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

Erik 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

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

Abstract

This paper studies some of the properties and fundamentals of static models of road traffic congestion that have triggered much debate in the literature. The first part of the paper focuses in particular on the difficulties arising with the backward-bending cost curve in the context of 'continuous congestion'. The relevance of the backward-bending segment of the cost curve for the static analysis of congestion is questioned by demonstrating that 'equilibria' on this segment produce upwards shifts of the cost curve itself, and can therefore not be of a stationary nature. Next, the implications for static models of 'peak congestion' are considered. In doing so, the paper addresses 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 and Kenneth Small for very useful and inspiring comments on an earlier version of this paper. Needless to say, the usual disclaimer applies.

1. Introduction

Notwithstanding the long history of the economists' and engineers' study of road traffic congestion and congestion pricing (see, for instance, Pigou, 1920; Knight, 1924; Wardrop, 1952; Walters, 1961; and Vickrey, 1969), it is fair to admit that 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 (Foster, 1974, 1975, versus Richardson, 1974, 1975; 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, 1997), 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 discusses some of the key issues that have dominated, explicitly or implicitly, the above mentioned debate in the literature, and that may cause analysts to arrive at rather different representations of congestion, and hence at different conclusions.

Within the class of economic models of road traffic congestion, two main streams can be distinguished, namely static and dynamic models. This paper mainly focuses on static models, especially those based on the so-called 'fundamental diagram' of traffic congestion, depicting the relation between density and speed. Although static models have obvious limitations in the analysis of traffic congestion, they are still often used for both research and educational purposes, and it is therefore worthwhile to consider the properties and implicit fundamentals underlying these approaches in some further detail.

In static models of congestion, the time dimension is not explicitly considered. However, as will become clear, 'static' does not mean that time as such plays no role in these models. It certainly does, often implicitly, and even causes part of the disagreement. 'Static' only means that these models do not explicitly study (or allow for) changes over time. A main source of disagreement in such static modelling of road traffic congestion concerns the choice of the output variable in the definition of demand and cost functions. Two main stances can be distinguished: '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 on the other, such as numbers of trips or densities (Alan Evans, 1992; Hills, 1993; Verhoef *et al.*, 1995ab, 1996ab). A second main issue concerns the question of whether the analysis applies to 'peak demand' or 'continuous demand'. This distinction will turn out to have important implications for the static modelling of congestion. Unfortunately, however, the question of which of these two types of demand is considered is usually not treated explicitly in such analyses. These two issues will turn out to be closely related, and both are dealt with in this paper.

Dynamic models of road traffic congestion typically focus on peak congestion, and explicitly describe equilibrium patterns of variables such as speeds, densities, and arrival rates over time during the peak. Two types of such dynamic approaches can be distinguished. The first one is the 'bottleneck approach', originally developed by Vickrey

(1969), and later on refined and extended in various directions by Arnott *et al.* (1993, 1997), Braid (1989, 1996), and others (see Arnott *et al.*, 1997, for a comprehensive overview). The second approach, originally proposed by Henderson (1974, 1981) uses ‘flow congestion’. In contrast to the bottleneck model, in this model travel delays are not completely eliminated in the social optimum (see Chu, 1995). In both approaches, the distribution of travel delays and scheduling costs over the peak and the duration of the peak in the unregulated equilibrium and the social optimum are determined endogenously, and both models have in common that the optimal toll is time-dependent, reaching its maximum for the drivers arriving at the desired arrival time.

As mentioned, the present paper is mainly concerned with static models of road traffic congestion, for which the disagreements among analysts have been particularly fierce and persistent. Nevertheless, in the process of investigating the underlying, often implicit assumptions that have caused these analysts to offer different and often opposing views, some elements that would perhaps sooner be associated with dynamic modelling, and therefore tend to be ignored in static analyses, will play an important role. Apart from the aleady mentioned distinction between peak demand and continuous demand, the question of whether a static equilibrium represents a dynamically stable configuration – a ‘stationary state’ – will be considered explicitly, and will be taken as a prerequisite for a static equilibrium to be consistent and meaningful. Furthermore, in the static model of peak demand, the scheduling cost structure necessary to render a static model applicable will be made explicit.

The paper is organized as follows. Section 2 discusses some main features of static approaches to the modelling 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 that have dominated much of the debate in the literature. The relevance of the backward-bending segment of the cost curve for the static analysis of congestion is questioned by demonstrating that ‘equilibria’ on this segment produce upwards shifts of the cost curve itself, and can therefore not be of a stationary nature. Section 4 studies the implications of this finding for static models of peak congestion. In doing so, it also addresses the question of 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. The interpretation of static models of road traffic congestion

As with most phenomena studied by economic science, road traffic congestion in reality is a complicated, inherently dynamic and spatially differentiated process, caused by the interplay between (and at the same time affecting the decisions of) a large variety of heterogeneous agents. The analyst studying congestion and congestion pricing is therefore soon confronted with the strategic choice of whether to opt for either an ‘as realistic as possible’ modelling approach, in which analytical solutions are difficult to obtain (see, for instance, the hydro-dynamic fluid model of Newell, 1988), or for an admittedly much simpler representation of reality, which, however, permits analytical solutions and, hopefully, 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.[1] Although the static model of road traffic congestion could be criticized for its inherent inability to deal with the dynamic features of congestion, it has a number of advantages, most of which are related to its apparent analytical simplicity and tractability. For the same reason, it probably more easily lends itself for further analytical extensions – other than related to dynamics – than does a dynamic model. Such extensions may include larger networks, or the study of the interactions between transport and the overall (spatio-)economic system. It is, given our analytical capabilities, probably too soon to throw the static model overboard, despite its limitations. Moreover, exactly because static models of road traffic congestion based on the fundamental diagram are still often used for both research and educational purposes, it is worthwhile to consider the properties and implicit fundamentals underlying static approaches in some further detail.

Notwithstanding the absence of an explicit time dimension in static models of road traffic congestion, time does play a role in these models. No matter how abstract it may seem in a ‘time-less’ approach, it is usually postulated that average generalized travel costs increase with usage[2], where this increase is due to decreased speeds and resulting time losses. Moreover, as in any static economic model of a market, the demand (or marginal benefit) and supply (or marginal cost) relations can only be specified after it is decided to which time period they apply – and these should of course be consistent for the two. For instance, in performing a standard diagrammatic analysis of the static equilibrium of a market for, let’s say, bread, under the admittedly simplifying assumptions that no fixed costs and seasonal effects in the production of bread are involved, the consideration of two periods instead of one would imply an outward rotation of both the supply curve and the demand curve by a factor 2, yielding a double equilibrium quantity, but the same equilibrium price. Especially when marginal costs are increasing, the confrontation of the demand curve pertaining to one period with the supply curve pertaining to two periods would produce a meaningless, flawed result. No matter how obvious this observation may seem, comparable flaws may plague static analyses of road traffic congestion, in particular when directed to the case of peak congestion (see below).

[1] Of course, one could envisage a formulation in which the peak is split up in a number of periods, with different demand functions, each yielding a different optimal toll. However, these periods, especially adjacent ones, will in reality of course not be independent. Considering such interdependencies would turn the model into a dynamic one; either based on discrete or continuous time. Strictly speaking then, a static formulation yields one single value for an optimal toll.

[2] Before turning to the discussion of which output measure to choose, the neutral phrasing ‘usage’ will be used.

A number of variables are normally used to characterize the non-intervention and optimal equilibria in the analysis of road traffic congestion. In defining these, it is in the first place important to consider one single, well-defined market. In particular, the product ‘trip’ should be homogeneous, which can most easily be assured by considering a one-link network, 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 are especially relevant in dynamic models, do not affect this homogeneity condition as long as such variations are reflected in differences in ‘generalized user costs’. In addition, apart from having a possibly different maximum willingness to pay for making a trip, giving rise to an elastic demand function for using this single link, the potential road users should be identical in all other respects. In particular, they share the same value of time, and they all contribute to congestion in the same way; for instance, they have the same vehicles and driving styles. These assumptions are necessary to postulate continuous cost functions, and should be relaxed only if one wishes to study the impacts of heterogeneity among users explicitly.

Furthermore, in a static equilibrium, all relevant variables will 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 usually vary over the peak (see Arnott *et al.*, 1997; and Chu, 1995), but it is simply a property of static models. It is, however, important to emphasize here the difference between the case of ‘peak demand’ (also referred to as models of ‘peak congestion’ in the sequel), where this divergence between the properties of static and dynamic models is evident, and the case of ‘continuous demand’ (or ‘continuous congestion’). This latter case would normally result in an everlasting ‘stationary state’ situation, where a road is continuously used at a constant intensity, so that the assumptions of single equilibrium values for speeds, flows, and densities could be less problematic. ‘Peak demand’, in contrast, refers to the case where a limited number of potential users consider using the road during the same period, the duration of which could be endogenized, where the total (equilibrium) quantity of actual users will depend on the (equilibrium) level of user costs. It will become clear in the sequel that this distinction is important for an unambiguous interpretation of static models of congestion. Much of the debate in the literature is partly caused by opposing authors apparently having different types of demand in mind when making their points. Therefore, in order to illustrate the difference between the two types of demand unambiguously: the intersection of the inverse demand curve with the horizontal axis gives (1) in case of peak demand: the total number of road users in case user costs were zero, with zero usage and an empty road before and after they have travelled; and (2) in case of continuous demand: the constant and everlasting number of users completing their trip per unit of time in case user costs were zero. Clearly, one could consider mixed cases, with a certain peak demand interfering with some continuous demand for the use of the same road, but the main point here is only to distinguish between these two basic types of demand.

Although one could rightly argue that road traffic congestion is usually due to peak demand, also continuous demand will be considered in this paper. In the first place, it is the model that is implicitly assumed to apply in many static models of congestion; often those models based on the fundamental diagram that use traffic flow as the output measure.

Secondly, it is closer to standard static economic models of markets, and thus shows some important differences between such standard models and markets for congested road usage, apart from the additional complications resulting from the limited duration of the peak.

Let us now turn to the various relevant variables. Consider first the set of possible measures of 'usage'. The first of these is *total road usage*, over the entire period considered, without specifying the duration of this period. 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. Total usage can be measured in numbers of road users completing the desired trip during the peak, and will be denoted N. 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), which refers to the number of users per unit of road space, where the total road space 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*.

Speed (S) is defined as distance travelled per unit of time, and is, after flow, the second variable in which a time component is explicitly present. Speed has, through the *value of time* (v), an important impact on the *generalized user costs*, because decreased speeds imply increased travel times. These generalized costs, which make up *average social costs* (AC), therewith include both monetized costs (for instances, expenses on fuel) and time costs. As in most models, only time costs will be considered in what follows.

