Verhoef, Erik Teodoor

Conference Paper

Time, speeds, flows and densities in static models of road traffic congestion and congestion pricing

38th Congress of the European Regional Science Association: "Europe Quo Vadis? - Regional Questions at the Turn of the Century", 28 August - 1 September 1998, Vienna, Austria

Provided in Cooperation with:

European Regional Science Association (ERSA)

Suggested Citation: Verhoef, Erik Teodoor (1998) : Time, speeds, flows and densities in static models of road traffic congestion and congestion pricing, 38th Congress of the European Regional Science Association: "Europe Quo Vadis? - Regional Questions at the Turn of the Century", 28 August - 1 September 1998, Vienna, Austria, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:
http://hdl.handle.net/10419/113457

TIME, SPEEDS, FLOWS AND DENSITIES IN STATIC MODELS OF ROAD TRAFFIC CONGESTION AND CONGESTION PRICING

Erik T. Verhoef*
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

Paper presented to the 38th meetings of the European Regional Science Association
Vienna, 28 August – 1 September 1998

Key words: Road traffic congestion, congestion pricing, hypercongestion
JEL codes: R41, R48, D62

Abstract
This paper deals with some of the features of static models of road traffic congestion that have caused much debate in the literature. It first focuses on the difficulties arising with the backward-bending cost curve defined over traffic flows in the context of 'continuous congestion'. The relevance of the backward-bending segment of this curve is questioned by demonstrating that the 'equilibria' on this segment of the cost curve are dynamically infeasible. Next, the implications for static models of 'peak congestion' are considered. In doing so, attention is paid also to the implicit assumptions, particularly on the nature of scheduling costs, that are necessary to render static models of peak congestion internally consistent. The paper ends with a brief discussion of the implications for dynamic models of peak congestion.

*Erik Verhoef is affiliated as a research fellow to the Tinbergen Institute. The author would like to thank Robin Lindsey, Kenneth Small, Olof Johansson and two anonymous referees for very useful and inspiring comments on earlier versions of this paper. The usual disclaimer applies.

1. Introduction

Notwithstanding the long history of the economists' and engineers' study of road traffic congestion and congestion pricing (see Pigou, 1920; Knight, 1924; Wardrop, 1952; Walters, 1961; and Vickrey, 1969), academia has not yet reached general consensus on the fundamentals that should underlie such analysis, witness the relatively large number of comments and replies that papers on these topics seem to trigger (Else, 1981, 1982, versus Nash, 1982; De Meza and Gould, 1987, versus Alan Evans, 1992; Andrew Evans, 1992, 1993, versus Hills, 1993; and Lave, 1994, 1995, versus Verhoef, 1995). However, especially now that the introduction of road pricing as a means of combating congestion becomes an increasingly realistic policy option at various places (Small and Gomez-Ibañez, 1998), the importance of at least transport scientists coming to a closer agreement on the analytical backgrounds underlying the phenomenon of road traffic congestion and the derivation of optimal fees increases likewise.

This paper is concerned with static economic models of road traffic congestion. Although static models have obvious limitations in the analysis of traffic congestion, they are still often used for both research and educational purposes. It is therefore worthwhile to consider these models in some further detail. This paper addresses some of the key questions that have dominated the debate on these models in the literature. These include issues related to hypercongestion and the backward-bending segment of the average cost curve that can be derived from the 'fundamental diagram of road traffic congestion'. A related question concerns the choice of the output variable in the definition of demand and cost functions. Two main stances can be distinguished here: 'flow-based' measures, where the output measure has an explicit 'per-unit-of-time' dimension (De Meza and Gould, 1987; Andrew Evans, 1992, 1993; Else, 1981, 1982 and Nash, 1982), and 'stock-based' measures, such as densities or numbers of trips (Alan Evans, 1992; Hills, 1993; Verhoef *et al.*, 1995ab, 1996ab).

Although the time dimension is by definition not considered explicitly in static models, it will turn out that this does not mean that time as such does not, or should not, play any role at all. 'Static' only means that these models do not explicitly study (or allow for) changes of variables over time. In what follows, some elements that would perhaps sooner be associated with dynamic modelling, and that are consequently often ignored in static analyses, will play an important role. Apart from making a fundamental distinction between 'peak demand' and 'continuous demand', the question of whether a proposed static equilibrium is dynamically consistent will be considered explicitly, and is taken as a prerequisite for the equilibrium to be meaningful.

The paper is organized as follows. Section 2 discusses some main features of static models of road traffic congestion, presents some definitions, and distinguishes between models directed towards the cases of 'continuous demand' and 'peak demand'. Section 3 proceeds by investigating the case of continuous demand, and focuses in particular on the difficulties arising with the backward-bending cost curve. The relevance of the backward-bending segment of this curve is questioned by demonstrating that 'equilibria' on this segment are dynamically inconsistent. Section 4 studies the implications for static models of peak congestion. In doing so, attention is paid also to the implicit assumptions, particularly on the nature of scheduling costs, that are necessary to render static models of peak congestion internally consistent. Because these assumptions turn out to be rather unrealistic, the section also discusses the

implications of the analysis in Section 3 for dynamic models of peak congestion. Finally, Section 5 concludes.

2. Static models of road traffic congestion: some introductory issues

Road traffic congestion in reality is a complicated dynamic process, and the analyst studying congestion and congestion pricing is soon confronted with the dilemma between using either an 'as-realistic-as-possible' modelling approach, in which analytical solutions are often difficult to obtain (such as Newell, 1988), or to apply a simpler representation of reality, allowing analytical solutions and the derivation of more or less general insights into the economic principles behind the problems studied. This paper is concerned with the latter type of approaches. Within this group of models, a distinction can be made between static and dynamic models. In static models, no explicit time dimension is present. Speeds, densities, generalized costs and the toll in case one is levied are, as it were, constant over time: they only have one single equilibrium value. This may often be at odds with standard results in dynamic models of road traffic congestion, where for instance travel times and optimal tolls usually vary over the peak (see Arnott *et al.*, 1993, 1998; Braid, 1989, 1996; Henderson, 1974, 1981; and Chu, 1995), but it is simply a property of static models. Still, the time factor does play an important role in these static models. No matter how abstract it may seem in a 'time-less' approach, the average generalized travel costs are assumed to increase with road usage because speeds decrease and travel times increase. Moreover, as in any static economic model of a market, a consistent pair of demand and supply relations can be specified only after the time period to which they pertain has been identified (e.g., the average daily demand for a certain good at a given price will be one-seventh of the average weekly demand).

Static models of road traffic congestion give a simplified representation of reality by definition. For an unambiguous interpretation of these models it is, however, important to make explicit exactly what type of real process they aim to represent. From that perspective, it is important to distinguish between models dealing with 'peak demand' on the one hand, and 'continuous demand' on the other. 'Peak demand' refers to the case where a limited number of potential users consider using the road during the same (peak) period, the duration of which could be endogenized. The equilibrium number of actual users will depend on the equilibrium level of user costs during this peak period, and the intersection of the inverse demand curve with the horizontal axis gives the total number of road users during the peak in the hypothetical case where user costs were zero – with zero usage and an empty road before and after they have travelled. 'Continuous demand', in contrast, refers to the case where the demand function is stable over time. This would normally result in an everlasting 'stationary state' situation, where a road is continuously used at a constant intensity. For the description of such a stationary state, the fact that speeds, flows, and densities each will have just one single equilibrium value in a static model could be less unrealistic than it might be for static models of peak demand. Here, the intersection of the inverse demand curve with the horizontal axis gives the constant, everlasting number of users completing their trip *per unit of time* in case user costs were zero. One could of course consider mixed cases, where a 'peak demand' interferes with some 'continuous demand', but the main point here is just to distinguish between these two basic types of demand.

Let us now turn to the various relevant variables. In defining these, one should in the first place consider one single, well-defined market. The product 'trip' should therefore be

homogeneous, which can be assured by considering a single road, to be used either completely or not at all by individual drivers. Hence, all trips are assumed to have equal length (L). Differences between trips in terms of speeds or arrival times, which could especially be relevant in models of peak congestion, do not affect this homogeneity condition as long as such variations are reflected in differences in generalized user costs.

A number of measures for 'road usage' can be distinguished. The first of these is total road usage (N): the total number of trips completed, over the entire period considered. This variable is relevant only for the case of peak demand. With continuous demand, total usage is either zero, or increasing with the time period considered. A second measure for usage is flow (F), measuring the number of vehicles passing a given point on the road per unit of time. A third measure related to usage is density (D): the number of users per unit of road space – where the total road space, in turn, can be measured as the product of two constants, namely length L and width W, usually the discrete number of lanes. D and F are relevant measures for peak demand as well as for continuous demand. Finally, the variable n will be used to represent the number of users that are simultaneously present on the road.

Road users are in the most basic model identical in all respects, except for having a possibly different maximum willingness to pay for making a trip. In particular, they share the same value of time, and they all contribute to congestion in the same way. The speed (S) has, through the value of time, an important impact on the generalized average social costs (AC). Like most models, only time costs will be considered in what follows, although other cost components could easily be introduced without affecting the results.

