SPEED-FLOW RELATIONS AND COST FUNCTIONS FOR CONGESTED TRAFFIC: THEORY AND EMPIRICAL ANALYSIS

Erik T. Verhoef[*]
Department of Spatial Economics, Free University Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam The Netherlands
Phone: +31-20-4446094 Fax: +31-20-4446004
Email: everhoef@econ.vu.nl

Abstract

A dynamic 'car-following' extension of the conventional economic model of traffic congestion is presented, which predicts the average cost function for trips in stationary states to be significantly different from the conventional average cost function derived from the speed-flow function. When applied to a homogeneous road, the model reproduces the same stationary state equilibria as the conventional model, including the hypercongested ones. However, stability analysis shows that the latter are dynamically unstable. The average cost function for stationary state traffic coincides with the conventional function for non-hypercongested traffic, but rises vertically at the road's capacity due to queuing, instead of bending backwards. When extending the model to include an upstream road segment, it predicts that such queuing will occur under hypercongested conditions, while the general shape of the average cost function for full trips does not change, implying that hypercongestion will not occur on the downstream road segment. These qualitative predictions are verified empirically using traffic data from a Dutch bottleneck. Finally, it is shown that reduced-form average cost functions, that relate the sum of average travel cost and average schedule delay costs to the number of users in a dynamic equilibrium, certainly need not have the intuitive convex shape, but may very well be concave – despite the fact that the underlying speed-flow function may be convex.

Keywords: Traffic congestion; Road pricing; Car-following theory; Speed-flow relations; Cost functions
Topic Area: C7 Traffic Simulation Models

1. Introduction

Notwithstanding the long history of the economists' and engineers' study of road traffic congestion and congestion pricing (see Pigou, 1920; Knight, 1924; Wardrop, 1952; Walters, 1961; and Vickrey, 1969), academia has not yet reached general consensus on the fundamentals that should underlie such analysis. This is illustrated by the relatively large number of comments and replies that papers on congestion modelling seem to trigger (Else, 1981, 1982, versus Nash, 1982; De Meza and Gould, 1987, versus Alan Evans, 1992; Andrew Evans, 1992, 1993, versus Hills, 1993; and Ohta, 2001ab, versus Verhoef, 2001b). Much of the debate centres around the modelling of cost functions for congested road use in the context of the conventional static economic model of traffic congestion, which – following Walters (1961) – uses the speed-flow function as the basis for the average cost function of

[*] Affiliated to the Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam. The research of Erik Verhoef has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. The author would like to thank OC&C for kindly providing the data used in this paper.

road use (see also Section 2 below). Two issues have been particularly heavily debated in the studies just mentioned: the economic interpretation of ‘hypercongestion’[1] and its stability as an equilibrium outcome, and the question of which argument should be used in the definition of demand and supply functions used for characterizing equilibria in the market for road trips. These issues are not entirely unrelated: an analyst who uses traffic density as the argument in the cost and demand functions may conclude that a unique, stable market equilibrium may occur at the hypercongested branch of the speed-flow function (*e.g.* Ohta, 2001a).

This paper reconsiders the derivation of cost functions from speed-flow relations from a dynamic perspective, with an emphasis on the phenomenon of ‘hypercongestion’. To that end, the paper first summarizes the main results obtained with a dynamic extension of the conventional model, presented in two earlier papers (Verhoef 2001a, 2003), and will then test these empirically by contrasting the model’s qualitative, aggregate predictions to insights obtained from empirical data for a well-known bottleneck on the Dutch road network (The Coenplein, near Amsterdam). This dynamic extension involves a simple car-following model, where the car-following equation is a first-order differential equation, linking a driver’s instantaneous speed choice to the distance from her leader. Because distance is the inverse of density, and because the conventional model is based on a relation between density and speed, the car-following model is easily made consistent with the conventional model for stationary state traffic. Two variants of this model will be considered: one in which there is a single homogeneous road of constant capacity, and one in which there is a bottleneck halfway this road due to a reduction in the number of lanes.