Apart from speeds and flows, there are three more time-related variables relevant for the static analysis of congestion, the first of which is the *duration of the peak* (T). This variable is usually ignored in static analyses of congestion, even if directed to the case of peak demand. However, it should be considered carefully in these cases, in order to avoid the type of fallacies described above for the case of the bread market. In particular, in case of a 'trip-based' demand function, giving the demand for using the road during the peak in terms of total number of trips as a function of the (equilibrium level of) generalized user costs, the specification of the cost function deserves careful attention. The reason is that the (average and marginal social) costs of having a number of users completing a trip will depend on the presumably endogenous duration of the peak: the longer the duration, the lower the costs. Likewise, in case of a 'flow-based cost function', giving the generalized costs as a function of passages per unit of time, the specification of the demand function should particularly be checked for consistency. For a given demand function defined in terms of the total number of trips completed during the peak, the position of the demand function defined over flows then depends on the endogenous duration of the peak itself (compare (2a) below). In what follows, two measures for the duration of the peak will be used. The duration T, without further qualification, will be used for the period between the first and last driver in the peak passing a certain point of the road; and 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 meaning of which is evident. In contrast to T and T_G, t is relevant both for peak demand and continuous demand. The grand duration of the peak, in a purely static model with peak demand, is equal to T_G=T+t.

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

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

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

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

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)–(1b) 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 (2c) only holds for places along the road at instants that it is actually used, and that during the first (and last) t time units, during which 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.

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).

An important question is whether the above static representation of peak congestion, in equations (2a)–(2c), can be internally consistent. As a matter of fact, some concentration of desired arrival times and some existence of scheduling costs have to be assumed in order to maintain that congestion will occur at all during the peak. Otherwise, users could always postpone their trips costlessly, and drive at free-flow costs. T and T_G could then be increased costlessly. Indeed, the specification of desired arrival times and scheduling costs is necessary to endogenize T and T_G, because the duration of the peak can be expected to depend on trade-offs between travel time delays, scheduling costs, and (time-varying) tolls in case these are levied. However, once these desired arrival times and scheduling costs are made explicit, one would normally end up with a dynamic model. Therefore, it seems difficult to consistently endogenize the duration of the peak in a static model of peak congestion. This question will be addressed in Section 4. First, however, the case of continuous demand will be considered.

3. The case of continuous congestion

The case of continuous congestion, where the demand for road usage is constant over time for an infinite time period, is probably not the most realistic representation of road traffic congestion in reality, but it is nevertheless the situation that is often, implicitly, assumed to apply in static models of road congestion and congestion pricing (see, for instance, Johansson, 1997). Indeed, this configuration can be seen as offering a basic 'bench-mark' model for studying the economics of congestion, the insights of which can be helpful in interpreting more realistic and complicated static or dynamic models. Moreover, even this probably most simple representation of road traffic congestion has triggered a remarkable level of disagreement, which justifies further consideration of the model. Finally, the model allows one to study some first differences between 'standard' static economic market models, and the model for the market of congested road usage, postponing the additional complications caused by the endogenous duration of the peak for the time being.

3.1. The standard analysis

Continuing the discussion in Section 2, it can first be observed that there are various interdependent variables: n, D, F, S, and t; while only L and W are constants (N is not relevant for continuous congestion). Two pairs of variables are unambiguously related through these constants, namely S and t by (1b), and D and n by (1c). It is usually assumed that speeds have some natural upper limit, S^*, which is the free-flow speed that a user would have when driving in a congestion-free situation. The equilibrium flow F can be written only as the ratio of two endogenous variables, namely n and t according to (1a); or as the product of two other endogenous variables, D and S, multiplied by the constant W according to (3). Therefore, in order to find the relation between equilibrium values of F and one of the other variables, one has to take account of the full equilibrium consequences of changes in that particular variable.

The relationship that is needed for that purpose is the one between speed and density. If speed were independent of density, flow would simply be linear in D. This would be the case if the road had an infinite capacity. Usually, however, speeds decrease with increasing densities, as is illustrated by the density-speed relation (DS-curve) in the first quadrant in Figure 1. This 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 secondly that there is some maximum density D_{max} at which speed falls to zero. Knowing that F is proportional to the product of D and S, it is now immediately clear that F will obtain a maximum value for some combination of speed and density, which are denoted $S^\#$ and $D^\#$. This gives rise to the familiar 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, via the value of time v, the speed-flow curve in Figure 1-IV can be combined with the relation between speed and travel times in (1b) to obtain the backward-bending average social cost function (AC) defined over flows, 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 ‘hyper-congestion’, corresponds with the lower section of the SF-curve; and, as speeds go to zero, generalized user costs go to infinity. 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 low and the speed is high, and a high one, where the opposite holds. It is especially this backward-bending cost curve that has lead to heated debate in the literature, in particular because the confrontation of this curve with a standard downward sloping demand curve may produce puzzling results.

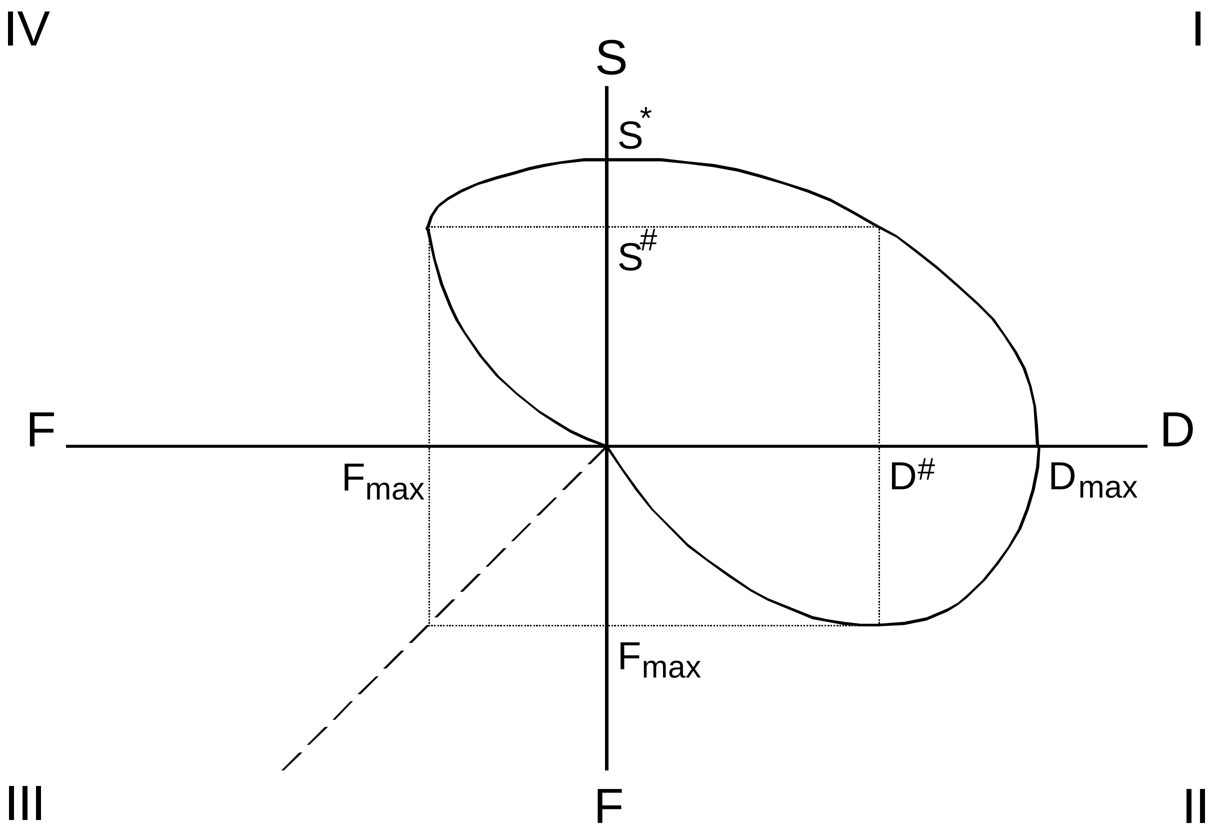

Figure A. The density-speed curve (I), the speed-flow curve (IV) and the density-flow curve (II)

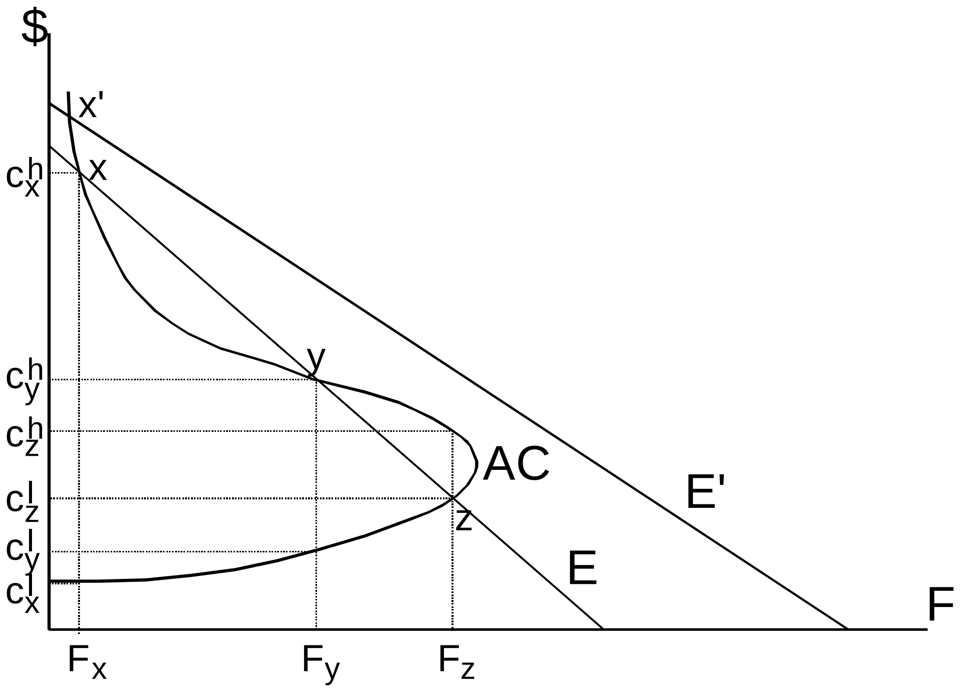

Figure B. *The backward-bending average cost curve (AC) and two demand curves (E and E) defined over traffic flow (F)*

Else (1981), Alan Evans (1992) and Hills (1993) have challenged the representation in Figure 2, in particular by questioning the appropriateness of flow as an output variable. Before addressing their arguments, we first proceed by investigating the non-intervention market equilibria, in absence of tolling, that are suggested by Figure 2. And, before doing that, it should be emphasized that the translation of the SF-curve into the AC-curve through the relation between speed and travel times by definition assumes that speeds, densities and flows are constant over time and along the road: the road should be in a stationary equilibrium. If speeds and densities would vary during the trip, one would first have to determine the total travel time implied by these varying speeds, before being able to determine generalized user costs. AC as a function of F would then not be very meaningful, as F itself would vary during the trip. So, the very nature of Figure 2, giving the average costs of making the entire trip as a function of a variable that describes the situation at a particular instant at a particular place along the road, presupposes this variable to be constant over time and place. Moreover, it is unlikely that drivers at a particular spot and instant will have, and maintain, the combination of speed and density implied by the fundamental diagram in Figure 1-I if the speed of the drivers just in front of them is either lower or higher than their own speed. The postulation of one single DS-relation therefore also presupposes speed to be constant.