Apart from speeds and flows, there are some other 'time-related' variables that may be relevant for the static analysis of congestion. The duration of the peak (T), for instance, is often ignored in static analyses of congestion, even when dealing with peak demand. However, when a 'trip-based' demand function is used, giving the total number of trips demanded during the peak as a function of the equilibrium level of generalized user costs, the related cost function can be defined only when the duration of the peak T is known. The reason is that the average and marginal social costs of having a number of users completing a trip during the peak will generally depend on the duration of the peak: the longer the duration, the lower these costs.[1] Two measures for the duration of the peak will be used below. The duration T, without further qualification, denotes the period between the first and last driver in the peak passing a given point along the road. The 'grand duration' T_G will give the time-span between the first driver's arrival time at the entrance of the road, and the last driver's arrival time at its exit. The last time-related variable is the duration of a trip (t): the time it takes to drive from the road's entrance to its exit. This measure is relevant both for peak demand and continuous demand. Therefore, the grand duration of the peak, in a purely static model with peak demand, is equal to $T_G=T+t$, because the peak starts and ends t seconds later at the road's exit than at its entrance.

With continuous demand, the following three identities can be given for a stationary state equilibrium:

[1] Alternatively, when using a 'flow-based' cost function in a static model of peak congestion, giving the generalized costs as a function of passages per unit of time, the transformation of the demand curve defined over total numbers of trips into the then relevant demand function defined over flows can also be made only when the duration of the peak is known.

$$F \equiv \frac{n}{t} \tag{1a}$$

$$S \equiv \frac{L}{t} \tag{1b}$$

$$D \equiv \frac{n}{L \cdot W} \tag{1c}$$

Equations (1b) and (1c) are evident; equation (1a) can be checked by observing that all users present on the road at a certain instant will have passed the point of exit after t time units.

Recalling that in a purely static model of peak congestion, all variables have one single value in equilibrium, and should therefore have the same value at each instant and at each place along the road as long as it is used at that instant at that place, (1a)–(1c) will have the following counterparts for the purely static model of peak congestion:

$$F \equiv \frac{N}{T} \tag{2a}$$

$$S \equiv \frac{L}{t} \tag{2b}$$

$$D \equiv \frac{N \cdot \frac{t}{T}}{L \cdot W} \tag{2c}$$

Equations (2a) and (2b) are evident; equation (2c) can be understood after realizing that D can be determined as if N·t/T vehicles were present simultaneously on the road during the time-span T. It should be emphasized that, like (2a) and (2b), (2c) only holds for places along the road at instants that it is actually used. During the first (and last) t time units, when a decreasing (increasing) segment of the road is empty, it would be incorrect to derive the density by dividing the number of users present at an instant by the total length of the road, because this would incorrectly assume these users to be distributed uniformly along the entire road. Outside the duration of the peak for a certain point along the road, we simply have F=D=0, and speed is not defined.

Finally, it is easily checked that both (1a)–(1c) and (2a)–(2c) are consistent with the well-known property that traffic flow is proportional to the product of density and speed:

$$F = D \cdot S \cdot W \tag{3}$$

W is often implicitly set at unity, and then disappears from (3).

3. The case of continuous congestion

The case of continuous congestion, where the demand function for road usage is stable over time for an infinite time period – or at least long enough to allow one to concentrate on stationary state equilibria – is probably not the most realistic representation of road traffic congestion. Nevertheless, it is the situation that is often, implicitly, assumed to apply in static models of congested road traffic; in particular those based on the ‘fundamental diagram of road traffic congestion’ (see, for instance, Johansson, 1997). The model can be seen as a basic ‘bench-mark’ model for studying the economics of congestion, the insights of which can be helpful in interpreting more realistic and complicated models in which, for instance, demand curves are not stable, or not independent, over time. Even this probably most simple representation of road traffic congestion, however, has triggered a remarkable level of

disagreement, which justifies further study of the model. The model also allows one to study some fundamental differences between 'ordinary' static economic market models and the market model of congested road usage, apart from the additional complications related to the duration of the peak. The latter will be considered explicitly in Section 4.

3.1. The standard analysis

From equations (1a)–(1c), it turns out that the equilibrium flow F can be written only as a function of at least two endogenous variables; for instance as the ratio of n and t according to (1a), or as the product of D and S, multiplied by the constant W, according to (3). Therefore, to find the relation between equilibrium values of F and one of these other variables, say S, one has to take into account the relations between these endogenous variables; for instance the relation S(D). It is normally assumed that speed decreases with an increasing density. This is illustrated by the density-speed relation (DS-curve) in the first quadrant in Figure 1, which is the so-called 'fundamental diagram' of road traffic congestion. As drawn, it is assumed that the free-flow speed S^* can be sustained for positive densities (the DS-curve starts with a flat segment); and that there is some maximum density D_{max} for which speed falls to zero.

Because F is proportional to the product of D and S by (3), F will obtain a maximum value for some combination of speed and density, denoted $S^\#$ and $D^\#$ in the diagram. This gives rise to the familiar backward-bending speed-flow curve (SF-curve) in the fourth quadrant of Figure 1, and the density-flow curve (DF-curve) in the second quadrant of Figure 1. Under the assumption that only time costs matter for generalized user costs, the speed-flow curve in Figure 1-IV can subsequently be combined with the inverse relation between speed and travel times in (1b) to obtain the standard backward-bending average social cost function (AC) depicted in Figure 2. The lower section of the AC-curve, where speeds are relatively high and travel times relatively short, corresponds with the upper section of the SF-curve in Figure 1-IV. Likewise, the upper section of the AC-curve, representing situations that are usually referred to as 'hypercongestion', corresponds with the lower section of the SF-curve. As speeds go to zero in Figure 1-IV, generalized user costs go to infinity in Figure 2.

Therefore, each level of flow, except the maximum level and zero flow, appears to be obtainable at two cost levels: a low one, where the density is relatively low and the speed relatively high; and a high one, where the opposite holds. It is especially the backward-bending cost or supply curve in Figure 2 that has led to heated debate, mainly because the confrontation of this curve with a standard downward sloping demand curve may produce puzzling results. Before discussing these, it should be emphasized that the translation of the SF-curve into the AC-curve through the relation between speed and travel times presupposes that speeds, densities and flows are constant over time and along the road. AC as a function of F would otherwise not be meaningful, because F and S themselves would then vary during the trip, and t=L/S could not simply be applied to derive the average cost for a trip.

In Figure 2, two demand curves, denoted E and E′, are included. These give the marginal willingness to pay for making the trip as a function of traffic flow, which, in a stationary equilibrium, is equal to the number of trips completed (and started) per unit of time. These demand curves therefore do not give the marginal willingness to pay to pass that

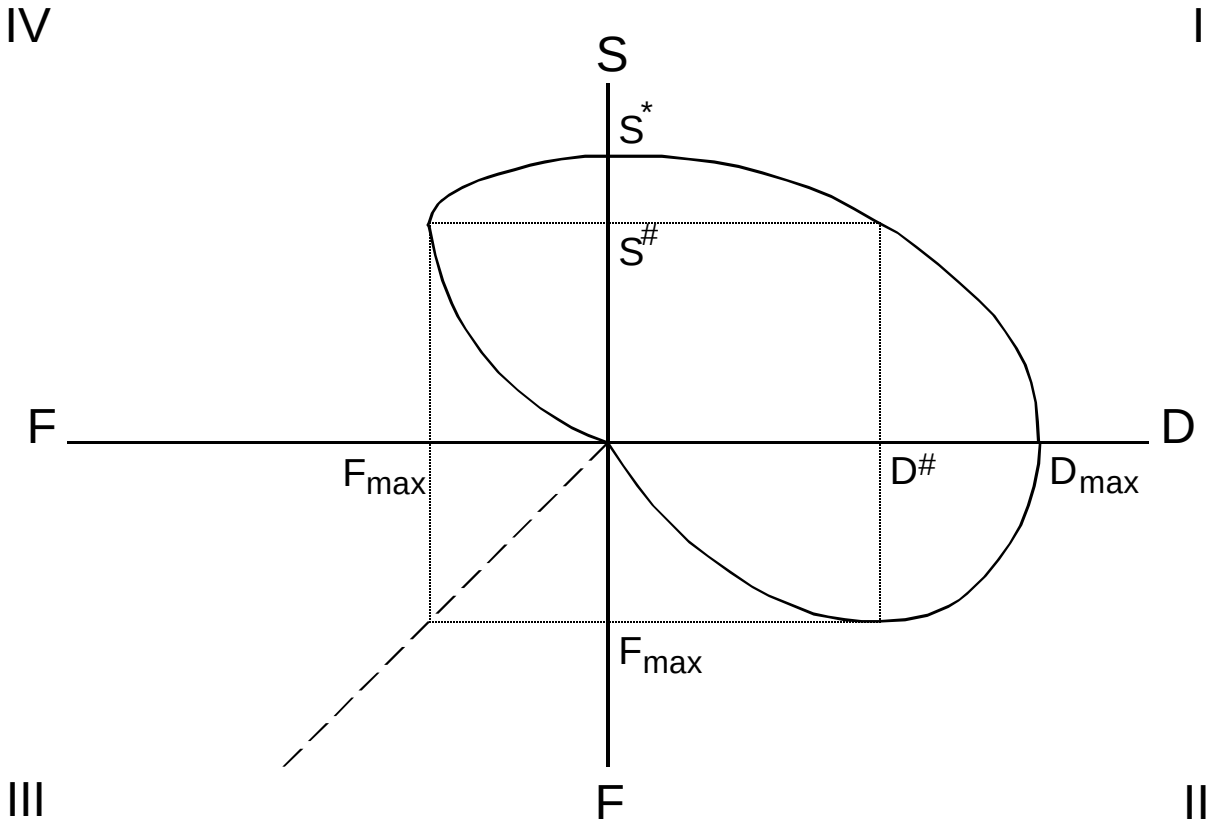

Figure 1. The density-speed curve (I), the speed-flow curve (IV) and the density-flow curve (II)

