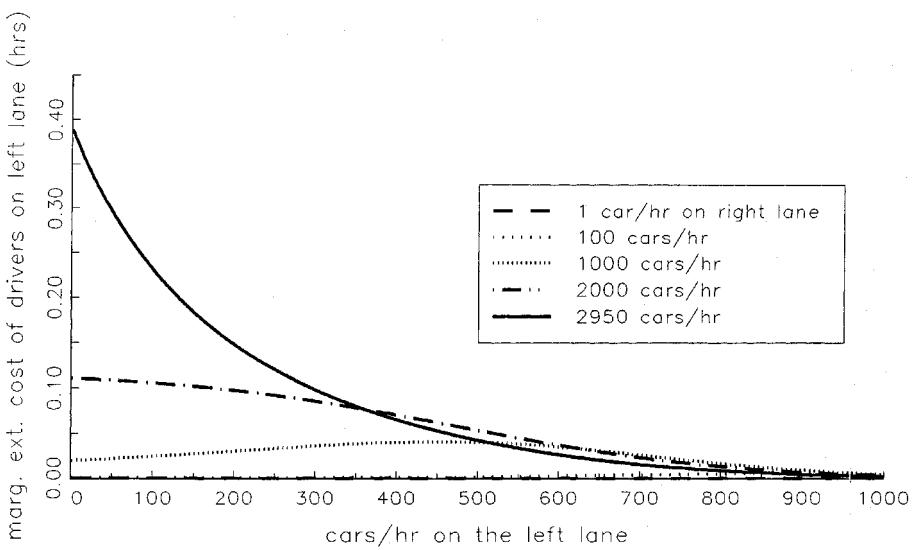
left lane becomes so intense that overtaking becomes impossible. Then it increases rapidly to the maximum travel time (remember the steady state context of the present model). For higher volumes of traffic on the right lane the curves start to increase earlier and approach the maximum value gradually.

The marginal external cost imposed by vehicles on the left lane on vehicles on the right lane is shown in Figure 9. This figure has been constructed on the basis of numerical differentiation. (Analytical methods can be used to calculate the third of equations (16) but imply tedious derivations.) Note that the curves shown in Figure 9 refer to the slope of the corresponding line in Figure 8, multiplied by μ_1, the number of fast vehicles on the right side of the road.[17] That is the reason why the marginal external costs are always virtually equal to zero when the number of fast vehicles on the right side equals 1, even though the travel time of this single vehicle increases rapidly when the number of vehicles on the left side is approximately 1,300, as shown by the corresponding line in Figure 8.

For larger numbers of vehicles on the right lane, marginal external costs of traffic on the left lane become larger, even though the cor-

Figure 9

The relationship between the marginal external cost of vehicles on the left lane and their number



[17] See the third of equations 16.

responding curves in Figure 8 are less steep. For 500 vehicles per hour on the right lane, the marginal external cost reaches a maximum when there are approximately 600 cars per hour on the left lane. Comparison with Figure 8 makes it clear that the travel time of fast drivers in the right lane increases rapidly at that level of demand on the left lane.

At higher levels of traffic demand on the right lane, external costs of cars on the left hand side are more substantial and the number of cars on the left lane for which marginal external costs are maximal shifts to the origin. Figure 9 shows that for 2,000 and 2,950 cars on the right lane the marginal external costs of vehicles on the left lane are a decreasing function of their number.

Relevance

We conclude that introducing the possibility of overtaking does not change the essence of the conclusions we drew in the previous section. Tolling of slow vehicles does not seem to be a promising policy since, depending on the number of slow vehicles, the alternatives *laissez faire* and "no slow vehicles" are probably almost as good in terms of total welfare and are much easier to implement.

However, there is an additional element in the model where overtaking is allowed: vehicles on the left lane generate external costs on the fast vehicles on the right lane because they hinder overtaking. Figure 9 shows that only when the right lane is very busy and the left lane is very quiet, are the marginal external costs of the left lane vehicles considerable. This suggests that only in this case does it make sense to impose a toll on users of the left lane. However, if the number of vehicles on the left lane is small, the policy of not allowing them to enter the lane would yield a welfare result that is not far from imposing an optimal toll and clearly better than the *laissez faire* alternative.