In Figure 2, two demand curves, denoted E and E′, are included. These curves 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 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.

Let us first consider the demand curve E, which is drawn so as to produce the possible types of intersection with the cost curve that can be distinguished. In contrast to standard graphical representations of market equilibria, in which unique equilibria are usually found, this curve seems to produce three possible equilibrium levels of flow, namely the three intersections of E and AC labelled x, y, and z, where the marginal benefits E (being the benefits of making the trip enjoyed by the marginal driver) are equal to the generalized costs AC (the 'price' considered by the road users). Furthermore, each of these flows could in principle be obtained at two cost levels, indicated with superscripts l (low) and h (high). Considering, for the sake of completeness, the resulting six configurations as candidates for an 'equilibrium', it can first be noted that z is nearest to standard market equilibria where supply curves slope upwards and demand curves slope downwards. It is also easily seen that, if the flow F_z would occur at the high generalized cost level c_z^h, the market could not be in equilibrium, as average costs then exceed marginal benefits. Next, consider point x more closely. It is, after a moment of reflection, evident that (F_x, c_x^h) could not be a stable equilibrium: a marginal increase of the flow makes marginal benefits exceed the average user cost, so the flow will continue to increase. To the left of x, it is evident that the flow would turn to 0, as benefits consistently fall short of average costs.[3] Also (F_x, c_x^l) gives no equilibrium, because marginal benefits exceed average costs. This eliminates two more candidates for equilibrium configurations. Finally, we have point y, which does seem to present a stable equilibrium at c_y^h. Clearly, the same flow could be obtained at a lower cost c_y^l. It is usually argued 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 obtained at the minimum cost. Apart from that, the toll should of course bridge the gap between AC and marginal social costs (MSC), but MSC is ignored for the time being in order to keep the diagram decipherable. Since (F_y, c_y^l) gives no equilibrium, because marginal benefits exceed generalized costs, we are left with two possible candidates for stable equilibria: y and z, the latter of which is clearly superior in producing a greater flow against lower costs. Nevertheless, also y appears to be a stable equilibrium, and the model so far cannot predict which of the two equilibria will come about.[4]

[3] Else's (1982, p. 301) claim that such a point in fact could be a stable equilibrium, because movements up the backward-sloping section imply increasing numbers of users on the road, and that therefore at x "marginal costs are increasing faster than marginal benefits" seems flawed. The statement either proves that, to the left of x, flow will indeed fall to zero (as indicated in the main text), or it ignores the fact that the demand curve he draws, in his Figure 3, should also be defined over flows; not over densities.

[4] The foregoing discussion of stability of equilibria considers small perturbations from an equilibrium in terms of *flows*, and uses the resulting changes in marginal benefits and average costs to determine the stability. It is interesting to note that if *price perturbations* are considered, in line with the Walrasian tâtonnement process, one might conclude that x is locally stable and y locally unstable. For a slightly higher price, an excess supply (demand) is then found for x (y), leading to a downward (upward) price adjustment by the auctioneer, and hence to a move back to (further away from) the initial equilibrium. Note that z is stable according to both approaches. However, it is questionable whether the auctioneer-approach offers a very realistic representation of the market under consideration. Moreover, should this approach be maintained, it can be argued that the supplier(s) of road space would offer the maximum flow at the prices

3.2. Reconsidering the standard analysis

The representation in Figure 2 has been challenged by a number of authors. One particular problem has received relatively little attention, and that is the situation where the demand curve E′ would apply. In that case, only the unstable equilibrium x′ remains. Beyond that flow, flows will continue to increase as marginal benefits consistently exceed the user costs. The 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. 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.

A first point of criticism on the above representation concerns the choice of flow as an output variable. Hills (1993), for instance, suggests that it should be the number of trips accomplished that is the relevant output variable, and Alan Evans (1992) suggests densities. After the discussion in Section 2, it will be clear that the total number of trips completed certainly makes sense in the case of peak congestion; however, in the case of continuous congestion, the case considered here, it seems that the number of trips completed *per unit of time* – flow, in a stationary state – should still be the relevant output measure. The number of trips as such is a meaningless concept in a stationary state (it is either equal to zero or increases with the time period considered), whereas the use of density as an output variable would imply that it is 'being on the road' that people wish, instead of 'completing a trip'. An increasing density can be seen as the result of more trips being undertaking per unit of time, which is the 'product' that potential road users actually want (only in case of touristic road usage, involving sightseeing from the car, could one argue differently). Therefore, it seems that one should maintain flow as the relevant output variable in the case of continuous congestion. This measure, which is equal to completed trips per unit of time in a stationary state, acknowledges that it is completed trips that users are interested in. The normalization with respect to the time dimension is perfectly consistent with common practice for the specification of demand and cost curves in standard static economic market models (Else, 1982 (p.300); see also the example of the bread market in Section 2).

A second point of criticism is that traffic flow is in fact an "endogenous variable, resulting from the characteristics of the road and interactions among road users" (Alan Evans, 1992, p. 212). Alan Evans then proceeds to claim that, for that reason, the demand curve should relate to density, a modelling principle that was rejected above. However, the point on endogeneity is valid. In particular, the backward-bending shape of the cost curve defined over flows results from the fact that an increase in flow arises from jointly occurring, interrelated changes in density, directly affected by individuals' decisions to use the road, and speed, which is only indirectly affected.

Alan Evans's (1992) and Hills's (1993) viewpoints are actually quite near those used below when offering a new interpretation of what happens on the backward-bending section of the cost curve. Two observations are made before doing so. The first is that, in

labelled with superscript h, causing excess supply for all price offers corresponding with the backward-bending segment of AC. Finally, since it will turn out in the next section that the backward-bending segment yields non-stationary equilibria anyway, the above discussion of stability of x and y under the incorrect assumption that the AC-curve itself is stable at these points is only of limited relevance.

using the backward-bending average cost curve, one should verify whether ‘equilibria’ on the upper segment satisfy one of the basic requirements for a consistent static equilibrium, namely that the functions considered should be stable. It will be claimed below that the upper segment of the AC-curve in Figure 2 does not satisfy this requirement: the AC-curve itself will shift upwards, due to an ‘equilibrium’ on this segment. The second observation is that, in a stationary state, it is necessary that the traffic flow be equal to the inflow, and to the arrival rate of new users at the entrance. So far, authors have implicitly assumed that this condition is always met, by ignoring what is going on at the entrance of the road. However, it will turn out that, for outcomes on the upper section of the AC-curve, this condition is violated. This also suggests that this segment of the AC-curve cannot be consistent with stationary equilibria. These two observations, of course, are closely related.

Let us start by considering the arrival rate of users at the entrance of the road, being the number of users ‘popping up’ at this entrance per unit of time. This variable will be denoted r. In a stationary equilibrium, we need r=F; and this should also be equal to the outflow: the number of users leaving the road at the exit per unit of time. In contrast to F, however, r itself is not defined as the product, or as the ratio, of the equilibrium values of two other endogenous variables; although in a stationary equilibrium, r will be equal to F. The density at the entrance of the road will be denoted d, and can for stationary equilibria, with a constant r, be written as d(r). The more users appear per unit of time, the higher the density at the entrance. It is tempting to define d(r) as r/W, but this ignores that density occurs not only in the ‘width dimension’, but also in the ‘length dimension’ of the road. Unlike the DF-function in Figure 1-II, however, it is evident that the relation between r and d for stationary states cannot be decreasing: ∂d/∂r≥0. In a stationary equilibrium, we will find d(r)=D if the entrance has the same physical characteristics as the rest of the link, and newly arriving users will, at the instant of their appearance, obtain a speed s(d) according to the DS-curve in Figure 2-I. In a stationary equilibrium, s(d)=S. Finally, we define f as the inflow on the road, so f gives the number of users that successfully pass the entrance per unit of time. In a stationary equilibrium, we need f=r=F. The relation between the arrival rate r and the inflow f in a stationary equilibrium (without queuing before the entrance; see below) can be written as:

$$f = s(d(r)) \cdot d(r) \cdot W \tag{4}$$

The s(d(r)) term in (4) is consistent with ‘Henderson congestion’, where the speed that drivers obtain is a function of the arrival rate at the instant of arriving at the entrance of the road; see Henderson (1974, 1981), and also Section 4.2 below. An important difference with Henderson’s formulation is that the role of density is made explicit here. Next, a necessary stationary state equilibrium condition is:

$$f = r \tag{5}$$

This implies that the partial derivative of the right-hand side of (4) with respect to r should be equal to 1 in a stationary state. Equation (6) gives this partial derivative:

$$\frac{f}{r} = \frac{d}{r} \cdot W \cdot \left(s + d \cdot \frac{s}{d} \right) \tag{6}$$

The first term on the right-hand side of (6) is non-negative, and the second and third term together gives $\partial f/\partial d$, which is – if the entrance has the same capacity as the rest of the road – the same as $\partial F/\partial D$ according to Figure 1-II. Clearly, for densities $d>D^{\#}$, this term will become negative. When this occurs, $\partial f/\partial r$ cannot be equal to 1, and the necessary condition for a stationary equilibrium (5) is violated. A crucial assumption here is that $\partial d/\partial r$ cannot be smaller than 0 in a stationary equilibrium: more users jointly arriving at the entrance per unit of time cannot produce a lower density at that entrance in a stationary state.