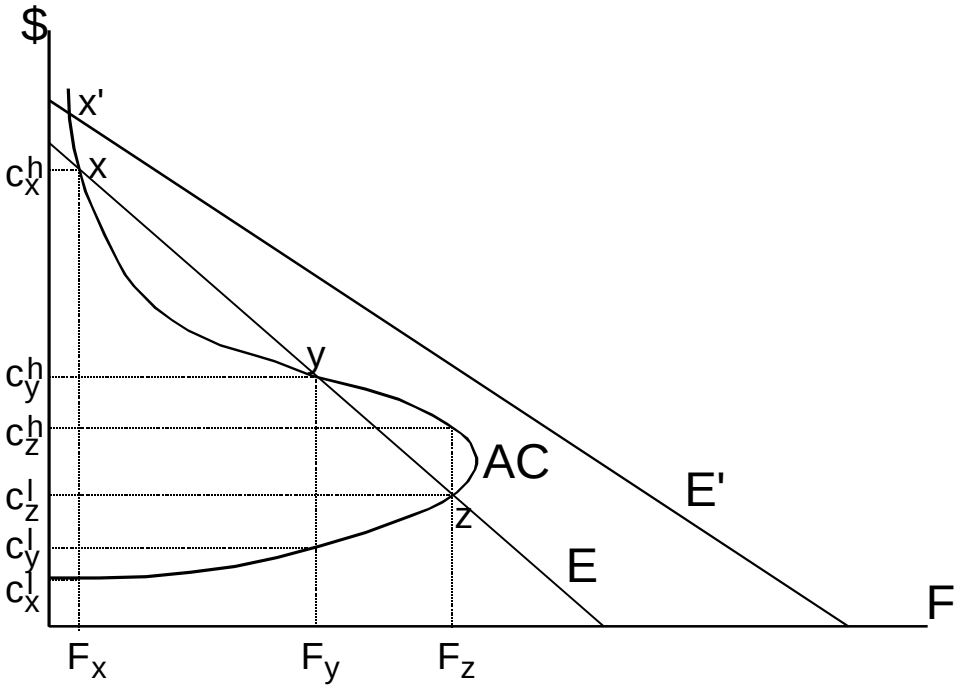

Figure 2. The backward-bending average cost curve (AC) and two demand curves (E and E¢) defined over traffic flow (F)

particular point where flow is measured, at the particular instant it is measured – unless the rest of the trip could be made against zero costs. Because the AC-curve was designed to give the generalized user costs of the entire trip as a function of traffic flow, also the demand curve should give the marginal benefits of the entire trip.

The demand curve E is drawn so as to produce the three possible types of intersection with the average cost curve that can be distinguished. In contrast to standard market models, where unique equilibria are usually found, this curve seems to suggest no less than three possible equilibria, denoted x, y, and z. At each of these points, the marginal benefits E are equal to the generalized costs AC. Among these points, z is nearest to a standard market equilibrium, with an upward sloping supply and a downward sloping demand curve. The intersections x and y, in contrast, suggest that market equilibria could also occur on the upper segment of the AC-curve, in which case a non-intervention stationary state equilibrium with hypercongestion would arise. A standard argument is that one of the objectives of a toll is to secure the transition to the lower segment of the AC-curve, because in an optimum, the flow should be realized at the minimum possible cost. Apart from that, the toll should bridge the gap between AC and marginal social cost (MSC). The latter, however, is ignored for the time being, in order to keep the diagram decipherable.

When multiple equilibria seem possible, a logical next step is to investigate the local stability of the candidates. Interestingly, in the present context, analysts do not agree on this question. For instance, Nash (1982) asserts that equilibria like y are stable, where the demand curve is steeper than the AC-curve, while Else (1982) proposes x, where the opposite holds. Although this issue is only of limited relevance for the sequel, because it will be argued that neither x nor y are dynamically consistent stationary states, one can explain the disagreement from the type of perturbations considered. For *price perturbations*, in line with the Walrasian tâtonnement process, x appears locally stable and y unstable. For a slightly higher price, an excess supply (demand) is then found at x (y), leading to a downward (upward) price adjustment by the auctioneer, and hence a move back to (further away from) the initial equilibrium. For *quantity perturbations*, the opposite holds. A slight increase in usage cause marginal benefits to exceed (fall short of) average cost at x (y), leading to a move further away from (back to) the initial equilibrium. Note that z is stable according to both approaches.

3.2. Reconsidering the standard analysis

The representation in Figure 2 has been challenged in the literature (see Chu and Small, 1996, for a recent contribution). One particular problem has received relatively little attention, and that is the situation where a demand curve like E′ would apply. In that case, only the equilibrium x′ remains. If one believes that quantity perturbations are the correct way to evaluate local stability – which actually does seem more appropriate in the absence of tolling – this equilibrium is unstable. Beyond that flow, flows will continue to increase, because marginal benefits consistently exceed the average user costs. Road users can ‘avoid the cost curve altogether’, and will presumably end up at the intersection of E′ with the horizontal axis, expecting free trips for ever. It is evident on intuitive grounds that this cannot be correct. However, the model presented so far is unable to explain what will happen in that case.

Several analysts have expressed unease with the choice of flow as an output variable in Figure 2, because flow is an “…endogenous variable, resulting from the […] interactions

among road users" (Alan Evans, 1992, p. 212). Hills (1993), for instance, suggests that the total number of trips accomplished should be the relevant output variable. After the discussion in Section 2, it will be clear that this measure certainly makes sense in the case of peak congestion. However, it is a meaningless concept in a stationary state equilibrium with continuous congestion, because it is then either equal to zero, or increases with the time period considered. Alan Evans (1992) proposes densities. However, this output variable has the unattractive implication that it is 'being on the road' that people demand, instead of 'completing a trip'. In particular, in the stationary state where density is at a maximum, and speed and flow are zero forever, a demand curve defined over density suggests maximum benefits, although not a single trip is ever completed. Density could be a reasonable output measure in models of, for instance, congested beach tourism. For road usage however, it seems that in case of continuous demand, one should maintain normalization with respect to the time dimension in the definition of the output variable. This is fully consistent with common practice for the specification of demand and cost curves in static economic market models (Else, 1982 (p. 300)). However, this does not mean that the endogenous variable flow, as used in Figure 2, should be the actual output measure. Instead, it will be proposed below to use the arrival rate of new users at the road's entrance: the number of trips started per unit of time. This variable, r, is equal to the flow only in stationary state equilibria. Hence, for stationary states, the demand curve E(F) should be the same as the demand curve E(r).

A second stance taken in this paper is that the 'dynamic consistency' of a static equilibrium should be taken as a prerequisite for this equilibrium to be meaningful. This requirement implies for the case of continuous congestion that a proposed static equilibrium, where all variables have one single equilibrium value, can only be meaningful if these values correspond to the long run stationary state values that these variables could or would obtain in a corresponding dynamic model. There are two conditions that guarantee a static equilibrium to be dynamically consistent. The *stationary state condition* is that the static equilibrium should not somehow imply growing or declining stocks, in which case the equilibrium would be nothing more than a 'snapshot' of an ever-changing system, rather than representing the system's long run equilibrium. The *feasibility condition* requires that the equilibrium is dynamically stable, and hence could result from at least some set of (internally consistent) initial conditions other than the conditions applying in that stationary state itself. If either of these conditions is violated, the static equilibrium loses most of its appeal and relevance, as it then does not represent a possible stationary state outcome of the dynamic process it aims to describe.