Dynamic stability analysis for the homogeneous road setting – the situation for which the conventional model predicts that all (positive) flows below capacity could in principle occur both as a hypercongested and as a normally congested equilibrium – demonstrates that hypercongested stationary states are in fact dynamically unstable. This means that there are no equilibrium paths to such a stationary state from any other stationary state (hypercongested or not). This finding will be substantiated in Section 3, and the implications for the shape of the average cost function (for full trips) that can be derived from the speed-flow function will be identified. These are that this function will coincide with the conventional function for normally congested speeds, but instead of bending backwards it will rise vertically at the road’s capacity (if queuing before its entrance is allowed).

In contrast, when applied for a road with a bottleneck, the model shows that in a dynamic equilibrium with endogenous departure times (*á la* Vickrey, 1969), hypercongestion will occur with certainty during the equilibrium peak, provided the equilibrium number of users over the peak, relative to the capacity of the downstream road segment, is sufficiently large. However, hypercongestion will prevail on the upstream, high-capacity segment of the road, whereas the speed on the downstream, low-capacity segment will asymptotically approach the constant value consistent with the maximum flow (*i.e.*, the capacity) of that segment. This will be discussed further in Section 4.

Section 5 will compare the main qualitative predictions of the theoretical model with empirical traffic data, which include observations upstream and downstream of a bottleneck. It is shown that these are largely consistent.

Finally, Section 6 will consider reduced-form average cost functions, that relate the sum of average travel cost and average schedule delay costs to the number of users in a dynamic equilibrium, for the proposed car-following model of traffic congestion.

[1] ‘Hypercongestion’ is the term used in the economics literature to indicate the lower branch of the speed-flow relation for which density has become so high and speed so low that the traffic flow (the product of speed and densities in stationary traffic conditions) has fallen below its maximum possible value (see also Section 2 below). The upper branch of the speed-flow function, where the same flow levels are obtained at a higher speed and a lower density, is called ‘normally congested’. In the engineering literature, this normally congested branch is often referred to as ‘uncongested’ and the hypercongested branch as ‘congested’. The economics terminology will be used in this paper.

2. The conventional model and a continuous-time – continuous place generalization

2.1. The conventional economic model of traffic congestion

The conventional model of traffic congestion considers stationary state traffic on a single road with identical users. The main advantage of this approach is its – relative – simplicity, owing to the fact that speeds (*S*), densities (*D*) and flows (*F*) are pre-supposed not to vary over time nor place (*i.e.*, along the road). Because in reality, stationary traffic conditions could occur only if the (inverse) demand function were stable over time, a logical implication of taking a stationary state perspective appears to be that the inverse demand function in such a model be defined with respect to the number of completed trips per unit of time, and hence traffic flow (strictly speaking measured at the road's exit, but due to the assumed constancy of flow along the road, this can be equated to 'flow' in general).[2]

The basis of the standard static model is the engineers' 'fundamental diagram of traffic congestion', which depicts how speed *S* falls monotonously with traffic density *D*. Because traffic flow *F* is identically equal to the product of speed and density in a stationary state (see, however, also footnote 4), the well-known backward-bending speed-flow function can be derived from this fundamental diagram (a sufficient condition being that a maximum density exists for which speed falls to zero). Walters (1961) observed that the inverse of speed (travel time per unit road length) multiplied by the constant length *X* of the road and by the value of time *vot* (assumed exogenous and constant) reflects the average (time) costs *AC* per trip (ignoring other costs of travel). The backward-bending speed-flow function thus implies a backward-bending AC-curve, which is depicted in Figure 1.