For values of $d<D^{\#}$, equation (6) implies that $\partial d/\partial r$ should increase with r to maintain the stationary equilibrium condition that $\partial f/\partial r = 1$, because $W\cdot\partial^2 f/\partial d^2<0$. It is now clear that for 'equilibrium' values of $D>D^{\#}$ and $S<S^{\#}$, the system cannot be in a stationary equilibrium, as the rate of arrivals r that produces this density is necessarily greater than the inflow f consistent with this rate of arrivals according to (4). The road itself starts to act as a bottleneck in that situation. The fact that the equilibrium cannot stationary in turn implies that the system cannot be analyzed straightforwardly with a static model. In a recent paper, investigating hyper-congestion in peak congestion with a dynamic approach, Chu and Small (1996) make a comparable claim.

One can now proceed with two assumptions concerning what is going on at the entrance when an initial 'equilibrium' on the backward-bending segment of the AC-curve in Figure 2 applies. The first assumption would be that the difference between vehicles arriving, r, and vehicles flowing in, f, form a queue in front of the entrance. The variable $q=r-f$ could then give the per-unit-of-time increase in vehicles waiting to start their trip. An alternative assumption is to stick to the original physical setting, and to deny the existence of a reservoir where vehicles could queue up and wait. In this case, the number of users who are present at the entrance to start their trip at a certain instant include, apart from r, all drivers that have not yet succeeded in departing. Clearly, this latter possibility is a limiting case of the queuing model, where the capacity for queuing is zero. However, following this approach, the static modelling framework used so far cannot be used to analyze what is going on at a point such as y, because (1) drivers having arrived at the entrance at different points in time interfere directly with each other, and in order to describe the resulting behavioural responses (for instance the change of arrival rates over time), one needs a dynamic model; and (2) densities and speeds at a given place along the road will not be constant in time, and, likewise, will not be constant along the road at a given instant. Equations (1a)–(1c), as well as Figures 1 and 2, all of which presuppose a stationary state if they are taken to represent the situation along the entire road, can no longer be used.

What can be inferred qualitatively in the absence of a queuing possibility, though, is that the point y with demand E cannot be a possible initial stationary equilibrium. To see this, observe that the demand curve specified over r should be the same as the one specified over F: it gives, at a given instant and at a given place along the road, the marginal user's willingness to pay for her trip. It should not matter whether this marginal willingness to pay is measured when arriving at the entrance, or somewhere halfway the trip. It is then clear that, because the arrival rate r giving rise to y must exceed F_{max} – otherwise, we would not be at the backward bending segment of the AC-curve – it must also exceed F_y. Therefore, the marginal benefits of arrivals at the entrance must be smaller then c_y^h. In other words, y would never arise as an initial equilibrium. The only consistent equilibrium remaining is z, where a stationary equilibrium occurs with the marginal benefit of arrivals being equal to the generalized costs.

The advantage of assuming a queuing possibility is that the static diagrammatic model can still be used to describe the stationary equilibrium also in case the demand curve E′ applies. In that case, if the arrival rate exceeds F_{max}, a queue will build up, which will lead to an upward shift of the AC-curve when defined over r, because also waiting implies time losses. The total duration of the trip t_t could then be subdivided in t_q, the time spent in the queue, and t_r, the time spent on the road. Road users are not mechanically ‘pushed on the road’ at the entrance, as with the no-queue model, so they do not interfere directly with drivers who have arrived earlier and are also still at the entrance, waiting to leave. As soon as the queue has reached its equilibrium length, a fully stationary state in terms of waiting time in the queue, and speeds, densities and flows on the road, is reached. Both properties ensure that from then onwards, the system can again be described with a static model. Although also with queuing a point such as y in Figure 2 will not arise as an initial equilibrium, the above properties of the queuing model are helpful in understanding what will happen in case the demand curve E′ depicted in Figure 2 applies. Figure 3 illustrates this stationary state.

One can actually not study the transition process towards the stationary state explicitly, because (1) no explicit formulation for the queuing process itself is used and only the resulting waiting times t_q are considered, and (2) the fundamental diagram in Figure 1-I cannot be used to predict actual speeds and densities during the non-stationary part of the process. However, a qualitative description of how this state is reached can be given. In doing so, it is best to start at the beginning of the process, where the road is still empty, being very inviting to all the drivers making up the demand curve E′ at that initial instant. One can then make two assumptions about the queuing process.

The first will be called ‘efficient queuing’ (EQ). Under this assumption, the queuing process takes the same form as is usually assumed in the bottleneck model (Arnott *et al.*, 1997), namely an inflow of $f=F_{max}$, and a queue growing at a rate $q=r–F_{max}$. This implies that, before the stationary state is reached, drivers arriving at exactly the same instant are not exactly equally well off (some have to wait an instant longer than others), but they are nevertheless better off than if they would behave very ‘greedy’ and would all enter the road at the same time, as the inflow f would then be necessarily smaller than F_{max}. The process then starts with an arrival rate r for which, approximately, $E′(r)=AC(F_{max})$. A queue starts building up because $r>F_{max}$, and the arrival rate at the entrance will fall over time because the AC-curve (defined over r), now including waiting time in the queue, starts shifting upwards. The process of the queue growing, r decreasing, and the AC-curve shifting upwards continues until $r=F_{max}$, and the stationary state is reached, which can be shown diagrammatically. That is the situation where the arrival rate r (now at the end of the queue) is equal to f and to F on each point of the network, all being equal to F_{max}. Excessive demand is kept away through queuing costs, given by the line segment m-n. Therefore, the average cost curve showing generalized costs for stationary equilibria only is the one denoted AC^*. Only the lower segment of the original curve AC is part of AC^*; at F_{max} it rises vertically, showing that any marginal willingness to pay for making trips exceeding the travel costs at speed $S^\#$, density $D^\#$ and flow F_{max} will simply be translated into queuing costs.[5]

The second queuing mechanism involves 'inefficient queuing' (IQ), where drivers arriving jointly at the entrance also enter the road at the same time. Suppose, for the sake of the argument, that the DS-curve in Figure 1-I applies also during the non-stationary part of the peak. If the DS-curve would shift over time during the non-stationary phase, the qualitative discussion would not be affected. This queuing process – in particular the assumption of 'zero group velocity' reflected in the stable DS-curve – is consistent with Henderson's (1974, 1981) representation of congestion, but adds a maximum possible inflow due to the structure of equation (4), and hence a queuing possibility (see Section 4.2 for a further discussion). Then, when the process starts, the very first cohort of drivers will arrive with a rate r implying a density d higher than $D^\#$, because there is no intersection with the upward sloping segment of the AC-curve. We know at the same time that d will be lower than D_{max}; otherwise, the road would be blocked completely at the entrance straight away, implying infinite travel times for these drivers. The equilibrium value of the first level of d will be such that the generalized costs for these users are equal to the marginal benefits at the arrival rate r implying d. A queue will start to build up immediately, as $r>F_{max}$. The queue, at this instant, means that these drivers do not leave quickly enough to create enough room for their successors to start straight away with their trips, because the first drivers' inflow f is necessarily smaller than the arrival rate r. This again shifts the AC-curve, defined over r, upwards for their successors, because of the implied waiting time. The drivers arriving just after the first ones will determine their r in a way comparable to how the first ones did so: by determining that particular r, implying that particular d and s once they can start, that makes the travel time cost, but now plus the implied waiting time, equal to marginal benefits. Because there is now a positive waiting time t_q, r is just slightly lower than it was for the first drivers. This process of r decreasing over time, and the queue growing, also continues until r and f are exactly equal to F_{max}. Hence, the EQ and IQ process both lead to the same stationary state. Since f is by definition smaller under IQ than under EQ during the non-stationary part of the process, the queue will build up quicker and the stationary state is reached sooner.[6]

[5] Note that it is assumed that the demand relations defined over r are unrelated in time during the non-stationary phase discussed qualitatively: users do not consider rescheduling. When not making the trip, a potential user chooses an alternative (e.g. an alternative mode, or not making the trip at all), but does not postpone the trip. This assumption assures the demand relations to be stable over time – which is in line with standard procedures in comparative static analyses of market equilibria. 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 here. The case where rescheduling would occur due to such average cost differences over time, however, could be considered. In that case, the demand curve over arrival rates at the entrance of the road is no longer stable over time, since some users will be attracted to the non-stationary phase of the process where queues are still smaller and therefore, as implied by the discussion in the main text, average costs including waiting costs in the queue would be lower without rescheduling. An equilibrium can then be envisaged in which the sum of average travel time costs plus waiting costs in the queue plus rescheduling costs is constant over time. This case is not discussed further here, because it adds complexity while not changing the discussion fundamentally.

[6] It is not evident that IQ will converge to the stationary state asymptotically: it is conceivable that 'undershooting', in the sense that $r<F_{max}$ for some period, occurs. For reasons of space, this issue is left aside.

The marginal social cost curve MC^* can now be derived, which, like AC^*, only holds for stationary equilibria. This curve is steeper than AC^* for stationary equilibria with $r=F<F_{max}$, and is vertical for stationary equilibria with $r=F=F_{max}$. 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 can be realized with a tax j–i. Note that, at F'_o, no queuing occurs. Finally, in case a demand curve such as E″ would apply, both the non-intervention and the optimal level of traffic flow are equal to F_{max}. An optimal tax k–n will then apply, serving only to avoid queuing, and not to affect equilibrium traffic flows $r=F_{max}$. Note that the tax k–n is then exactly equal to the queuing costs that would hold in the non-intervention case, which is consistent with one of the standard results in the bottleneck model (Vickrey, 1969; Arnott *et al.*, 1997). Furthermore, observe that the AC^*-curve has to have a kink at $r=F_{max}$. Otherwise, MC^* would approach the vertical line only assymptotically, implying that marginal social costs become infinite as soon as queuing sets in. Although this would be true in steady state equilibria, where queues would be growing infinitely long, this cannot be true if queues have a finite stationary state equilibrium length. Therefore, $r=F_{max}$ indeed is the optimum in case E″ applies.

Therefore, there appears to be a fundamental fallacy in the usual analysis of a diagram such as Figure 2, which is the incorrect implicit assumption that the upper segment of the AC-curve is consistent with stationary equilibria. This ignores that, for a certain flow to be sustained at the cost level indicated by the upper segment, one needs an arrival rate at the entrance exceeding the inflow equal to that flow itself. This cannot be a stationary state, as a queue will then build up. What the present analysis has in common with those by Alan Evans (1992) and Hills (1993) is the questioning of the usual assumption that the equilibrium traffic flow F is the correct measure to base demand and cost functions on in the determination of market equilibria. What is different is the choice of arrival rates as such a variable, which are only equal to the traffic flow at each point of the link in stationary equilibria. Therefore, in contrast to the arguments put forward by the authors just mentioned, the present model in principle accepts F as the output measure, provided only stationary equilibria are considered.