In most static market models, this question of dynamic consistency is ignored, because it is implicitly assumed that a certain good is produced, traded and consumed within a single time period. As a result, no stocks accumulate or decline over time; prices, production levels and consumption levels at different instants do not interact; and there are no inter-temporal externalities. However, for the market considered here, such an assumption is clearly unrealistic. The speed, and hence the costs, that a driver obtains during the trip will generally not be independent of the travel conditions on the road (just) before he starts. By distinguishing between r and F, and by acknowledging that F is also dependent on previous values of r, instead of imposing beforehand that for all points along the road we have F=r at every instant (as is implicitly done in Figure 2), this inter-temporal cost interdependence can be

taken into account. Hence, the consideration of both r and F, and the imposed prerequisite of dynamic consistency, are closely connected.

A number of propositions will be used to assess the dynamic consistency of all points on the AC-curve. The first of these are related to the stationary state condition, and can be made without making explicit the drivers' behaviour – and hence the model's behaviour – during transitional phases. The only assumption that should be made explicit is that the road's maximum capacity F_{max} is constant along the road, including the entrance.

Proposition 1a All points on the AC-curve in Figure 2 can be stationary states.

Proof It will be proven that, starting from an initial stationary state with a consistent set of S_0, D_0 and $F_0=S_0 \cdot D_0<F_{max}$ according to the AC-curve, if we have an arrival rate of new users at the entrance $r_0=F_0$, the stationary state equilibrium sustains itself. This can be shown by considering what happens near the road's entrance during an arbitrarily short time frame of τ seconds between two clock-times τ_0 and τ_1, with $\tau_1-\tau_0=\tau$. At τ_1, the last drivers that arrived at τ_0 will have moved a distance of $d=\tau \cdot S_0$ meters. The available road space for those arriving between τ_0 and τ_1 is therefore $W \cdot \tau \cdot S_0$. The number of newly arrived cars is $\tau \cdot r_0$, implying that the (average) density D_d over the first $d=\tau \cdot S_0$ meters can be written as:

$$D_d = \frac{t \cdot r_0}{W \cdot t \cdot S_0} = \frac{r_0}{W \cdot S_0} = \frac{F_0}{W \cdot S_0} = D_0 \qquad (4)$$

(compare (3) for the last step). Note that the result is independent of the time frame τ considered, and therefore also holds for $\lim_{t \to 0} D_d$. Provided $r_0=F_0$, the density near the entrance will therefore remain constant and equal to the density D_0 that is consistent with the initial speed S_0 and flow F_0.■

Proposition 1b If $r>F_{max}$, the system cannot be in a stationary state equilibrium.

Proof If $r>F_{max}$, somewhere a stock must be accumulating at a rate $q \geq r-F_{max}>0$.■

In order to test the dynamic consistency according to the feasibility condition, one has to be more specific about the model's behaviour during transitional phases than the fundamental diagram allows. This diagram presupposes and subsequently produces stationary states only. Speed is only defined for a constant density along the road, and because all drivers along the road will as a consequence obtain the same speed, the density will also *remain* constant over time and place (drivers do not get closer to, or further away from each other). In testing the dynamic feasibility of equilibria, one would ideally use a full-fledged dynamic model, which should in stationary states be consistent with the fundamental diagram underlying the static model. Verhoef (1998) presents one such model, based on the identities that density D as given in (1c) and (2c) is the inverse of the distance between two subsequent cars for a single-lane road, and that the arrival rate r is the inverse of the time elapsed between the arrival of two subsequent cars at the entrance. A 'car-following' model is then be specified, which for stationary states yields a density-speed relation like the one given in Figure 1-I.

In such a model, however, the determination of the position of subsequent cars along the road over time involves solving differential equations, and the model unfortunately does not yield analytical expressions that can easily be used for the present purpose. Moreover, since the aim here is to prove that the points on the upper segment of the AC-curve are infeasible, the car-following assumption that only the distance to the preceding car (not to the following car) matters in the speed choice could be perceived as too restrictive. Therefore,

only a minimum number of rather mild assumptions are used below to describe the drivers' behaviour during transitional phases. For this purpose, the 'forward local density' $d_{f,x}$ ('backward local density' $d_{b,x}$) at a point along the road is defined as the average density over the first x meters downstream (upstream), where x can have any value as long as it does not exceed the distance to the road's exit (entrance). Next, $d_{f,max}$ ($d_{b,max}$) gives the maximum value for $d_{f,x}$ ($d_{b,x}$) that can be found by varying x. The assumptions made then are that:

(1) At a certain instant, a driver will not drive slower than $S(\max\{d_{f,max}, d_{b,max}\})$, where $S(\cdot)$ gives the density-speed relation as given in Figure 1-I.
(2) When a driver, who previously drove in stationary state conditions, observes that the nearest driver behind him slows down, so that $d_{b,x}$ decreases for some relatively small but positive values of x, he will not slow down himself but maintains the stationary state speed, even if $d_{b,x}$ simultaneously increases for some relatively large values of x.
(3) A driver will not voluntarily cause hypercongestion himself:
    (a) at the instant of starting a trip, when $d_{b,x}$ is not defined, a driver will not select a speed below $S(d_{f,max})$, for instance in anticipation of the high $d_{b,x}$ that would subsequently result from this choice itself;
    (b) during a trip, a driver will not select a speed below $S^{\#}$ when all speeds downstream exceed $S^{\#}$ and when hypercongestion would only be building up behind him because of his own choice to drive slowly.

The second assumption makes sure that decreasing speeds upstream do not work as a 'vacuum' by hindering speeds downstream, and thus limits the potential consequences of the very mild assumption (1) somewhat. The following propositions can now be derived:

Proposition 2a Starting from a stationary state 0 $\{S_0, D_0, r_0=F_0=S_0\cdot D_0<F_{max}\}$ without hypercongestion, there is no arrival rate $r_1\leq F_{max}$ that would lead to hypercongestion.

Proof First, define state 1 $\{S_1, D_1, r_1=F_1=S_1\cdot D_1\}$ as the 'non-hypercongested' stationary state consistent with r_1, and define τ_0 as the clock-time from which moment onwards r_1 applies. If the last drivers that arrived just before τ_0 would maintain S_0 throughout their trips, at clock-time $\tau_0+\tau$ the average density D_d over the first $d=\tau\cdot S_0$ meters can, for each τ, be written as:

$$D_d = \frac{t\cdot r_1}{W\cdot t\cdot S_0} = \frac{r_1}{W\cdot S_0} < \frac{F_{max}}{W\cdot S^{\#}} = D^{\#} \qquad (5a)$$

where the inequality follows from $r_1\leq F_{max}$ and $S_0>S^{\#}$.

If $r_1<r_0$, then $D_1<D_d<D_0<D^{\#}$ for all τ. The assumption that the drivers who arrived before τ_0 maintain S_0 is therefore in accordance with assumption (1): for those drivers $d_{f,max}=D_0>d_{b,max}$, because $D_d<D_0$ for all τ.

If $r_1>r_0$, one should take account of the possibility that the drivers who arrived just before τ_0 will not maintain S_0, because a density higher than D_0 is building up behind them. However, the drivers who started their trip again just before these drivers will maintain S_0 by assumption (2). Therefore, since $r_0<r_1$ and $S_0>S_1$, the last drivers who arrived just before τ_0 will have speeds not exceeding S_0 but strictly exceeding S_1 by assumptions (1) and (3b). In particular, observe that because they have started their trips at a speed S_0, $d_{b,max}<D_1$ for these drivers throughout their trips. Hence, (5a) can be written as:

$$D_d < \frac{t\cdot r_1}{W\cdot t\cdot S_1} = \frac{r_1}{W\cdot S_1} < \frac{F_{max}}{W\cdot S^{\#}} = D^{\#} \qquad (5b)$$

Since (5ab) hold for all values of τ, a density consistent with hypercongestion can never build up on the road, and as a consequence, speeds will consistently remain larger than $S^{\#}$. In particular, observe that both for (5a) and (5b) we find:

$$\lim_{t\to 0} D_d < D^{\#} \tag{5c}$$

implying that at τ_0, the speed at the entrance remains greater than $S^{\#}$. This in turn implies that the reasoning leading to (5ab) can be reapplied for the entrance for every instant after τ_0.■

Proposition 2b Starting from *any* initial situation – including non-stationary ones – without hypercongestion, where speeds exceed $S^{\#}$ and densities are below $D^{\#}$ along the entire road, there is no arrival rate $r_1 \leq F_{max}$ that would lead to hypercongestion.

Proof Proposition (2b) can be proven analogous to Proposition (2a), after replacing $\tau \cdot S_0$ ($\tau \cdot S_1$) by $\int_0^t S^+ du$ in the denominator of (5a) ((5b)), where S^+ gives the speed at instant τ of the drivers that arrived at τ=0. Since $S^+ > S^{\#}$ at τ_0, the same reasoning as underlying (5abc) can be applied.■

Proposition 3 Starting from a stationary state 0 $\{S_0, D_0, r_0=F_0=S_0 \cdot D_0 < F_{max}\}$ with hypercongestion, a change in the arrival rate to any r_1 will not lead the system to converge to a new stationary state 1 with hypercongestion $\{S_1, D_1, r_1=F_1=S_1 \cdot D_1\}$.