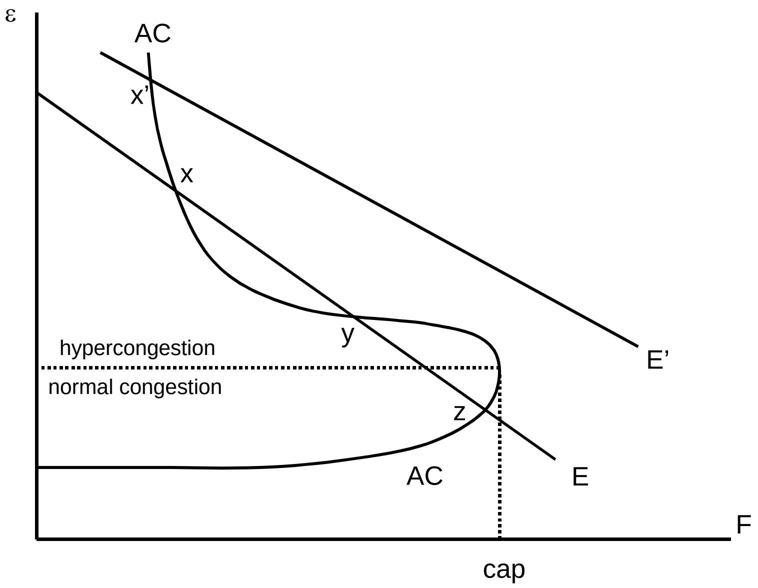


Figure 1. The backward-bending average cost curve (AC) and two inverse demand curves (E and E′) defined over traffic flow (F)

Every positive flow below capacity (*cap*) is attainable as a 'hypercongested' equilibrium with high costs (above the horizontal line) and as a 'normally congested' one with

[2] The possible alternative argument of density (*D*) would be appropriate for an inverse demand function only in the implausible circumstance (at least for commuting) where people's utility were derived directly from *being on the road*, instead of from completing a trip (and hence from *leaving the road*); whereas the 'total number of trips accomplished' (*N*) is not defined in a stationary state: this number increases linearly with the time period considered. It should be acknowledged, however, that this issue of identifying the appropriate argument is a source of ongoing debate in the literature. For instance Ohta (2001a) proposes density, and Hills (1993) proposes the total number of trips explicitly uncorrected for a time dimension; both considering a static congestion technology.

low costs (below the horizontal line). The confrontation with a downward-sloping inverse demand function (E) yields puzzling results, feeding the debates mentioned earlier (see also Chu and Small, 1996). Three candidate equilibria are suggested: x and y (both hypercongested), and z (normally congested). Candidates y and z appear stable for quantity perturbations (a slight increase in flow leads to $AC>D$, so that flow will be pushed back, and *vice versa* for a slight decrease in flow), and x and z for price perturbations (a slight decrease in the price (AC) leads to excess demand, so that the price will be pushed upwards, and *vice versa* for a slight increase in the price). Even stability analysis therefore appears inconclusive, and the model in general cannot predict whether or not hypercongestion will occur in equilibrium, regardless of whether price or quantity perturbations are considered the appropriate type to consider. Moreover, with a demand curve such as E', only one (hypercongested) candidate equilibrium x' remains, which is, however, unstable for quantity perturbations. If one believes that quantity perturbations would be the appropriate type to consider for a congested road in absence of tolling, this would discomfortingly and counter-intuitively imply that no stable equilibrium exists for a demand function such as E', and the model simply cannot predict what will happen.

An important shortcoming of the above stability analysis is that the stationary state approach only allows consideration of simultaneous and identical changes in traffic conditions along the entire road. This ignores that in reality, traffic flows, densities and speeds along the road will be endogenous variables, dependent on the (history of) arrival rates of ‘new’ users at the road’s entrance – somewhat confusingly referred to as the departure rate (from home) in the sequel. Downstream conditions will typically not respond immediately to a change in this departure rate. To incorporate this idea into the model, its presupposed inherently static and non-spatial nature (F, S and D obtain the same value over time and along the entire road) nature has to be abandoned. The next sub-section describes how this can be accomplished.

2.2. A continuous-time – continuous place generalization of the conventional model based on first-order car-following modelling

One simple transformation suffices to obtain the simplest plausible continuous-time – continuous-place generalization of the static model discussed above. This involves converting traffic density into its inverse: distance between cars. The proposed formulation stipulates that an individual driver i, at each instant t, chooses a speed S_i^t as a function of the distance δ_i^t from the car in front of him – his ‘leader’, in car-following terminology. Consistency (for stationary states) between a standard static model based on a speed density relation $S^D(D)$ and its proposed continuous-time – continuous-place counterpart can easily be established. This simply requires using a car-following equation for which $S_i(\delta_i^t)=S^D(D)$ for $\delta_i^t=1/D$.