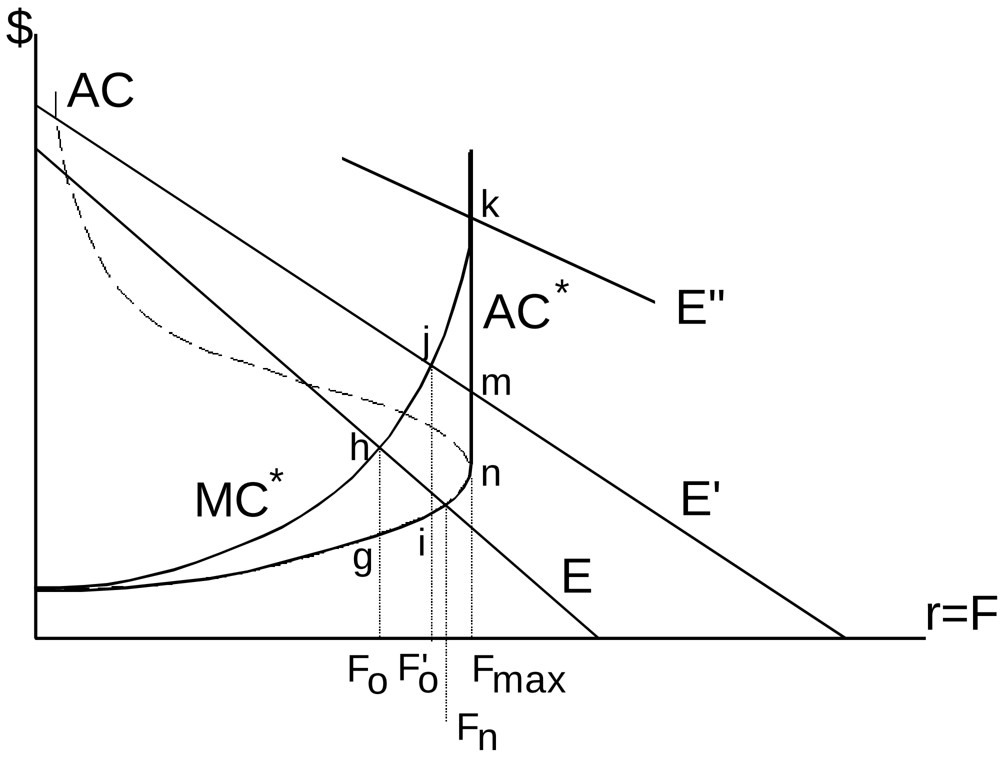

Figure C. Demand, average costs and marginal costs in stationary equilibria for the static model of continuous congestion

Finally, the above analysis implies that a static model of continuous congestion can be used directly for any optimal market equilibrium provided the average cost curve AC* is used, and for non-intervention equilibria provided the intersection of the demand curve with the average cost curves occurs on its lower segment. For other non-intervention equilibria, a queuing possibility has to be assumed in order to keep on using a static model. Furthermore, only the stationary state, with the queue being at its equilibrium length, can be fully described with that model; not the transition process towards that state. Hyper-congestion does not occur in the stationary state, and occurs in the non-stationary part only if IQ is assumed.

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 is usually a peak event. Most models of congestion are therefore, implicitly or explicitly, taken to describe peak congestion. In contrast to dynamic models of peak congestion, in which the duration of the peak is one of the endogenously determined variables, static models based on the fundamental diagram are usually remarkably careless in the treatment of the duration of the peak. However, this duration is actually a crucial variable for the consistent modelling of the market for peak road usage, also (or even particularly) for static models based on the fundamental diagram. Ignoring the duration may easily lead to the type of fallacies described in Section 2 for the case of the bread market. To see why, it can first be observed that the demand for peak travelling would naturally refer to the total number of trips accomplished during the peak (in the spirit of Hills, 1993). Using such a 'trip-based' demand function, however, the specification of the cost function deserves careful attention, because the capacity of a road increases, and therefore cost curves change, when the presumably endogenous duration of the peak changes. Likewise, in case of a ‘flow-based cost function’, giving the generalized costs as a function of passages per unit of time (see Figure 3), the specification of the demand function should particularly be checked for consistency, as the position of the demand function defined over flows then depends on the endogenous duration of the peak itself. Therefore, for a consistent static economic model of peak congestion, it is necessary to take full account of the impact of the duration of the peak on the position of either the flow-based demand function, or the trip-based cost function.

However, it seems difficult to consistently endogenize the variables T and T_G in a static model, because the duration of the peak will, in equilibrium, depend on trade-offs between travel time delays, scheduling costs, and (time-varying) tolls in case these are levied. Once the desired arrival times and scheduling costs are made explicit, however, one would normally end up with a dynamic model, with speeds and densities varying over time, unless it is somehow assumed that scheduling costs are constant over the peak. Two ways out of this dilemma for the static modelling of peak congestion based on the fundamental diagram can therefore be envisaged. The first possibility involves the explicit assumption that the grand duration of the peak, T_G, is exogenously given: all peak-travelling related activities have to take place within a given time span. To find an economic justification for this, one has to assume that scheduling costs discretely jump to a sufficiently high level for users departing earlier or arriving later than the times implied by T_G; and this particular assumption will indeed explicitly be made when considering this case in Section 4.1. T_G is then ‘endogenized’, or rather ‘consistent’, because it can be taken as an exogenously determined constant: the entire peak has to take place within the boundaries implied by T_G. By assuming the scheduling costs to be constant within T_G, the system during this interval may indeed be representable with a static approach. Both T and t then become endogenous variables.[7] Secondly, 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. This could correspond to a situation where commuters could not start working before they have all arrived. This assumption is actually as unrealistic as the usual assumption in static models of peak congestion that speeds and travel times are constant during the peak, and could therefore be defended on basis of the same flawed arguments as those that underlie the acceptance of these assumptions – which, of course, are primarily arguments relating to mathematical simplicity. Nevertheless, because this assumption is so unrealistic, the resulting model is relegated to Appendix 2.

It is perhaps surprising that either one of these rather peculiar assumptions on the pattern of scheduling costs implicitly must underlie the static models of peak congestion that have been presented in the literature. The alternative implicit assumption that

[7] Alternatively, if T were taken to be exogenously given, T_G would become endogenous. The reasons for considering the case with exogenous T_G rather than exogenous T in the sequel are (1) that it makes more sense to consider prohibitively high scheduling costs for a departure from home before a given time, than for an arrival at work before a given time; and (2) that it implies an exogenous duration in which all peak travelling activities have to take place, which is a more intuitive measure to be set exogenously.

scheduling costs simply do not exist will not work, since in that case T and T_G could be increased costlessly and no congestion would occur during the peak.[8] Although static models may therefore give a rather distorted picture of peak congestion, because other scheduling cost structures are more likely to apply in reality, the implied model with exogenous grand duration is nevertheless considered here, in particular to see to what extent the analysis in the previous section carries over to a static model of peak congestion based on the fundamental diagram, which has played an important role in the discussion in the literature. The usually implicit assumptions on the nature of scheduling costs, necessary to render the model a consistent representation of peak congestion, are made explicit by considering the step-wise scheduling cost function explicitly. The section will conclude with a qualitative discussion of to what extent the analysis could carry over to a dynamic model of peak congestion.

4.1. A static model of peak congestion with exogenous grand duration

Let us 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, $t_{D,F}^*$, which is assumed to be the same as the arrival time at the entrance of the road, and a latest arrival time $t_{A,L}^*$ at the exit of the road, which is where the workplace is. Therefore, $t_{A,L}^*– t_{D,F}^*=T_G^*$. Next, observe that in a static model, speeds, densities and flows are in principle not allowed to vary during the peak. In order to construct a model in which both features apply, it is assumed that the scheduling costs are constant – we assume zero – for those drivers departing after $t_{D,F}^*$ and arriving at the road's exit before $t_{A,L}^*$, no matter exactly when they travel; and to be prohibitively high for others, even if they depart an instant earlier than $t_{D,F}^*$, or arrive at the road's exit an instant later than $t_{A,L}^*$. An important equilibrating principle in models of peak congestion is namely that in equilibrium, no driver should be able to benefit from rescheduling: generalized user costs, including scheduling costs, should be constant over time. Because of the assumed constancy of scheduling costs in the present model, also total travel times (including the time spent in the queue, if there is one) will therefore be constant. In absence of a queue, this, in turn, implies constant speeds and densities during the peak, which are required for a static model.

This pattern of scheduling costs, involving penalties for early departures and late arrivals, therefore implies that the duration T_G is exogenously given (provided N is sufficiently large), and yields constant travel times t during the peak, at least in equilibria where no queuing occurs; implying a level of T through the equality $T_G^*=T+t$. Both T and t, however, will have to be determined endogenously. Finally, to avoid irregularities at the end of the peak, it is assumed that a driver departing later than another driver cannot arrive at the exit of the road at the same time. In other words: a driver driving in the last cohort (arriving at $t_{A,L}^*$) cannot benefit from departing later, drive at a higher speed, join the last cohort before reaching the exit of the road, and still arrive at $t_{A,L}^*$. The implied step-wise scheduling cost function can be seen as a reasonable approximation for the situation where

[8] Also the assumption that desired arrival times are distributed uniformly over time for a given period will not yield a configuration that is representable with a static model, unless the scheduling costs make discrete jumps such as described in the case of exogenous T_G. Without this additional assumption, there will be periods of increasing (and decreasing) traffic before (and after) the first (and last) preferred arrival time – unless we have the uninteresting case where everybody could drive at free-flow speed.

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 discreteness assumption, however, is a necessary requirement for using a static formulation, because scheduling costs are then independent of the arrival time as long as no-one drives outside T_G^*.

Because T is endogenous due to the equality $T_G^*=T+t$, a representation in terms of flows or arrival rates becomes problematic. In comparison to the model with continuous demand, the nature of the demand is fundamentally different, being now a demand for a total number of trips, to be fully accomplished within the time-span T_G^*. For each stationary state equilibrium without queuing, this demand in terms of N can of course be evaluated in terms of r or F by an inward rotation with a factor T, compare equation (2a). This in turn implies that for every pair of equilibrium values of T and t, a different demand relation defined over r or F would apply. It is therefore easier to specify the demand in terms of total numbers of trips accomplished over the entire peak, as proposed by Hills (1993). The consequence is that also the cost functions, which were up to now defined over flows or arrival rates, have to be defined in terms of N.