Proof If $r_1 > r_0 = F_0$, then $S_1 > S_0$ and $D_1 < D_0$ (see Figures 1 and 2). However, assuming that the last drivers that arrived at τ_0 either maintain S_0 or reduce their speed in response to the higher $d_{b,x}$ building up behind them, for all τ the average density over the first $d=\tau \cdot S_0$ meters is:

$$D_d \geq \frac{t \cdot r_1}{W \cdot t \cdot S_0} = \frac{r_1}{W \cdot S_0} > \frac{r_0}{W \cdot S_0} = D_0 > \frac{r_1}{W \cdot S_1} = D_1 \tag{6a}$$

Likewise, if $r_1 < r_0$, we find:

$$D_d = \frac{t \cdot r_1}{W \cdot t \cdot S_0} = \frac{r_1}{W \cdot S_0} < \frac{r_0}{W \cdot S_0} = D_0 < \frac{r_1}{W \cdot S_1} = D_1 \tag{6b}$$

Taking $\lim_{t\to 0} D_d$, which exceeds D_0 for (6a) and is smaller than D_0 in (6b), and subsequently using the result when reapplying (6a) and (6b) for drivers arriving later than τ_0, it is clear that we find for $r_1 > r_0$ consistently higher and increasing densities (and lower and decreasing speeds) near the entrance, whereas state 1 requires lower densities and higher speeds. A decrease in the arrival rate, in contrast, will lead to consistently lower and declining densities (and higher and increasing speeds) near the entrance, which are also inconsistent with the required values for stationary state 1. Therefore, the system diverges from the densities and speeds consistent with stationary state 1 after the arrival rate takes on the value $r_1=F_1$.■

According to Proposition 2, coming from any non-hypercongested initial situation, hypercongestion cannot be explained as long as $r \leq F_{max}$. The intuition is that the road space that becomes available per unit of time near the entrance is relatively large because of the relatively high initial speeds. Therefore, the number of new users needed per unit of time in order to build up a density consistent with hypercongestion is relatively large, partly because of the relatively high speed that new users will obtain themselves. Since at an initial speed of $S^{\#}$ one already needs the maximum inflow F_{max} in order to sustain the density $D^{\#}$ (compare Proposition 1), it is intuitively clear – and directly follows from (5ab) – that at initial speeds exceeding $S^{\#}$, one would need an inflow exceeding F_{max} in order to build up a density of $D^{\#}$ or larger. This inflow is impossible by definition.

Propositions 2 and 3 imply that as long as the arrival rate never exceeds F_{max}, the 'non-hypercongested' stationary states are the only feasible stationary states between which the system can move. Coming from any initial situation without hypercongestion, the system cannot reach a hypercongested stationary state at all. Even *if* the initial situation would be a hypercongested stationary state, a reduction in the arrival rate will then take the system to higher instead of lower speeds, while the opposite holds for increasing arrival rates. In the former case, the system may as a consequence leave the hyper-congested regimes 'for good'. Therefore, a hypercongested stationary state 1 will never result from a process where, starting from any other initial stationary state (hypercongested or not), the arrival rate takes on any value $r_1<F_{max}$. Hypercongested stationary states in contrast are 'razor's edge' dynamic equilibria, which can only result from an initial situation in which that particular stationary state's equilibrium conditions already apply. These conditions can never arise if the road was once opened empty (without-hypercongestion), and arrival rate's and inflows below F_{max} have always applied. Since an inflow exceeding F_{max} is inconsistent with the maximum capacity of the road over the first meters, we conclude that the upper segment of the AC-curve in Figure 2 is dynamically inconsistent according to the feasibility criterion. Note that the standard static model of course does not test for such questions related to the dynamic stability of equilibria.

3.3. The average and marginal cost curves for dynamically consistent equilibria

Because the upper segment of the AC-curve in Figure 2 is dynamically inconsistent according to the feasibility criterion, the only possible dynamically consistent stationary state equilibrium remaining in case the demand curve E applies is z. This, however, does not yet tell us what will happen in case E' applies. It seems that only a dynamically infeasible equilibrium x' remains. In particular for this question, the consideration of r instead of F as an output variable is helpful in determining the stationary state equilibrium. For this purpose, it should first be made explicit that the static equilibrium we want to find is the stationary state for a dynamic system where the stable demand curve E' applies from $\tau=0$ onwards. Furthermore, it is postulated that the initial situation at $\tau=0$ is an empty road.

The first drivers, starting at $\tau=0$, therefore expect to complete their trip at a speed not above S_{max} by definition, but which will at the same time not be lower than $S^{\#}$ during the trip, because the maximum inflow at and after $\tau=0$ is F_{max} (compare Proposition 2). The implied generalized costs provoke an initial arrival rate $r_0>F_{max}$. Now if the model does not somehow allow a queuing possibility for excess arrivals, this arrival rate cannot be accommodated, and the model breaks down. Note that this conclusion does not critically depend on the assumption that at $\tau=0$, the road is empty. It holds for any initial stationary state with $S_0\leq S^{\#}$ at the moment from which onwards the demand relation E'(r) applies. Also note that the system could not end up at x' in this case. With an arrival rate r=F(x'), a speed much higher than the speed associated with this configuration would arise by Proposition 2a.

The model can be 'saved' only if we do allow a queue to develop before the road's entrance. It is assumed that whenever there is a queue, the maximum possible inflow on the road f_{max} will apply. Usually, $f_{max}= F_{max}$; only in a stationary state with hypercongestion would it be smaller. Under this assumption, the queuing process takes the same form as is assumed in the bottleneck model (Vickrey, 1969; Arnott *et al.*, 1998). It is consistent with drivers

minimizing the time span between their predecessor's and their own entrance on the road, which is the inverse of the flow at the entrance.

Under the queuing assumption, with an arrival rate $r_0>F_{max}$, a queue immediately starts growing at a rate $q_0=r_0-F_{max}$ at $\tau=0$. As a consequence, the total travel time (t_t) for drivers starting their trips later than τ_0 will exceed the time spent on the road $t_r=L/S^{\#}$, because of the implied waiting time in the queue (t_q). Owing to the not perfectly inelastic demand function, the arrival rate therefore immediately starts declining at $\tau=0$. As long as r_τ exceeds F_{max}, the queue will keep on growing, and the arrival rate decreases. A stationary state is reached when $r_\tau=F_{max}$ and the queue has a constant length.[2] When queuing is allowed, the stationary state equilibrium with the demand curve E′ applying therefore involves $S^{\#}$, $D^{\#}$, and F_{max} on the road; an arrival rate $r=F_{max}$; and hence a stationary queue of constant length $Q>0$ which serves to keep away excessive demand through the implied waiting time costs. When queuing is not possible, the model has no equilibrium solution, because it then cannot handle $r_0>F_{max}$.

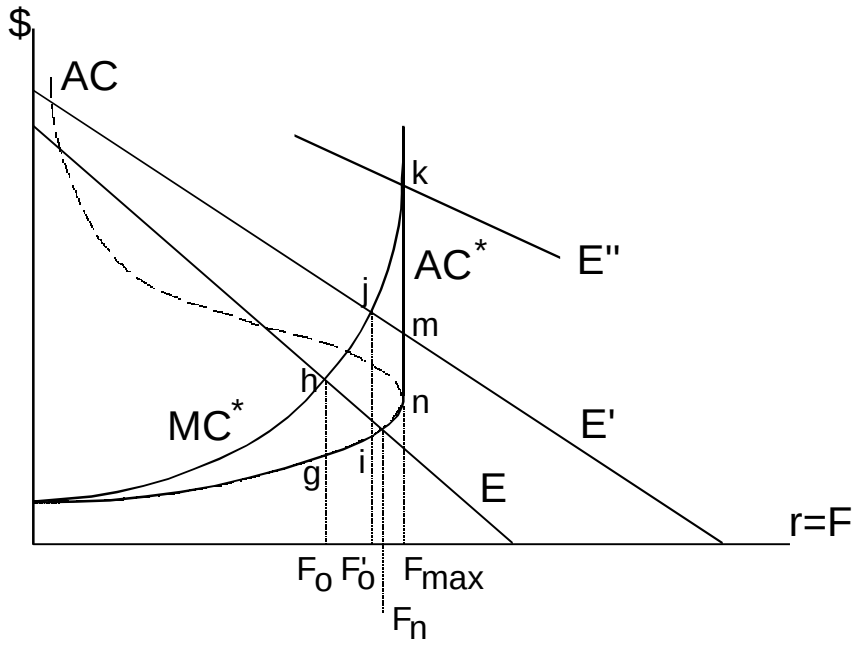

Figure 3. Dynamically consistent stationary state equilibria for the model with continuous congestion and a queuing possibility

The AC^*-curve in Figure 3 therefore shows the possible average cost levels for all dynamically consistent feasible stationary state equilibria when queuing is allowed. The output variable r is used. Only the lower segment of the standard AC-curve defined over F, representing dynamically consistent stationary states cost levels, is part of $AC^*(r)$. At F_{max}, however, $AC^*(r)$ rises vertically, showing that any marginal willingness to pay for making trips exceeding the travel costs at speed $S^{\#}$ will in stationary states simply be translated into queuing costs (m–n in

[2] Note that it is assumed that the demand relations defined over r are unrelated in time during the non-stationary phase: users do not consider rescheduling. Hence, the implication that drivers in the first, non-stationary part of the process are better off (have lower generalized costs) than those in the stationary part causes no problem. The case where rescheduling would occur during transitional phases due to average cost changing over time would merely add complexity while not changing the conclusion fundamentally.

case of E′). Mun (1994) obtains a similar cost curve with a serial two-link network model, based on kinematic wave theory.