For a given initial state (*e.g.* an empty road, or an exogenously defined stationary state), and for any exogenously (as in Section 3) or endogenously (as in Section 4) determined time profile of departures from home, the model then endogenously determines traffic conditions for each location x and each time t. Typically, an individual driver’s speed varies over time and location, the speed at a given location varies over time, and the speeds at a given time vary with location. This dimensional richness (*i.e.*, the continuous-time – continuous place congestion technology) comes at a price: no analytical closed-form solutions exist other than for stationary states. The model namely consists of a set of N (the number of users considered) interdependent first-order partial differential equations[3] of the type:

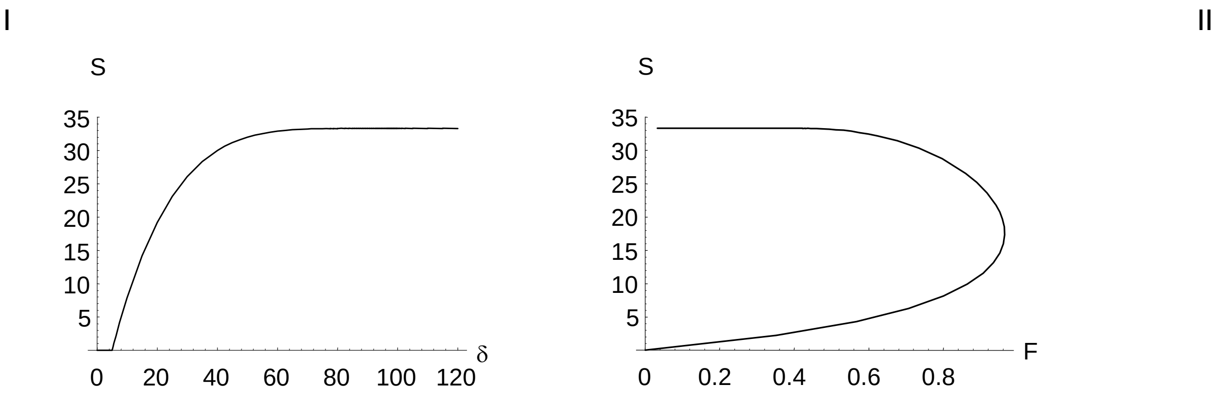
[3] The first-order structure of the car-following equation distinguishes it from standard car-following models, where typically a driver’s acceleration is defined as a function of the speed difference with and distance from his leader. However, also in the present model, a driver will be accelerating when driving slower than his leader, and no infinite de- or acceleration will occur provided the first driver modelled never drives at infinite speed.

$$S_i^t \equiv \dot{x}_i^t = S(\delta_i^t) \equiv S(x_{i-1}^t - x_i^t) \tag{1}$$

Nevertheless, a number of general properties of the model were derived in Verhoef (2001a, 2003), that do not depend on a specific choice for the function $S(\delta_i^t)$, but only require it to be continuous, smooth for positive speeds, increasing for speeds below the free-flow speed, to obtain a constant maximum 'free-flow' speed for distances exceeding some value, and to obtain a value of zero for some minimum (positive) distance, approximately the length of a vehicle. For the purpose of illustrating the results, a numerical version of the model was used in these papers, and will also be used below, where the distance-speed function $S(\delta_i^t)$ depicted in Figure 2-I is used (the units are meters m for distance and seconds s for time):

$$\begin{aligned} &S(\delta) = 0 \quad \text{if } \delta \le 5 \\ &S(\delta) = 33\tfrac{1}{3} - \tfrac{33\frac{1}{3}}{(100-5)^5} \cdot (100-\delta)^5 \quad \text{if } 5 < \delta \le 100 \\ &S(\delta) = 33\tfrac{1}{3} \quad \text{otherwise} \end{aligned} \tag{2}$$