A real world example that looks like this situation is the morning peak with large traffic flows in one direction and small flows in the other. The instrument of switch lanes (used for traffic in one direction during the morning peak and for traffic in the other direction during the afternoon peak) comes close to this outcome of allocating road capacity exclusively for vehicles in one direction.

Another obvious conclusion that follows from the analysis is that imposition of a ban on overtaking leads to very small time losses when traffic is heavy in both directions, whereas the gains in terms of improved safety (not addressed in this paper) may be substantial.

Conclusions

In this paper we analysed congestion caused by speed differences in situations where two groups of vehicles with different preferred speeds use a single lane. This relatively simple situation allowed us to develop a structural model of this type of congestion. Our main result is that this model indicates that the relationship between marginal external costs and numbers of vehicles are very different from the increasing convex curves familiar from the Pigou–Knight analysis. In Section 2 we considered the case in which overtaking is impossible and found that marginal external costs of slow vehicles are a *decreasing* function of their number, while optimal tolls for this group are virtually zero, unless their number is small. Clearly tolling is not a very useful policy in this situation. We suggested that the actual policy, which prohibits slow traffic entering some primary roads, while allowing all kinds of traffic on a denser network of secondary roads, might be optimal.

In Section 3 we found that introduction of the possibility of overtaking changes some aspects of the model, but not the downward sloping relationship between the marginal external costs of slow vehicles and their number. Moreover, we found that the relationship between marginal external cost and the number of vehicles on the left lane is only monotonous when the number of vehicles on the right lane is close to the capacity of that lane. Then we find again that these costs are a decreasing function of the number of cars. For lower numbers of vehicles on the right side, we find that these marginal external costs are initially increasing.

Since our analysis makes use of numerical simulations, we have carried out sensitivity analyses by changing the values of key variables namely the desired speeds and the minimum required distance. In all cases the results were qualitatively the same as those reported in the text.[18]

A main conclusion that can be drawn from our analyses is that tolling slow vehicles tends to have rather small effects on total welfare. Depending on the number of slow vehicles, the alternatives *laissez faire* or "no slow vehicles" are almost as good in terms of total welfare and

[18] The original values of these parameters were chosen deliberately so as to represent a road that may be thought of as connecting two villages in a rural area, and the alternative values of the parameters were chosen so as to maintain the possibility of this interpretation. The differences between the alternative values and the ones used originally was therefore limited (for speeds the maximum difference was 20 km/hr, for the distance between cars 10 metres).

probably better if the costs of implementation are also taken into account. It may also be noted that we have tacitly assumed that it would be possible to toll the two groups of vehicles differently, which requires that a policy maker should be able to identify these groups (for instance, on the basis of observable characteristics). If this assumption is false, it will be even more difficult to improve welfare by means of tolling.

Although our results refer to a special and simplified situation, they are nevertheless interesting because they suggest that it may be quite worthwhile to develop a more structural analysis of congestion and the appropriate policy instruments to deal with it. Especially in situations where congestion is not related to the existence of a bottleneck (for this situation a well-developed structural model exists), this may result in new insights. The traditional Pigou–Knight analysis which is commonly used as a kind of benchmark model for thinking about congestion in such situations suggests a relationship that turns out to be quite different from the one studied in this paper.

Finally, we would like to stress that our analysis does not refer to a theoretical curiosity, but is likely to have policy relevance. Our examples were constructed so as to resemble the traffic situation on roads (not highways) outside the main cities in the Netherlands, which is probably close to that in comparable areas in most other advanced economies. Congestion caused by speed differences may also occur because of interaction between various traffic modes. A typically Dutch example would be the interaction between cyclists and motor vehicles. In the Netherlands the (parts of) the infrastructure used by cars and cyclists are often separated, but in other countries interaction between cars and slower traffic modes seems to be an important source of congestion, especially in large cities.

The analysis of speed differences may also be relevant in situations that are different from the one discussed in the present paper. For instance, Newell's (1998) analysis of a moving bottleneck suggests a study of the welfare economic effects of congestion caused by speed differences on a highway where two (or more) types of vehicles use two lanes that are exclusively used for traffic in one direction. If the capacity of one lane is insufficient for allowing the fast vehicles to overtake the slow ones, there will be queueing behind the slow vehicles. If there are three types of vehicle with different preferred speeds, congestion occurs also when a vehicle with medium speed overtakes one with a slow speed and thereby forces a vehicle with a high speed to slow down.