A matching marginal social cost curve MC^* can now be derived. This curve lies above and is steeper than AC^* for stationary equilibria with $r=F<F_{max}$, and asymptotically approaches a vertical line at $r=F=F_{max}$. For the demand curve E, the non-intervention traffic flow is F_n; the optimum F_o, where marginal benefits are equal to marginal social costs, can be realized with a tax h–g. For E′, as described above, the non-intervention traffic flow is F_{max}, with queuing costs m–n, and the optimum $F'_o< F_{max}$ can be realized with a tax j–i. Note that at F'_o, no queuing occurs. Finally, in case a demand curve E″ applies, the optimal traffic flow is – approximately – equal to F_{max}. The optimal tax k–n then mainly serves to avoid queuing, but hardly affects the non-intervention stationary state traffic flow $r=F_{max}$. In the limit, the optimal tax k–n is equal to the queuing costs that apply in the non-intervention case, which is in line with one of the standard results in the bottleneck model (Vickrey, 1969; Arnott *et al.*, 1998).

4. The case of peak congestion

Although the case of continuous congestion discussed above offers a useful starting point for the economic modelling of road traffic congestion, congestion in reality is usually a peak event. Most models of congestion, therefore, implicitly or explicitly aim to describe peak congestion. This section considers the implications of the above analysis for static models of peak congestion. In doing so, it addresses the implicit assumptions, particularly on the nature of scheduling costs, that are necessary to render static models of peak congestion dynamically consistent. In this context, dynamic consistency is defined by the condition that during the peak period, (congested) speeds, densities, flows and travel costs should indeed be constant over time, as is implicitly assumed by a static representation. Because the assumptions necessary to render a static model of peak congestion dynamically consistent turn out to be rather unrealistic, also the implications of the analysis in Section 3 for dynamic models of peak congestion are discussed.

4.1. A static model of peak congestion

In contrast to dynamic models of peak congestion, where the duration of the peak is one of the endogenously determined variables, static models based on the fundamental diagram are often remarkably careless in the treatment of the duration of the peak. However, this duration is actually a crucial variable for the consistent modelling of a market for peak road usage. The reason is that the demand for peak travelling would naturally refer to the total number of trips accomplished during the peak, while the cost function for road usage would naturally be defined over flows or arrival rates (see Figure 3). Therefore, as argued in Section 2, for a consistent static economic model of peak congestion, it is necessary to take full account of the impact of the presumably endogenous duration of the peak, as an argument in either the flow-based demand function, or in the trip-based cost function.

Unfortunately, it seems hard to consistently endogenize the duration of the peak in a static model. This duration will in reality depend on users' trade-offs between time delays, scheduling costs, and – possibly – time-varying tolls. Once the desired arrival times and the scheduling costs are made explicit, however, one would normally end up with a dynamic model, where speeds and densities continuously vary over time. The only way out of this

dilemma for the static modelling of peak congestion is to assume that scheduling costs are constant over the peak. In equilibrium, no driver should be able to benefit from rescheduling: generalized user costs, including scheduling costs, should be constant over time. Therefore, only with constant scheduling costs will travel times also be constant. This, in turn, implies constant speeds and densities during the peak, which are required for a static model.

There are two assumptions that render constant scheduling costs over a well-defined duration of the peak. The first possibility, explored further below, involves the assumption that either T (the duration) or T_G (the 'grand duration') is somehow exogenously given. Below, T_G is assumed to be exogenous: all peak-travelling related activities have to take place within a given time span, within which scheduling costs are constant, and outside which they are prohibitively high.[3] Alternatively, a rather artificial endogenization of T in a static model could be accomplished by assuming that average scheduling costs are the same for all users during the peak, and increase with T. Because the implied assumption that all scheduling costs are a purely public bad is rather unrealistic, this possibility will be ignored here.

It is perhaps surprising that either one of these peculiar assumptions on the pattern of scheduling costs implicitly must underlie the static models of peak congestion that have been presented in the literature, and that assume constant speeds during the peak. An alternative assumption that scheduling costs do not exist simply will not work, since the duration of the peak could then be increased costlessly, and congestion would not occur at all.

Consider the case where the grand duration of the peak is exogenous and denoted T_G^*. T_G^* is then defined by an earliest departure time from home, τ_0, which is assumed to be the same as the arrival time at the entrance of the road, and a latest arrival time τ_1 at the exit of the road, which is where the workplace is. Therefore, $\tau_1-\tau_0=T_G^*$. The scheduling costs are constant – we assume zero – for those drivers departing after τ_0 and arriving at the road's exit before τ_1, no matter exactly when they travel; and prohibitively high for others. The implied step-wise scheduling cost function can be seen as an approximation for the case where morning peak commuters have no specific desired arrival time, but do not want to leave home before a given time, nor to arrive at work after a given time. In reality, one would then expect the scheduling costs to increase sharply, but not discretely. The comparable scheduling cost structure assumed by Ben-Akiva *et al.*, 1986, where scheduling costs are constant for some period and rise linearly outside that period, allows this. As already stated, however, the present discreteness assumption is a necessary requirement for using a static formulation. Finally, two additional assumptions should be made in order to avoid irregularities at the beginning and ending of the peak. The first assumption is that the very first driver(s) already choose the equilibrium speed. The second is the no-overtaking condition that a driver cannot arrive earlier or at the same time as someone who started the trip earlier, but will always arrive later.

Because the demand is defined in terms of total numbers of trips accomplished over the entire peak, also the cost functions now have to be defined in terms of N. To make this transformation, observe that the relation between F and N can be found by rewriting (2a) as:

$$N = F \cdot (T_G^* - t) \tag{7}$$

[3] Alternatively, if T were taken to be exogenously given, T_G would become endogenous. The reason for considering the case with exogenous T_G rather than exogenous T in the sequel is that it seems more realistic to consider prohibitively high scheduling costs for a departure from home *before* a given clock time, and for an arrival at work *after* a given clock time.

On the right-hand side of (7), both F and t are endogenous. By converting the speed-flow curve in Figure 1-IV into a travel time-flow (TTF-)curve, and using the fixed relation between speed and travel time given in (1b), a function F(t) can be constructed that depicts the equilibrium combinations of these two variables. This TTF-curve is shown in Figure 4-I. Writing F as F(t), the derivative of the right-hand side of (7) with respect to t can then be taken to investigate the equilibrium relation between t and N during the peak:

$$\frac{\partial N}{\partial t} = \frac{\partial F}{\partial t} \cdot (T_G^{\,*} - t) - F \qquad (8)$$

From (8), it follows that the maximum number of users that can travel over the grand duration of the peak, N_{max}, is found for a flow smaller than the maximum flow F_{max}, and hence a travel time below $t^{\#}$: $\partial N/\partial t=0$ requires $\partial F/\partial t>0$. F^* and t^* in Figure 4-I could give that particular combination of F and t consistent with the maximum number of users N_{max}. This implies that the average cost curve defined over N is backward-bending. Furthermore, it implies that the cost level for which this average cost curve has an infinite derivative with respect to N, at N_{max} in Figure 4-II, is lower than the minimum cost level consistent with the maximum flow F_{max} shown in Figure 3. To see this, note that t^* (for N_{max}) is smaller than $t^{\#}$ (for F_{max}).