Let us start by considering stationary state equilibria without queuing; so, involving cost levels associated with the lower segment of the AC-curve in Figure 3. In order to make the transformation towards an AC-curve defined over N, it can first be observed that the relation between F and N for such equilibria can be found by rewriting (2a) as follows:

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

On the right-hand side of (7), both F and t are endogenous. A function F(t) can be constructed that depicts the equilibrium combinations of these two variables, by converting the speed-flow curve in Figure 1-IV into a travel time-flow (TTF-)curve, using the fixed relation between speed and travel time given in (1b). 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 be taken to investigate the equilibrium relation between t and N during the peak:

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

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 levels of flow smaller than the maximum flow F_{max} ($\partial N/\partial t=0$ requires $\partial F/\partial t>0$), and therefore for a travel time smaller than $t^{\#}$. 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 cost level for which the average cost curve defined over the total number of users has an infinite derivative with respect to N, at N_{max} in Figure 4-II, is lower than the (minimum) cost level consistent with F_{max} in Figure 3. The lower segment of the AC-curve in Figure 4-II implies a marginal social cost curve MSC. These can 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) and E″ applies (N_{max}, requiring a toll k–n).

The question that remains to be answered is what happens in the non-intervention case when a demand curve such as E′ or E″ applies. These curves imply that at N_{max}, there will be some drivers left with a marginal benefit for making the trip exceeding the average cost level at N_{max}. Clearly, N_{max} itself will then not be a stable equilibrium, since these drivers have the incentive to use the road. Of course, they would then depart at such an instant that they will not incur the prohibitive scheduling costs of arriving later than $t_{A,L}^*$ themselves.

As a matter of fact, the AC-curve in Figure 4 can be completed with a backward-bending section, showing that a given level of N is also obtainable at higher average costs, possibly involving queuing. In the latter case, the total travel time t_t, being the sum of the time spent in the queue and on the road (t_t=t_q+t_r) should be the same for all users, because scheduling costs are constant within the time-span T_G, and costs should be equal for all users. Therefore, the queuing process that should be assumed is inefficient queuing (IQ); see Section 3. In case efficient queuing (EQ) were assumed, with a constant inflow f=F_{max}, an equilibrium with equal user costs will not result because users departing earlier are then always better off: t_r would be the same for all, whereas t_q would increase over time.[9] Such configurations, both with and without queuing, imply values of $N<N_{max}$. For outcomes without queuing, given by the solid part of the backward-bending segment, this follows directly from (8): for $F>F^*$, the positive effect of a larger F on N is outweighed by the shorter duration at the exit due to the condition $T=T_G^*–t$. For equilibria with queuing, given by the dotted curve, N is even smaller than the level consistent with F_{max}, which is given by $N^{\#}=F_{max}\cdot(T_G–t^{\#})$. With queuing, the arrival rates for the first drivers, who have not experienced the queue at its maximum length, is going to be smaller than F_{max}, and because their travel times will exceed $t^{\#}$, also the duration of the peak at the exit of the road is smaller than $T_G–t^{\#}$. Since the others cannot have an arrival rate exceeding F_{max}, the result is evident.

This implies a backward-bending section of the AC-curve defined over N, showing that decreasing levels of N are consistent with increasing total travel times t_t and hence higher costs. Note that this AC-curve gives potential equilibrium combinations of N and AC, where total travel costs are equal for all users. For demand curves that intersect the AC-curve only once, and only at the backward-bending segment, the intersection (for instance point x in Figure 4-II) then gives the unregulated market equilibrium. This may seem odd in the light of the discussion in Section 3.1, Figure 2, where the configuration x – before considering the question of stationarity in Section 3.2 – was classified as unstable because beyond this 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. Since a departure at any of the instants available would imply marginally higher travel costs at that particular instant because of the higher density, the marginal private

[9] The reason that the EQ does not work here, whereas it does in the bottleneck model of peak congestion, is that in the latter case scheduling costs are not constant over time.

costs are increasing for additional drivers at each instant of arrival at the entrance, and therefore cannot coincide with the falling average social costs in this region.

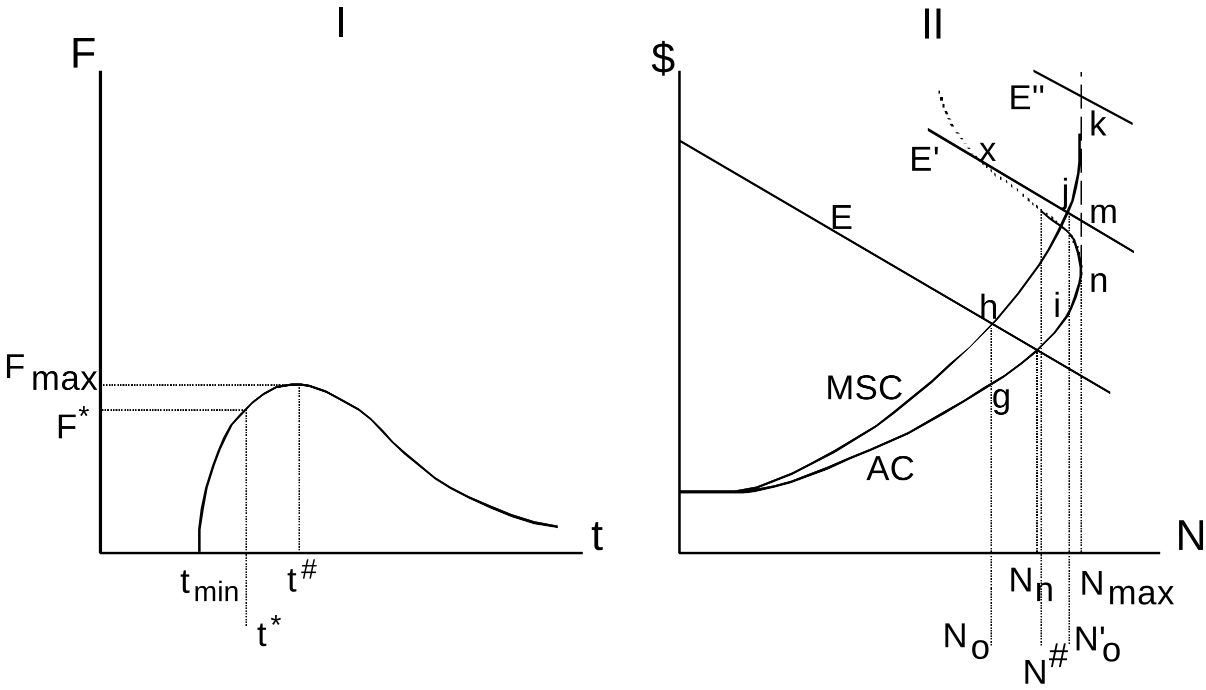

Figure D. The travel-time flow relation (I) and the diagramm atic representation of the static model of peak congestion with exogenous grand duration (II)

Note that in these equilibria, total travel times t_t are relatively large and the duration T is relatively short, implying that also the departures from home are relatively concentrated in time. This reflects that those agents using the road in the equilibrium, by departing relatively shortly after $t_{D,F}^*$, avoid arriving late, and at the same time know that through the relatively high cost the threat of potential road users entering the market is no longer manifest. Therefore, these users are indeed prepared to accept a relatively large t_t. Interestingly, the optimal taxes j–i and k–n in these cases lead to a higher number of road users in the optimum than in the non-intervention equilibrium. The reason is that the necessity of selecting early departure times in order to avoid late arrivals is no longer there, since the potential road users on the right-hand side of N_{max} no longer have the incentive to use the road, because of the tax.

Finally, in the event that the demand curve intersects the AC-curve defined over N three times (compare Figure 2) or twice (for instance in case of a kinked demand), things become somewhat complicated. As outlined in Appendix 1, the question of which configuration will finally come about as the unregulated market equilibrium then depends on the size of the penalty for travelling outside T_G^*. The lower the penalty, the more likely the more favourable equilibria are to come about (see Appendix 1).

Although configurations on the lower segment of the AC-curve defined over N obtained with the above model may be a reasonable representation for peak congestion in reality, it is fair to admit that when demand is relatively high, the model becomes problematic. Most equilibria on the backward-bending segment of the AC-curve defined over N involve queuing, so the primary purpose of creating a static model of peak congestion according to equations (2a)–(2c) actually has failed because the equilibrium involves the building up of a queue. As argued in the previous section, this process, where densities exceed $D^{\#}$ in Figure 1-I and speeds and densities are not constant over time and place raises the question of the applicability of the fundamental diagram.

Nevertheless, it is striking that for the 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 without queuing. Hyper-congestion does not occur in optima, nor in (purely static) non-intervention outcomes without queuing, even if these involve equilibria on the backward-bending segment of the AC-curve defined over total numbers of road users. Because IQ is the only queuing process consistent with constant scheduling costs, however, hyper-congestion does occur in (non-static) non-intervention outcomes involving queuing.

4.2. Implications for dynamic models of peak congestion

Finally, it can be mentioned that 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 requires rather heroic assumptions on the structure of scheduling costs. Moreover, as demonstrated in the previous analysis, as soon as demand is relatively high relative to the exogenous grand duration of the peak and the maximum possible flow, the model predicts a queuing process, and the stationary state requirement for a static model to be meaningful is not fulfilled. The assumed stepwise scheduling cost function then becomes the main driving force in the model, and it becomes increasingly worthwhile to relax this assumption and to consider situation in which 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; where c gives the travel time costs (including time spent in the queue). For flow congestion, this means that one has to find a formulation that can replace the relation between speed and density in Figure 1-I 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.[10] 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 of these authors 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 assumed below).

Although Chu (1975) has pointed out that overtaking could be a problem in Henderson's (1974, 1981) formulation, for the present qualitative discussion based on the previous sections it is convenient to consider the 'Henderson-type' of flow congestion, and to make the additional assumption, like Henderson does, that overtaking is not possible.[11] Using the notation introduced in Section 3, suppose that the speed for a certain cohort as a function of the arrival rate r is given by s(d(r)), as implied in (4), so that the inflow has a maximum value f_{max} and the average travel costs function c(r) has a shape comparable to the AC(r) function depicted in Figure 3. As soon as the arrival rate exceeds the maximum inflow f_{max}, a queue therefore develops before the entrance of the road. Two types of non-intervention equilibria can then be considered, the first of which involves $r_t=f_t=s_t(d_t(r_t))\cdot d_t(r_t)\cdot W\leq f_{max}$ for all arrival times t at the entrance. 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.*, 1997), face scheduling costs k_{max} for which:

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

where $c_{min}(f_{max})$ is the minimum travel time cost consistent with the maximum inflow (hence, it does not include 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 can subsequently be derived (Henderson, 1974, 1981; Chu, 1995):

$$_t = r_t \cdot \frac{c}{r_t} \qquad (10)$$

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. Assuming the EQ queuing process as outlined in Section 3, one could at first

[10] 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.