Appendix

Expected travel time for fast vehicles when overtaking is impossible

We start the analysis by introducing an additional simplifying assumption, which will be dropped later on. The assumption states that all cars can be considered as points (that is, do not occupy road space) and can approach each other indefinitely close.

Consider a fast driver who enters the road and let τ be the time that elapsed since the last slow driver before him entered the road segment. This slow car has therefore travelled $s_2\tau$ kilometres when the fast driver enters. The distance between the fast and the slow driver will start to decrease immediately and after $s_2\tau/(s_1 - s_2)$ time units it is equal to zero. The fast car will then have travelled $s_1 s_2\tau/(s_1 - s_2)$ kilometres. If this number exceeds the length of the road l the fast car will not experience congestion. Total travel time w is then equal to l/s_1. The critical value for τ can be derived as $l(s_1 - s_2)/s_1 s_2 = l/s_2 - l/s_1$.

If τ is smaller than this critical value, the fast car will be forced to move at a low speed for the last $l - \tau s_1 s_2/(s_1 - s_2)$ kilometres of the road. Total travel time can then be computed as:

$$w = \frac{1}{s_1}\frac{s_1 s_2}{s_1 - s_2}\tau + \frac{1}{s_2}\left(l - \frac{s_1 s_2}{s_1 - s_2}\tau\right),$$

$$= \frac{l}{s_2} - \tau.$$

A standard result from the theory of Poisson processes says that τ is distributed exponentially with parameter λ_2. It should be recalled that this refers only to the case with τ below the critical value.

If X is a random variable with an exponential distribution (with parameter λ) that is truncated at k, its expected value can be derived as follows:

$$E(X) = \int_0^k \lambda x e^{-\lambda x}\, dx + k \text{ Prob } (X > k),$$

$$= \frac{e^{-\lambda k}\lambda k + e^{-\lambda k} - 1}{\lambda} + k e^{-\lambda k},$$

$$= \frac{1 - e^{-\lambda k}}{\lambda}.$$

The expected value of travel time for a fast driver is:

$$E(w) = \frac{l}{s_2} - E(\tau),$$

$$= \frac{l}{s_2} - \frac{1}{\lambda_2}\left(1 - e^{-\lambda_2(1/s_2 - 1/s_1)}\right)$$

The second line results from application of the formula derived above to the case with $X = \tau$, $\lambda = \lambda_2$ and $k = l/s_2 - l/s_1$.

Now we drop the assumption that cars do not occupy any space and assume that a distance d^* between (the noses of) subsequent cars is always maintained. This results in two complications:

- a Poisson process is no longer suitable as a model for arrivals since it implies that two cars can enter at (almost) the same time;
- the number of fast cars between an arbitrary fast car and the first slow car in front of it becomes of importance, since it determines the location where the fast car has to switch to the speed.

Arrival of cars is now modelled as a source that can be on and off. If the source is on, a car arrives after a time that is distributed exponentially with parameter $\lambda_1 + \lambda_2$. After a car arrives, the source will be off for d^*/s_2 times units. This guarantees that the first d^* metres of the road are always free if the source is on. (It is possible to change this second characteristic in such a way that the source will be off for a smaller length of time until a fast car arrives, but this complicates the derivations and adds nothing significant to the analysis.)

Now consider a fast car that is entering the road. In order to determine the location where speed has to switch to the lower value, notice the following facts:

- the number of fast cars that entered the road after the last slow car that entered plus that slow car itself is distributed geometrically with parameter $\lambda_2/(\lambda_1 + \lambda_2)$: call this number K;
- if $K = k$, then the distance between the fast car and the slow car at the moment the former has to switch to the low speed is kd^*;
- The random variable $\Sigma_{i=1}^{K} X_i$, with the Xi's mutually independent exponentially distributed with parameter λ, and K geometrically distributed with parameter p, is exponentially distributed with parameter λp.