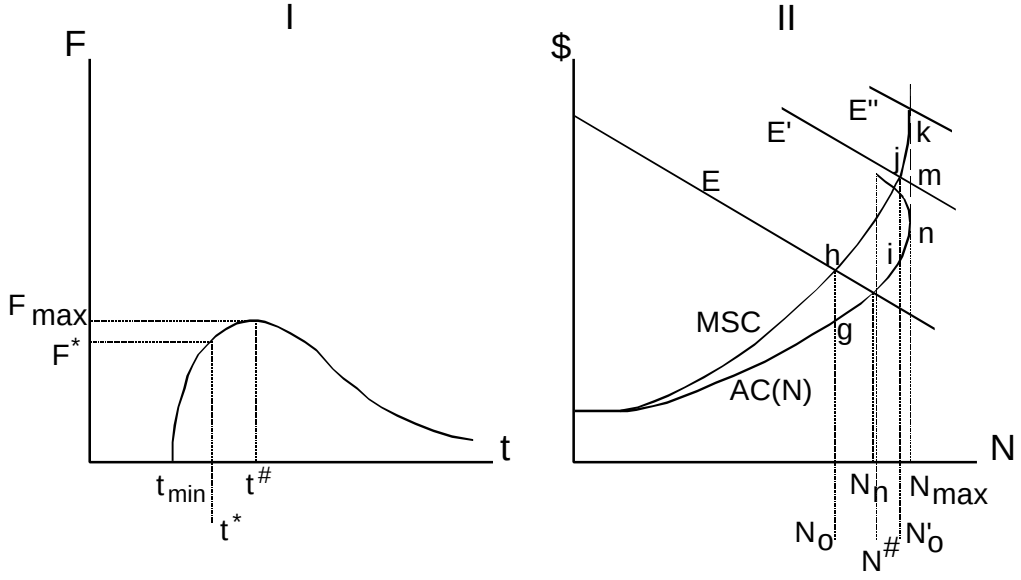

Figure 4. The travel-time flow relation (I) and the diagrammatic representation of the static model of peak congestion with exogenous grand duration (II)

In Figure 4-II, the backward-bending segment is drawn up to the (minimum) cost level consistent with F_{max} in Figure 3 (at $N^{\#}$). The reason is that configurations with higher cost levels necessarily involve queuing. The equilibrium principle that average costs should be constant over the peak is then violated: a queue will immediately be building up at τ_0. The waiting time t_q then cannot possibly be constant over the peak, whereas the time spent on the road $t_r^{\#}=L/S^{\#}$ will be constant. As a result, a consistent static equilibrium, with equal average user costs for all users, does not exist for such cases. It should therefore be emphasized that the backward-bending segment of the AC(N)-curve in Figure 4-II does not involve hypercongestion. The minimum speed for which this curve is defined is $S^{\#}$ (for $N^{\#}$), while it can be recalled from Figure 1-IV that hypercongestion sets in only at speeds below $S^{\#}$.

The lower segment of the AC-curve in Figure 4-II implies a marginal social cost curve MSC. These two curves can then be used to derive the non-intervention and optimal total numbers of users (N_n and N_o, requiring a toll h–g) in case the demand curve E applies, as well as the optimal total number of road users in case E′ applies (N'_o, requiring a toll j–i) or E″ applies (approximately N_{max}, requiring a toll k–n). For the non-intervention case when a demand curve such as E′ or E″ applies, the model simply has no static equilibrium solution with AC equal for all users, and equal to the marginal benefits.

Because of the backward-bending shape of the AC-curve defined over N, multiple intersections with the demand curve are in principle possible. Unless the demand curve is rather irregular, one would normally expect a maximum of two intersections, because the AC(N)-curve has no inflection point.[4] The question of which configuration will then finally come about as the unregulated market equilibrium depends on the size of the penalty for travelling outside T_G^*. As is argued and proven in the appendix of a longer version of this same paper, the lower the penalty, the more likely the more favourable equilibria are to arise.

It should also be noted that even if the demand E is flatter than AC(N), an intersection of the demand curve with the backward-bending segment of the AC(N)-curve can be a locally stable unregulated market equilibrium when considering quantity perturbations (provided, of course, the penalty is sufficiently high). This may seem odd in the light of the discussion in Section 3.1, Figure 2, where the configuration x was classified as unstable for quantity perturbations because beyond that point, drivers would keep on entering the road as average costs consistently fall short of marginal benefits. The reason that this argument does not apply here is that drivers not only have to decide *whether* to use the road, but also *when* to depart. An additional departure at any of the relevant instants available in the equilibrium would, given the departure times of the other drivers, imply marginally higher travel costs for those starting at that instant, because of the implied higher density. Therefore, in any of the possible equilibrium configurations depicted by AC(N), with constant travel costs during the peak, the 'marginal private costs' are increasing at each possible instant of arrival at the entrance, and therefore cannot coincide with the falling average social costs on the backward-bending part.

Indeed, considering quantity perturbations, the 'marginal private costs' are higher than what is suggested by AC(N) also for configurations on its upward sloping part. They would coincide with AC(N) only if all other drivers would respond optimally to perturbations, and new equilibria with constant average costs would result. A perturbation, however, is a disequilibrium concept by definition, and it would be inconsistent with the assumption of price-taking behaviour to assume that the perturbing driver would (rightly) expect all others to respond optimally to his own (unexpected) decision to make the additional trip. Hence, when studying perturbations, one should not ignore that the AC(N)-curve gives the average costs only for equilibrium departure patterns, where average costs are constant during the peak.

Although configurations on the lower segment of the AC(N)-curve described above may still be a reasonable approximation for real peak congestion, the model clearly becomes problematic when demand is relatively high. A static non-intervention equilibrium then even

[4] Note that by differentiating (8) once more, it follows that the inverse of AC(N), N(t), has a strictly negative second derivative: $d^2N/dt^2 = d^2F/dt^2 \cdot (T_G^* - t) - 2 \cdot dF/dt < 0$ ($dF/dt > 0$ and $d^2F/dt^2 < 0$ in the relevant region $t_{min} < t < t^{\#}$, as is shown also in Figure 4-I). I owe this observation to one of the anonymous referees.

may not exist. Still, it is striking that for this static model of peak congestion with exogenous grand duration, based on the fundamental diagram, two conclusions arise from the analysis in terms of numbers of road users that are usually drawn from the model of continuous congestion defined over flows, and that were rejected in Section 3. These conclusions are (1) that the average cost function is backward-bending, and (2) that the optimal toll may lead to an increase in road usage. The static model can properly describe all optima, as well as non-intervention outcomes with $S>S^{\#}$. Hypercongestion does not occur in optima, nor in (purely static) non-intervention outcomes, even if these involve equilibria on the backward-bending segment of the AC(N)-curve defined over total numbers of road users.

4.2. Implications for dynamic models of peak congestion

The above analysis actually sets the stage for a dynamic model of road traffic congestion based on the fundamental diagram, which would integrate elements from the bottleneck model with flow congestion models (see Rouwendal, 1990, for an earlier attempt along these lines). A full treatment would require a paper all of its own, but some discussion is warranted, in particular because an internally consistent static model of peak congestion appeared to require rather heroic assumptions on the structure of scheduling costs. Moreover, as demonstrated above, when demand is relatively high, the assumed stepwise scheduling cost function becomes the main driving force in the model, and it becomes increasingly worthwhile to relax this assumption and to consider the situation where the duration of the peak is endogenized.

As soon as scheduling costs (denoted k) are a continuous function of the difference between the actual arrival time and the jointly preferred arrival time, travel delays and average speeds should vary over the peak in order to obtain a non-intervention equilibrium in which total travel costs k+c are equal for all users (c gives the travel time costs). For flow congestion, this means that one has to find a formulation that can replace the relations in Figure 1 also for non-stationary processes. Henderson (1974, 1981) and Chu (1995) both make the convenient assumption of ‘zero group velocity’, in which case the speed experienced by a driver is a function of the arrival rate at the entrance of the road at the instant the trip is started (Henderson, 1974, 1981), or at the exit of the road at the instant the trip is ended (Chu, 1995). Drivers drive at constant speeds, so varying speeds can be observed along the road at every instant.[5] Such formulations have as an advantage over the bottleneck model that it is no longer assumed that up to the maximum inflow, no travel delays occur. However, neither model explicitly considers the above mentioned process that if arrival rates at the entrance exceed the maximum possible inflow, a queue will build up (note that, although bottleneck congestion is a limiting case of the congestion function considered by Chu (1995), in order to reach this limit the elasticity of travel delay has to approach infinity, so that the model then only has bottleneck congestion, and no flow congestion, as will be assumed below).

Although Chu (1995) has pointed out that overtaking could be a problem in Henderson’s (1974, 1981) formulation, for the present purpose it is convenient to consider the ‘Henderson-type’ of flow congestion, and to make the additional assumption, like Henderson

[5]Alternatively, Agnew (1973) assumes ‘infinite group velocity’, where at every instant speeds are constant along the entire road. This seems an even less realistic representation of the dynamics of road traffic congestion than zero group velocity (Henderson, 1974). Reality is likely to be somewhere in between these two extremes.

does, that overtaking is not possible.[6] Suppose that the speed s is a function of the arrival rate r at the instant τ the trip is started and is given by $s(r_\tau)$, that the inflow has a maximum value F_{max}, and the average travel costs function $c(r_\tau)$ is comparable to the AC(r) function depicted in Figure 3: as soon as $r_\tau > F_{max}$, a queue develops before the entrance of the road. Two types of non-intervention equilibria can then be considered, the first of which involves $r_\tau \leq F_{max}$ for all τ. This is consistent with the situation where the very first and very last drivers, who experience no congestion in dynamic models of traffic congestion and hence drive at free-flow travel costs c^* (Chu, 1995; Arnott *et al.*, 1998), face scheduling costs k_{max} for which:

$$k_{\max} + c^* \leq c_{\min}(F_{\max}) \tag{9}$$

where $c_{min}(F_{max})$ is the minimum travel time cost consistent with the maximum inflow (hence, with zero queuing costs). We know that those users arriving at the desired arrival time and facing zero scheduling costs can then not have experienced a queue, owing to the constancy of user costs over the peak. This produces the standard Henderson (1974, 1981) model, for which the following optimal time-varying tolls apply (Henderson, 1974, 1981; Chu, 1995):

$$toll_t = r_t \cdot \frac{\partial c_t}{\partial r_t} \tag{10}$$

If, however, (9) does not hold, we know that those users arriving at the desired arrival time and facing zero scheduling costs must have experienced a queue in the non-intervention situation. One could then at first glance expect a situation in which 'Henderson' tolls would apply for the first and last phases of the peak where $r_\tau \leq F_{max}$; and 'Vickrey'-bottleneck tolls would be necessary to avoid all queuing for the middle period where $r_\tau > F_{max}$ in the non-intervention outcome. Interestingly, however, the optimal Henderson toll in (10) prevents this to occur, since the optimal toll approaches infinity as r approaches F_{max}. Therefore, as in the standard bottleneck model, queuing will not occur in the optimum. In contrast to the pure bottleneck model however, with flow congestion, the entrance of the road will in the optimum always operate below the maximum capacity F_{max}. Furthermore, not all travel delays are eliminated, as optimal flow congestion is positive.