[11] Moreover, because (1) overtaking cannot occur in equilibrium anyway, because the drivers being overtaken are necessarily worse off than the 'overtakers', and (2) both Chu (1995) and Henderson (1974, 1981) assume zero group velocity, it can be argued that both models produce exactly the same equilibria provided capacity is constant along the road.

glance then expect a situation in which ‘Henderson’ tolls would apply for the first and last phases of the peak where $r_t \leq f_{max}$; and ‘Vickrey’-bottleneck tolls would be necessary to avoid all queuing for the middle period where $r_t > f_{max}$ 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. This type of integrated modelling, as also proposed by Rouwendal (1990), certainly deserves further attention in future work.

5. Conclusion

Although the static model of road traffic congestion based on the ‘fundamental diagram’ has been around for some 75 years now, there still seems to be major disagreement about the exact formulations that should be used. An important source of disagreement concerns the choice of the output variable in the definition of demand and cost functions. Two main stances can be distinguished here, namely ‘flow-based’ measures and ‘stock-based’ measures. The foregoing analysis suggests that the choice of the correct output measure is directly related to the type of demand that is assumed to apply: ‘continuous demand’ on the one hand, requiring a formulation in flows by definition, or ‘peak demand’ on the other, which can most naturally be analyzed using a formulation in total numbers of trips accomplished during the peak. In performing the analyses, it was acknowledged that a meaningful static equilibrium should satisfy the condition of stationary stability, and only static configurations that do represent stationary stable configurations were accepted.

Section 3 investigated the case of continuous demand, focusing in particular on the difficulties arising with the backward-bending average cost curve defined over flows that have dominated much of the debate in the literature. The relevance of the backward-bending segment of the cost curve for the static analysis of congestion was questioned by demonstrating that ‘equilibria’ on this segment produce upwards shifts of the cost curve itself, and can therefore not be of a stationary nature. In particular, the analysis ignores that, for a certain flow to be sustained at the cost level indicated by the upper segment, one needs an arrival rate at the entrance exceeding the inflow equal to that flow itself. This cannot be a stationary state, as a queue will immediately build up. Using the notion that, in a stationary equilibrium, the arrival rate of drivers at the entrance should be the same as the flow on any point along the link, the model presented in principle accepts traffic flows as the output measure for the static analysis of continuous congestion, provided only stationary equilibria are considered. The analysis implies that a static model of continuous congestion can be used directly for any optimal market equilibrium provided the lower segment of the average cost curve defined over flows is used, and for non-intervention equilibria provided the intersection of the demand curve with the average cost curves occurs on its lower segment. When such an intersection does not exist, because the marginal benefit exceeds the average cost at the maximum flow, the model can only be used if it is assumed that queuing before the entrance of the road is possible. This causes the average cost curve to rise vertically at the maximum flow, due to queuing costs. For the associated non-intervention equilibria, involving queuing, only the stationary state, with the queue at its equilibrium length, can be described with the model; not the transition process towards this state. Hyper-congestion does not occur in the stationary state, and occurs in the non-stationary phase only with 'inefficient queuing' (IQ).

In Section 4, it turned out that the analysis of peak demand with a static model based on the fundamental diagram can be quite problematic, because for a consistent static model of peak congestion, it is necessary to fully take account of the impact of the duration of the peak on the position of either the flow-based demand function, or the trip-based cost function. However, it seems difficult to consistently endogenize the duration of the peak in a static model, because this duration will depend on trade-offs between travel time delays, scheduling costs, and (time-varying) tolls in case these are levied. Once these desired arrival times and scheduling costs are made explicit, however, one would normally end up with a dynamic model, unless it is assumed that scheduling costs are constant during the peak. To solve this dilemma for the static modelling of peak congestion, the explicit assumption was made that the grand duration of the peak is exogenously given, because of a particular step-wise scheduling cost function. No matter how unrealistic this assumption may appear, it seems that it is at least preferable to models in which the duration of the peak is completely ignored. Two conclusions arose 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. Hyper-congestion does not occur in optima, nor in purely static non-intervention outcomes without queuing, even if these involve equilibria on the backward-bending segment of the AC-curve defined over total numbers of road users. Because IQ is the only queuing process consistent with equilibria involving constant user costs when scheduling costs are kept constant by assumption, hyper-congestion does occur in non-intervention outcomes involving queuing.

Finally, it was concluded that the static framework presented in Section 4.1 can be extended to a dynamic model which combines elements of flow congestion with bottleneck congestion. When the scheduling cost curve is sufficiently steep and demand is sufficiently high, this model produces non-intervention outcomes where the road operates as a bottleneck in the middle of the peak, and flow congestion occurs in the first and last stages. Under the standard assumption in bottleneck models that queuing is 'efficient' (EQ), hyper-congestion does not occur in this model. All queuing is eliminated under optimal tolling, but optimal travel delays are positive. Inflows below the maximum inflow apply throughout the optimized peak, and therefore only 'Henderson tolls' apply. This model certainly deserves further attention in future work.

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 (1997) "Recent developments in the bottleneck model". In: K.J. Button and E.T. Verhoef (1997) *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* Edward Elgar, Cheltenham (forthcoming).

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.

Foster, C. (1974) "The regressiveness of road pricing" *International Journal of Transport Economics* **1** 133-141.

Foster, C. (1975) "A note on the distributional effects of road pricing: a comment" *Journal of Transport Economics and Policy* **9** 186-187.

Gibbons, R. (1992) *A Primer in Game Theory*. Harvester Wheatsheaf, New York.

Hargreaves Heap, S.P. and Y. Varoufakis (1995) *Game Theory: a Critical Introduction*. Routledge, London.

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 D: Transport and Environment* forthcoming.

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.

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.

Richardson, H.W. (1974) "A note on the distributional effects of road pricing" *Journal of Transport Economics and Policy* **8** 82-85.

Richardson, H.W. (1975) "A note on the distributional effects of road pricing: a rejoinder" *Journal of Transport Economics and Policy* **9** 188.

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

Small, K.A. (1992) *Urban Transportation Economics*. Fundamentals of Pure and Applied Economics **51**, Harwood, Chur.

Small, K.A. and J.A. Gomez-Ibañez (1997) "Road pricing for congestion management: the transition from theory to policy". In: K.J. Button and E.T. Verhoef (1997) *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., 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.

Appendix 1: Multiple equilibria in the static model of peak congestion with exogenous grand duration

This appendix considers the case where in the static model of peak congestion with exogenous grand duration (see Section 4.1), the demand curve defined over N intersects the AC-curve defined over N three times (compare Figure 2) or twice (for instance in case of a kinked demand).

Before addressing this issue, first observe that in general for an equilibrium to arise in dynamic models of peak congestion, individual road users should somehow be capable of creating the equilibrium they all expect. In a model of peak congestion, where individual agents should not only make the choice of *whether* to use the road, as in the model of continuous congestion, but also *when* to depart from home, this would require coordination if individuals are supposed to play 'pure strategies' in terms of selecting one particular departure time. If private costs do not vary over the peak, and it is immaterial exactly when one is travelling, it would otherwise be unclear what mechanism should secure the equilibrium pattern of departure times to arise, as there would be countless Nash equilibria. It is therefore more natural to assume that an equilibrium requires a 'mixed strategy' to be played by every individual, specifying a probability density function of departure times. In a symmetric game, the same probability density function should then be selected by all individuals, and it should correspond with the distribution of departure times that actually produce the candidate Nash equilibrium under consideration. If the number of individuals is sufficiently large, the equilibrium may then be expected to actually arise. This representation of equilibria will also be followed here.

The question of which configuration will finally come about as the unregulated market equilibrium in this static model of peak congestion then depends on the size of the penalty for travelling outside T_G^*. Suppose there are three intersections x, y and z, where $N_x<N_y<N_z$ and $AC_x>AC_y>AC_z$ (compare Figure 2). Consider the smallest group of users N_x, and suppose that they only consider three possible equilibria to eventually arise, namely the potential equilibrium configurations x, y or z where user costs are the same for all users and marginal benefits are equal to average costs, and that they ignore the impact of their own decision on the determination of the equilibrium. That is, they are pure price-takers. For reasons of exposition, their strategy set is restricted to the following choices: three equilibrium strategies S_x S_y S_z, involving departure time probability density functions consistent with x, y and z; and a fourth strategy S_4 involving a discrete departure time outside T_G^* implying, with certainty, travel costs consisting of the sum of free-flow costs plus the penalty. As soon as an N_x-driver finds S_4 more attractive than S_x if all other N_x-drivers would play S_x, he will realize that this also holds for the other N_x-drivers. All N_x-drivers will then infer that x cannot be a Nash-equilibrium, and will expect either y or z to arise.

Now if the penalty for late arrivals is 'sufficiently high' in the sense that AC_x is smaller than the sum of the free-flow costs plus the penalty for late arrivals, and if the N_x-drivers attach only the slightest possibility to x actually occurring, they will want to avoid a late arrival in all cases, and they will all be driven to play S_x, so that x is the Nash equilibrium. S_x then minimizes their expected costs of travelling, even if y or z would arise, as it certainly avoids the penalty. Other drivers, knowing this, then have no incentive to join the market. However, if the sum of the penalty plus the free-flow costs penalty is less than c_x but more than c_y, so that the penalty is only 'sufficiently high' in the sense that it prevents travelling outside T_G^* in case equilibrium y arises, x is no Nash equilibrium because N_x-drivers would prefer S_4 to S_x in configuration x. Only y and z remain as possible candidates for the Nash equilibrium. Depending on whether similar reasoning applies for the comparison between y and z, either y or z can then finally be identified as the Nash-equilibrium, with either N_y or N_z playing the associated strategies in terms of selecting the appropriate departure time probability density function. So, the level of the penalty determines the selection of the equilibrium. The lower the penalty, the more likely the more favourable equilibria are to come about. Note that the sum of free-flow costs and the penalty should of course at least exceed AC_z.