We now determine the location of the slow car, conditional upon $K = k$. This location is $s_2 T$ with T the elapsed time since the slow car

entered the road. Conditional upon $K = k$ we have:

$$T = X_1 + d^*/s_2 + X_2 + d^*/s_2 \cdots + X_k + d^*/s_2,$$

$$= kd^*/s_2 + \sum_{i=1}^{k} X_i.$$

This equation expresses the fact that the time between subsequent starts equals $d^*/s_2 + X_i$, where X_i is distributed exponentially with parameter $(\lambda_1 + \lambda_2)$. The location L of the slow car is therefore:

$$L = s_2 T = kd^* + s_2 \sum_{i=1}^{k} X_i.$$

Making use of the second fact, we know that the fast car has to reduce speed when the distance to the preceding slow car equals kd^*. The time during which the fast speed can be maintained is therefore equal to the time it takes to reduce the distance to the slow car by $s_2 \Sigma X_i = T_s$ metres. This time is equal to:

$$T_s = \frac{s_2}{s_1 - s_2} \sum_{i=1}^{k} X_i.$$

Now use the third fact, with $\lambda = \lambda_1 + \lambda_2$ and $p = \lambda_2/(\lambda_1 + \lambda_2)$. This implies:

$$T_s = \frac{s_2}{s_1 - s_2} \sum_{i=1}^{k} T_s^*,$$

with T_s^* distributed exponentially with parameter λ_2. Notice that this derivation is still conditional upon k.

In order to complete the derivation of the distribution of the travel time of the fast car, one additional assumption has to be made. If the slow car in front of the fast cars leaves the road, which speed will the fast cars immediately behind it use until they leave the road? We assume that they maintain the low speed s_2. We now complete the derivation by observing that during the first T_s time units the car has speed s_1, while during the remaining time units speed will be s_2. The distance travelled after T_s time units is $((s_1 s_2)/(s_1 - s_2))T_s^*$ kilometres. If this is larger than l, travel time is l/s_1, otherwise $l - ((s_1 s_2)/(s_1 - s_2))T_s^*$ kilometres have to be travelled with

speed s_2. Analogous to the derivation under the simplified assumptions, the following conclusions may be drawn:

(a)
$$W = \frac{l}{s_1} \qquad \text{if } \frac{s_1 s_2}{s_1 - s_2} T_s^* > l,$$
$$= \frac{l}{s_2} - T_s^* \qquad \text{otherwise}$$

(b) W has the same distribution function as $l/s_2 - Z$, with Z an exponentially distributed distribution which is truncated at $(s_1 - s_2)l/(s_1 s_2)$.

In other words the distribution of W is equal to that derived under the earlier simplifying assumption. This could have been conjectured already because of the cancellation of the two complicating factors in the formula for T_s^*: the random variable K does not appear in it. The expected value for W may therefore be written as:

$$w_1 = (E(W) =) \frac{l}{s_2} - \frac{1}{\lambda_2}\left(1 - \exp\left(-\lambda_2 \left(\frac{l}{s_2} - \frac{l}{s_1}\right)\right)\right)$$

and this is equation (6) of the main text.

Before concluding this part of the appendix we observe some less obvious properties of this function.

(c)
$$\lim_{\lambda_2 \to 0} w_1 = \frac{l}{s_2} - \lim_{\lambda_2 \to 0} \frac{1 - \exp\left(-\lambda_2\left(\frac{l}{s_2} - \frac{l}{s_1}\right)\right)}{\lambda_2}$$
$$= \frac{l}{s_2} - \lim_{\lambda_2 \to 0} -\exp\left(-\lambda_2\left(\frac{l}{s_2} - \frac{l}{s_1}\right)\right)\left(\frac{l}{s_2} - \frac{l}{s_1}\right)$$
$$= \frac{l}{s_1}$$

(the second line makes use of l'Hopital's rule).

(d) w_1 is increasing in λ_2. This should of course be expected, but it is not completely obvious from the equation. The first-order derivative is:

$$\frac{\partial E_1(w)}{\partial \lambda_2} = \frac{1}{\lambda_2^2}\left(1 - [1 + \lambda_2 A]e^{-\lambda_2 A}\right)$$

with $A = l/s_2 - l/s_1$. We have to show that this is positive. Observe that for $\lambda_2 = 0$, the expression in brackets is positive. Computation of the relevant partial derivative of the expression in brackets shows that it is increasing in λ_2. Thus we conclude that $\partial E_1(w)/\partial \lambda_2$ is positive for positive values of λ_2. It follows that $\partial E_1(w)/\partial \lambda_2$ is always positive for

$\lambda_2$non-negative. Moreover, it can be shown that:

$$\lim_{\lambda_2 \to 0} \frac{\partial E(w)}{\partial \lambda_2} = \frac{1}{2}\left(\frac{l}{s_1} - \frac{l}{s_2}\right)^2,$$

which is also positive.