This concludes our brief excursion to dynamic models of road traffic congestion. It can be concluded that the static framework presented in Section 4.1 indeed can be extended to a dynamic model which combines elements of flow congestion with bottleneck congestion. This requires an alternative to the fundamental diagram, which is actually only valid for stationary states with constant speeds, flows and density. A first possibility was presented above, based on the notion of zero group velocity. A second possibility is a car-following model, as studied in Verhoef (1997). This type of integrated modelling, also proposed by Rouwendal (1990), certainly deserves further attention in future work.

5. Conclusion

This paper addressed some of the key questions that have dominated the debate on static models of road traffic congestion. A distinction was made between models that deal with

[6] Moreover, because both Chu (1995) and Henderson (1974, 1981) assume zero group velocity and constant speeds, it can be argued that with the non-overtaking restriction, the Henderson formulation should in pinciple replicate equilibria in Chu's model: the arrival rate a driver experiences at the entrance of the road is then for both models equal to the flow he experiences during the trip and the arrival rate he experiences at the road's exit.

'continuous demand', normally resulting in stationary state equilibria, and those that aim to describe 'peak demand'.

In the context of the former, it was demonstrated that the backward-bending section of the standard average cost curve is dynamically inconsistent, because the implied configurations are infeasible: they are dynamically unstable, and moreover, in order to get there, inflows on the road should have exceeded the maximum possible inflow at some point in the past. A *practical* consequence is that whenever 'hyper-congested' speeds are observed in reality, it is unlikely that the cause is to be found in 'flow congestion' on the road itself. Instead, the true reason for such speeds may often be a downstream bottleneck. Therefore, optimal pricing rules should then not primarily be based on the road's characteristics, but rather on the bottleneck's capacity. A *theoretical* consequence is that the standard backward-bending supply curve is flawed. Instead, it was argued that when replacing the endogenous output variable of 'traffic flow' by the arrival rate of new cars at the entrance of the road – two variables that should be equal to each other in stationary states only, but that do not presuppose this stationary state like the traditional output variable 'flow' does – a non-backward-bending supply curve can be found, which coincides with the standard curve only for its lower segment, but rises vertically at the road's maximum capacity.

For static models of peak demand, it was argued that for such models to be dynamically consistent, rather heroic assumptions on the pattern of scheduling costs have to be made. Interestingly, once these assumptions are made, a backward-bending cost curve defined over numbers of road users was derived from the non-backward-bending cost curve defined over arrival rates that was found earlier for the case of continuous demand.

Finally, because the assumptions necessary to render a static model of peak congestion dynamically consistent turn out to be rather unrealistic, also the implications of the analysis in Section 3 for dynamic models of peak congestion were discussed. A dynamic model, which combines elements of flow congestion with bottleneck congestion, was outlined. Such integrated modelling deserves further attention in future research.

References

Agnew, C.E. (1973) "The dynamic control of congestion – prone systems through pricing". Report No. 6, Stanford University Center for Interdisciplinary Research.

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 (1998) "Recent developments in the bottleneck model". In: K.J. Button and E.T. Verhoef (1998) *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* Edward Elgar, Cheltenham (forthcoming).

Ben-Akiva, M., A. de Palma and P. Kanaroglou (1986) "Dynamic model of peak period traffic congestion with elastic arrival rates" *Transportation Science* **20** 164-181.

Braid, R.M. (1989) "Uniform versus peak-load pricing of a bottleneck with elastic demand" *Journal of Urban Economics* **26** 320-327.

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

Chu, X. (1995) "Endogenous trip scheduling: the Henderson approach reformulated and compared with the Vickrey approach" *Journal of Urban Economics* **37** 324-343.

Chu, X. and K.A. Small (1996) "Hypercongestion" Paper prepared for the meeting of the American Real Estate and Urban Economics Association, New Orleans, Jan. 1997.

De Meza, D. and J.R. Gould (1987) "Free access versus private property in a resource: income distributions compared" *Journal of Political Economy* **95** (6) 1317-1325.

Else, P.K. (1981) "A reformulation of the theory of optimal congestion taxes" *Journal of Transport Economics and Policy* **15** 217-232.

Else, P.K. (1982) "A reformulation of the theory of optimal congestion taxes: a rejoinder" *Journal of Transport Economics and Policy* **16** 299-304.

Evans, Alan W. (1992) "Road congestion: the diagrammatic analysis" *Journal of Political Economy* **100** (1) 211-217.

Evans, Andrew W. (1992) "Road congestion pricing: when is it a good policy?" *Journal of Transport Economics and Policy* **26** 213-243.

Evans, Andrew W. (1993) "Road congestion pricing: when is it a good policy?: a rejoinder" *Journal of Transport Economics and Policy* **27** 99-105.

Henderson J.V. (1974) "Road congestion: a reconsideration of pricing theory" *Journal of Urban Economics* **1** 346-365.

Henderson J.V. (1981) "The economics of staggered work hours" *Journal of Urban Economics* **9** 349-364.

Hills, P. (1993) "Road congestion pricing: when is it a good policy?: a comment" *Journal of Transport Economics and Policy* **27** 91-99.

Johansson, O. (1997) "Optimal road-pricing: simultaneous treatment of time losses, increased fuel consumption, and emissions" *Transportation Research* **2D** (2) 77-87.

Knight, F.H. (1924) "Some fallacies in the interpretation of social cost" *Quarterly Journal of Economics* **38** 582-606.

Lave, C. (1994) "The demand curve under road pricing and the problem of political feasibility" *Transportation Research* **28A** (2) 83-91.

Lave, C. (1995) "The demand curve under road pricing and the problem of political feasibility: author's reply" *Transportation Research* **29A** (6) 464-465.

Mun, S.-I. (1994) Traffic jams and the congestion toll" *Transportation Research* **28B** (5) 365-375.

Nash, C.A. (1982) "A reformulation of the theory of optimal congestion taxes: a comment" *Journal of Transport Economics and Policy* **26** 295-299.

Newell, G.F. (1988) "Traffic flow for the morning commute" *Transportation Science* **22** 47-58.

Pigou, A.C. (1920) *Wealth and Welfare*. Macmillan, London.

Rouwendal, J. (1990) "An integrated model of traffic congestion". Manuscript, Wageningen University.

Small, K.A. and J.A. Gomez-Ibañez (1998) "Road pricing for congestion management: the transition from theory to policy". In: K.J. Button and E.T. Verhoef (1998) *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* Edward Elgar, Cheltenham (forthcoming).

Verhoef, E.T. (1995) "The demand curve under road pricing and the problem of political feasibility: a comment" *Transportation Research* **29A** (6) 459-464.

Verhoef, E.T. (1998) "An integrated model of road traffic congestion based on simple car-following theory". Discussion Paper TI 98-030/3, Tinbergen Institute, Amsterdam-Rotterdam.

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

Verhoef, E.T., P. Nijkamp and P. Rietveld (1995b) "The economics of regulatory parking policies" *Transportation Research* **29A** (2) 141-156.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1996a) "Second-best congestion pricing: the case of an untolled alternative" *Journal of Urban Economics* **40** (3) 279-302.

Verhoef, E.T., R.H.M. Emmerink, P. Nijkamp and P. Rietveld (1996b) "Information provision, flat- and fine congestion tolling and the efficiency of road usage" *Regional Science and Urban Economics* **26** 505-529.

Vickrey, W.S. (1969) "Congestion theory and transport investment" *American Economic Review* **59** (Papers and Proceedings) 251-260.

Walters, A.A. (1961) "The theory and measurement of private and social cost of highway congestion" *Econometrica* **29** (4) 676-697.

Wardrop, J. (1952) "Some theoretical aspects of road traffic research" *Proceedings of the Institute of Civil Engineers* **1** (2) 325-378.