Appendix 2: A static model of peak congestion with endogenous duration

As mentioned in the main text, the endogenization of a variable duration of the peak T in a static model of peak congestion requires the assumption that scheduling costs are constant over the peak; depending on the duration of the peak only, and not on the actual arrival time within this peak. If not, the model would be internally inconsistent, as the assumption of constant travel times associated with a static approach would then imply that users arriving nearer the desired arrival time would be better off then others, so that the system could not be in equilibrium. [12]

Accepting this assumption for the sake of being able to investigate the implic ations of the analysis in Section 3 for the case of a static model with peak demand and an endogenous duration T, the following model can be constructed. First, the generalized user costs AC now consists of two components: travel time costs c, and scheduling costs k. Travel time costs c, as a function of flow F, have the same shape as the AC^*-curve in Figure 3. Therefore, for cost levels c>n, one could again subdivide travel time costs into time costs incurred when waiting in the queue, and time costs when driving on the road. The scheduling costs are constant over the peak, and increase with T. Congestion then manifests itself as two externalities: travel time delays and scheduling costs. Hence, average costs can for equilibria without queuing be written as:

$$AC = c(F) + k(T) \tag{A1}$$

An important equilibrium condition that can be used to infer the specific combination of F and T that will arise for a given total number of road users N, is that in a Nash equilibrium no driver should be able to be better off – obtain lower average costs – through rescheduling. This may lead, for a given N, to various possible equilibrium levels of F and T, depending on some assumptions to be made explicit below. The set of possible equilibrium combinations of F and T may vary from the situation where average costs are minimized, to the limiting case that free-flow speeds apply throughout the peak. To see why this is the case, first two bench-mark representations of the market process will be considered.

The first bench-mark case concerns perfect information and perfect foresight, where all drivers know all other drivers' departure times with certainty. Consider an initial situation where T is smaller and hence F is higher than the average cost minimizing levels for that particular N. It is then attractive for one driver to travel just an instant later than in the initial situation. This increases T, and hence k, for all drivers, whereas the travel time for that single last driver is smaller than for the rest. Because information is perfect, others will join him, until travel times are again equalized over the peak. The average costs are now smaller for all drivers, as we started in the situation where T was smaller and F larger than the cost minimizing combination. The first driver postponing his trip will therefore not regret his decision. This process will continue until the cost minimizing combination of T and F is reached. Then, no driver has an incentive to unilaterally postpone his trip, because he knows that others would keep on joining him to enjoy his smaller travel time, until travel times are again equalized. In the resulting equilibrium, also this driver himself would then be worse off, because we now started from the average cost minimizing combination of T and F. So, for each level of road usage N, the cost minimizing combination of T and F will arise when information and foresight are perfect.

The other bench-mark case concerns the situation where the morning peak is a one-shot game, where all agents do know the others' pay-off functions, but not their actual actions. In this case, the agents end up in a prisoners' dilemma, because the above described average cost minimizing combination of T and F is not a Nash equilibrium. A driver could then unilaterally benefit from driving an instant earlier or later, if all others would stick to the above strategy. He will then be able to drive at the free-flow speed, so he will make a non-marginal travel time gain,

[12] Alternatively, one could of course assume that average scheduling costs are constant during the peak and depend on T_G. Analytically, this would lead to comparable results as the model discussed in the main text. For reasons of space, and because it is probably less realistic than the case discussed in the main text, this variant is not considered.

and will incur only marginally higher scheduling costs. It is then intuitively clear that in the resulting Nash equilibrium, F will be smaller and T will be higher than in the perfect information case. Moreover, if the drivers are pure price-takers and do not take into account the impact of their own decision on k(T), the only Nash equilibrium is the situation where all drivers drive at free flow speed.

Neither representation is particularly realistic for the choice of departure times. It is likely that the game could be best represented as an infinitely or indefinitely repeated game. According to the 'Folk Theorem' in game theory, 'anything goes' in such games (Gibbons, 1992; Hargreaves Heap and Varoufakis, 1995). For the present case, this means that both extremes mentioned above, namely the minimum average cost outcome and the free-flow speed outcome, as well as all intermediate combinations of F and T, could be sustained as Nash equilibria. Whereas the argument for the free-flow speed equilibrium is the same as the one given above, for the minimum average cost equilibrium this requires the agents to play a trigger strategy, where the joining of the agent that postponed his trip will occur in the next stage, and will be maintained ever after. With a discount factor sufficiently close to one (i.e. a discount rate sufficiently close to zero), this boils down to the same type of reasoning as given above for the perfect information/perfect foresight setting.

Even in the average cost minimizing equilibrium, where flows are relatively high compared to the one-shot Nash equilibrium, only the lower segment of the travel time cost function is relevant, and outcomes with queuing will not arise. Only equilibria where $\partial c/\partial F$ (the slope of the AC-curve in Figure 4) is finite and positive will generally arise. This can be demonstrated by substitution of T=N/F in equation (A1); compare equation (2a). This allows us to derive, for a given N, the average cost minimizing level of F, since the necessary first-order condition can be written as:

$$\frac{AC}{F}=\frac{c}{F}+\frac{k}{T}\cdot\frac{-N}{F^2}=0 \tag{A2}$$

Queuing, which occurs in equilibria with $\partial c/\partial F=\infty$ according to Section 3, will not arise in the minimum average cost Nash equilibrium as long as $\partial k/\partial T<\infty$. In general, the second term in the middle expression of (A2) is negative, so that $\partial c/\partial F$ must be positive. Because F is lower and T is higher in all other possible Nash equilibria, it can be concluded that queuing will not arise in the static model of peak congestion with endogenous duration.

Figure 5 shows the diagrammatic implications. First, panel 5-I shows why only configurations involving the lower segment of the AC-curve over flows, from Figure 3, are relevant. Two sorts of curves are depicted, the first of which are iso-AC-curves, showing combinations of traffic flow F and the duration of the peak T for which the average cost according to (A1) take on the same value. These curves can be derived by observing that c(F), as given in Figure 4, increase with flow. In order to maintain the same level of AC, therefore, increasing savings in scheduling costs are required. At F_{max}, where c(F) is vertical according to Figure 4, also the iso-AC-curve in Figure 5-I is vertical. Next, the iso-N-curves give those combinations of F and T satisfying equation (2a): $N=F\cdot T$. Average cost minimizing combinations of F and T can be found as points of tangency of an iso-N-curve and an iso-AC-curve. In accordance with (A2), it also follows from Figure 5-I that these points of tangency involve outcomes involving the lower segment of the (travel time) cost curves over flows shown in Figures 3 and 4. Two such combinations are indicated, denoted with subscripts 1 and 2. For other Nash equilibria, the possible combinations of T and F are given by points on the iso-AC-curves to the north-west of these minimum average cost equilibria. Also in those equilibria, queuing does not occur.

Next, Figure 5-II shows that for minimum average costs Nash equilibria, the average cost will be an increasing function of N as long as $\partial k/\partial T<\infty$, because both c(F) and k(T) will be larger for larger values of N. Confronting this AC-curve defined over N, and the implied MSC-curve, with the demand curve E shows that E and AC have only one intersection. The non-intervention level of road usage during the peak is N_n, and the optimal level is N_o, requiring an optimal toll h–g.

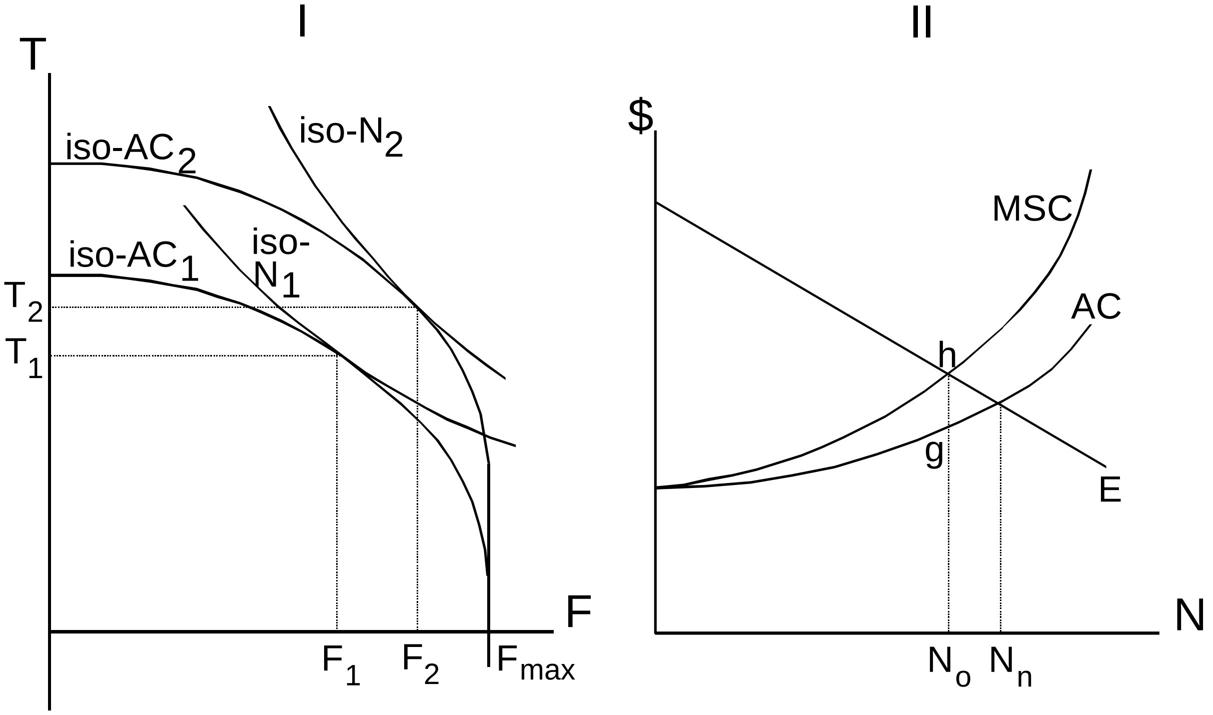

Figure E. model *Iso-AC and iso-N curves (I) and the diagrammatic representation of the static of peak congestion with endogenous duration (II)*

If the agents do not play the optimal trigger strategy, the AC-curve will obviously be higher. If the regulator is not able to affect departure time decisions, however, the diagrammatic representation is qualitatively identical to Figure 5-II, be it that equilibrium values of N_n and N_o will be lower, and the optimal tax h–g will be higher. If the regulator could affect departure time decisions, by setting sufficiently high tolls for arrivals before or after the optimal duration T, the optimum is the same as under the optimal trigger strategy, and would involve a downward shift of the AC-curve to the position consistent with this strategy.

In conclusion, the static model of peak congestion with endogenous duration has no unique Nash equilibrium for non-intervention when it is assumed that agents play an infinitely or indefinitely repeated game. However, non of the possible Nash equilibria involves queuing or hyper-congestion.