(e) w_1 is a concave function of λ_2. The second order derivative of w_1 is:

$$\frac{\partial^2 E(w_1)}{\partial \lambda_2^2} = -\frac{1}{\lambda_2^3}\{1 - [1 + \lambda_2 A]e^{-\lambda_2 A} + \lambda_2^2 A^2 e^{-\lambda_2 A}\}.$$

The expression between curled brackets is positive. The sum of the first two terms has been shown to be positive when determining the sign of $\partial E_1(w)/\partial \lambda_2$ and the third term is clearly also positive.

Optimal tolls

In order to derive the optimal tolls, we maximise the sum of the consumer surpluses of the two groups of vehicles, plus the toll revenues under the side conditions that the generalised travel costs are equal to the sum of time costs and toll at the optimum.

The Lagrangian is:

$$L = \int_0^{\mu_1} g(m_1)\, dm_1 - v_1 w_1 \mu_1 + \int_0^{\mu_2} g_2(m_2)\, dm_2 - v_2 w_2 \mu_2$$

$$+ \eta_1(g_1(\mu_1) - v_1 w_1 - z_1) + \eta_2(g_2(\mu_2) - v_2 w_2 - z_2) + \delta_1 \mu_1 + \delta_2 \mu_2,$$

where the last two terms refer to the constraints that both μ_1 and μ_2 should be non-negative. If demand of both fast and slow vehicles is positive, the two δ-s are zero, and the solution given in the text (10) can be obtained easily.

In order to analyse the possibility that one or both demands are zero at the optimum, observe:

- that the sum of the consumer surpluses will be zero when both demands are zero, which is never optimal as long as (at least) one of the two surpluses is positive with both tolls equal to zero;
- that it can never be optimal to set demand by fast vehicles equal to zero if $g_1(0) > l/s_2$, since the presence of fast vehicles do not cause congestion for slow vehicles.

This implies that in the situation in which $g_1(0) > l/s_2$ and $g_2(0) > l/s_2$ (that is, both surpluses are positive in a situation without tolling), the only

possible corner solution is the one in which demand by slow vehicles equals zero.

In such a situation there will be no congestion for the fast vehicles, which implies that their toll equals 0. The toll for slow vehicles is so high as to imply $w_2 + z_2 > g_2(0)$.

Since in practice the travel time for fast vehicles becomes close to w_2 if slow vehicles are allowed to drive on the road, what the foregoing essentially implies is that there is a choice between two equilibria:

(a) an equilibrium in which both fast and slow vehicles are permitted to enter the road, and travel times for both are practically equal to $w_2 = l/s_2$, while tolls are practically equal to zero,
(b) an equilibrium in which only fast vehicles are permitted to enter the road, and their travel time equals l/s_1, while slow vehicles do not enter the road because of a prohibitively high toll.

For the computation of the first partial derivative, $\partial E_1(w)/\partial\lambda_2$, we use equation (6). The result can be written as:

$$\frac{\partial E_1(w)}{\partial \lambda_2} = \frac{1}{\lambda_2^2}\left(1 - (1 + \lambda_2 A)e^{-\lambda_2 A}\right),$$

with $A = l/s_2 - l/s_1$. Note that the partial derivative $\partial E_1(w)/\partial\lambda_2$ depends on demand from both fast and slow vehicles through the arrival rate λ_2.

We use equation (4) in order to compute the other two derivatives that we need:

$$\frac{\partial \lambda_2}{\partial \mu_1} = \frac{\mu_2}{c}\frac{1}{\left(1 - \left(\frac{\mu_1 + \mu_2}{c}\right)\right)^2},$$

$$\frac{\partial \lambda_2}{\partial \mu_2} = \frac{\partial \lambda_2}{\partial \mu_1}\left(1 + \frac{c - (\mu_1 + \mu_2)}{\mu_2}\right).$$

These derivatives have been written in this particular way in order to show clearly that the toll for slow vehicles is always larger than that for the fast vehicles, and especially so in situations in which the road is not heavily used ($c - (\mu_1 + \mu_2)$ is large) and demand for trips from slow vehicles (μ_2) is low.

Derivations for the situation with overtaking

In this part of the appendix we derive expressions for φ_1 and φ_2, the expected values of the periods during which the high and low speeds are used. We start with φ_1.

If a slow car is overtaken, the fast driver drives a distance of at least $2d^*$ at the high speed. The distance between two slow cars is exponentially distributed with parameter λ_2/s_2. The expected value of this distance is therefore s_2/λ_2. Note also that the fast car that overtakes a slow car uses already d^* units of this distance as soon as overtaking is completed.[19] The distance that can be travelled at the high speed before another slow car has to be overtaken or speed has to be reduced is therefore on average equal to $d^* + s_2/\lambda_2$. The time involved for the fast driver will be denoted as τ_1:

$$\tau_1 = \frac{d^* + s_2/\lambda_2}{s_1 - s_2}.$$

If the distance to the next slow car becomes minimal, the high speed can only be maintained if overtaking can start immediately. We therefore have to determine the probability π_1 that there is a sufficiently large gap in traffic on the left lane at the moment the fast driver needs to start overtaking. To do so, first observe that overtaking can never take place if there is another car present on the left lane at the location where overtaking must start. The probability that such a situation occurs is equal to $d^*/(d^* + s_3/\lambda_3)$.[20] If no car is present at the location on the left lane where overtaking has to start, the distance to the first car on the left lane has to be equal to or greater than g. The probability that this is the case equals $\exp(-\lambda_3 g/s_3)$. We therefore find:

$$\begin{aligned}\pi_1 &= \left(1 - \frac{d^*}{d^* + s_3/\lambda_3}\right) e^{-\lambda_3 g/s_3}, \\ &= e^{-\lambda_3 g/s_3}/(d^*\lambda_3/s_3 + 1).\end{aligned}$$

Finally, we have to take into account that the total length of the period during which the high speed can be maintained is a multiple of (sub)-periods with expected duration τ_1:

$$\varphi_1 = \tau_1(1 - \pi_1) + 2\tau_1\pi_1(1 - \pi_1) + 3\tau_1\pi_1^2(1 - \pi_1) + \cdots,$$

[19] We ignore complications that may arise from the possibility that two slow cars are so close to each other that overtaking them one by one is impossible. We simply assume that the last of these slow cars throttles back in order to give the overtaking car enough distance to remove to the right lane if this is necessary and moves back to the original position behind the first of the two slow cars as soon as possible.

[20] The length of a car has here been taken to be equal to d^*.

and this can be elaborated as follows:

$$\varphi_1 = \tau_1(1-\pi_1)\sum_{i=0}^{\infty}\sum_{j=i}^{\infty}\pi_1^j$$

$$= \tau_1(1-\pi_1)\sum_{i=0}^{\infty}\pi_1^i/(1-\pi_1)$$

$$= \tau_1/(1-\pi_1).$$

We move on to the derivation of φ_2. Such a period starts when a fast car is forced to slow down. Immediate overtaking may be impossible, either because the left hand side is occupied by a car moving in the opposite direction, or because a car on the left lane is within distance g. In the first case, the expected value of the time until a gap occurs on the left lane equals $d^*/2(s_2+s_3)$, which is small, and we ignore this part of the total delay in what follows.

If a car is approaching on the left lane within distance g, the expected time until it passes the fast driver who is forced to slow down on the right lane is equal to $E(x|x<g)/(s_2+s_3)$. We elaborate the expected value:

$$E(x|x<g) = s_3\frac{\int_0^{g/s_3} x\lambda_3 e^{-\lambda_3 x}\,dx}{\int_0^{g/s_3}\lambda_3 e^{-\lambda_3 x}\,dx},$$

$$= s_3\frac{1/\lambda_3\left(1-(1+\lambda_3 g/s_3)e^{-\lambda_3 g/s_3}\right)}{1-e^{-\lambda_3 g/s_3}},$$

$$= \frac{s_3}{\lambda_3} - g\frac{e^{-\lambda_3 g/s_3}}{1-e^{-\lambda_3 g/s_3}}.$$

If it is taken into account that the fast driver also has to wait until the oncoming car has passed, we can write down the expected value of the time τ_2 it takes until the first oncoming car is passed and overtaking may become possible:

$$\tau_2 = \frac{1}{s_2+s_3}\left(d^* + \frac{s_3}{\lambda_3} - g\frac{e^{-\lambda_3 g/s_3}}{1-e^{-\lambda_3 g/s_3}}\right).$$

The probability π_2 that the gap that follows the first oncoming car will be at least equal to g equals:

$$\pi_2 = e^{-\lambda_3 g/s_3}.$$

The expected value φ_2 of the period during which speed has to be equal to s_2 can now be written as:

$$\varphi_2 = \tau_2\pi_2 + 2\tau_2(1-\pi_2)\pi_2 + 3\tau_3(1-\pi_2)^2\pi_2 + \cdots,$$

and this can be elaborated to:

$$\varphi_2 = \tau_2/\pi_2.$$

In order to show that the expression for the expected values of speed and travel time are appropriate, call t_1 the time during which speed s_1 can be maintained and t_2 the time during which the lower speed s_2 is relevant. Let n be the number of events. The total number of kilometres driven during n events is:

$$s_1 \sum_{i=1}^{n} t_1^i + s_2 \sum_{i=1}^{n} t_2^i, \tag{i}$$

whereas the total time this takes is:

$$\sum_{i=1}^{n} t_1^i + \sum_{i=1}^{n} t_2^i. \tag{ii}$$

Average speed equals the ratio of (i) and (ii), and if numerator and denominator are both divided by the number of events n, we find the following expression for average speed:

$$\frac{s_1\left(\sum_{i=1}^{n} t_1^i/n\right) + s_2\left(\sum_{i=1}^{n} t_2^i/n\right)}{\left(\sum_{i=1}^{n} t_1^i/n\right) + \left(\sum_{i=1}^{n} t_2^i/n\right)}. \tag{iii}$$

Now observe that:

$$\lim_{n\to\infty} \sum_{i=1}^{n} t_1^i/n = \varphi_1, \qquad \lim_{n\to\infty} \sum_{i=1}^{n} t_2^i/n = \varphi_2,$$

and the result for $A(s)$ follows by taking the limit of (iii) for $n \to \infty$ The result for $A(w)$ is derived similarly from the expression for the ratio of the total time spent travelling during n events and the number of kilometres driven, that is, the inverse of (iii).

Optimal tolls when overtaking is possible

Again, we maximise social surplus. Now we have three groups of vehicles. The Lagrangean is:

$$L = \sum_{i=1}^{3} \int_{0}^{\mu_1} g_i(m_i)\, dm_i - v_i w_i \mu_i + \eta_i \left(g(\mu_i) - v_i w_i - z_i\right)$$

(where we ignore the non-negativity constraints).

The travel times of the vehicles belonging to groups 2 and 3 are constant. The travel time of vehicles belonging to group 1 is equal to $w_1 = lA(w)$(see equation (15) of the main text) with l the length of the road. Hence w_1 is a function of the arrival rates λ_2 and λ_3 and, through them, of the demands μ_1, μ_2 and μ_3.

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*, 83 161–79.

Barzily, Z. and M. Rubinovitch (1979): "On platoon formation on two-lane roads," *Journal of Applied Probability*, 16, 347–61.

Huisman, T. and R. J. Boucherie (2001): "Running times on railway sections with heterogeneous train traffic," *Transportation Research*, B35, 271–92.

Newell, G. F. (1998): "A moving bottleneck," *Transportation Research*, B32, 531–37.

SWOV (1990): *Naar een duurzaam veilig wegverkeer (Towards sustainably safe road traffic)* (in Dutch) Leidschendam.

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

Verhoef, E. T., J. Rouwendal and P. Rietveld (1999): "Congestion caused by speed differences," *Journal of Urban Economics*, 45, 533–56.

Wardrop, J. (1952): "Some theoretical aspects of road traffic research," *Proceedings of the Institute of Civil Engineers*, 1, 325–78.