Equation (2) assumes that the speed falls to zero when the distance between cars is 5 meter (approximately the length of a car), while the maximum free-flow speed of $33\frac{1}{3}$ m/s (120 km/hr) is obtained if $\delta \geq 100$ meters (or when no leader is present). For intermediate values of δ, an arbitrary polynomial function is used, for which $S(\delta)$ is continuous at δ=5 and δ=100, and smooth at δ=100. The implied speed-flow curve for *constant speed* stationary states[4] (for a single lane) is shown in Figure 2-II (flow is calculated as F=$S(\delta)/\delta$). The maximum flow of 0.965 veh./s occurs at a speed of 17.55 m/s (= 63.18 km/h) and a distance of 18.195 meters (a density of 0.055 veh./m). This maximum flow is appreciably higher than the usually empirically measured maxima of 2000 – 2500 vehicles per hour per lane (*e.g.* Small 1992, Figure 3.4), but this deviation is not expected to affect the qualitative properties of the model.



Source: Verhoef (2001a)

Figure 2. The distance-speed function (I) and the implied speed-flow function for stationary states (II) for the numerical simulation model

[4] A stationary state for the dynamic model is defined as a situation where the flow is constant over time for every point along the road. This implies that the flow must be constant along the road, and that S and hence δ at a certain location does not change over subsequent driver. Verhoef (2001a) demonstrates that speed and density (or its inverse, δ) need not be constant *along* the road in a stationary state; *i.e.*, acceleration or deceleration during trips is possible in a stationary state. This will for instance occur when the stationary state involves a vertical queue before the road's entrance, so that drivers start their trips at a zero speed. Interestingly, the standard definitional relation F=S·D or its equivalent F=S/δ do not apply when a stationary state involves acceleration or deceleration (Verhoef, 2001a). The stationary states shown in Figure 2-II, however, are constant speed stationary states.

Two more assumptions ought to be made explicit, involving the road's exit and its entrance. First, near the exit of the road, a driver's speed could not be determined if his leader were assumed to vanish into thin air once completing his trip (δ is no longer defined). To avoid kinks in individual drivers' 'clock-time speed' functions – which could easily become a strong driving force in the model if they cause upstream moving shock-waves – drivers' speeds near the exit are calculated as if all preceding drivers continue their trips after the exit, on a road with the same capacity as the (final segment of) the road modelled. Second, if upon arriving at the road's entrance, a driver's leader has not yet travelled the minimum distance required for a positive speed, it is assumed that a vertical queue with FIFO queuing discipline arises before the road's entrance (this assumption is relevant for Section 3 below).

3. The dynamic instability of hypercongested equilibria for stationary state traffic

Verhoef (2001a) studied what will be called the 'dynamic stability' of the stationary state candidate equilibria found along the AC-curve in Figure 1. Dynamic stability, in this context, is defined as follows: can the 'target' stationary state equilibrium considered, denoted with superscripts 1 and defined by $\{F^1 \equiv S^1 \cdot D^1,\ S^1 = vot \cdot X / AC^1,\ D^1 = 1/\delta^1\}$, be reached or approached asymptotically from any other initial stationary state (denoted with superscripts 0), different from the target state 1, following any possible change in the departure rate (ρ) of new users? If the answer is yes, the target stationary state is said to be dynamically stable; otherwise it is classified dynamically unstable.

Note that the question framed above implicitly assumes that both the normally and hypercongested equilibria can indeed sustain as stationary states, when disregarding dynamic stability, in the present car-following model. For completeness, this is proven to be true in Verhoef (2001a). Any initial stationary state 0 along the AC-curve, hypercongested or non-hypercongested, will sustain (without queuing before the road's entrance) if and only if $\rho^0 = F^0$ continues to hold true following the initial state.

Knowing that any target stationary state 1 can only be consistent with a constant departure rate ρ^1, the stability analysis was simplified to a procedure of starting with an initial stationary state 0, setting the arrival rate ρ^1 equal to the target flow F^1, and proving which stationary state, if any, would be approached (asymptotically). As we will see, the results make it highly implausible that target stationary states that are classified dynamically unstable according to this type of stability analysis would appear dynamically stable when allowing ρ to follow a more complicated pattern following the initial stationary state (*e.g.*, when over- or undershooting F^1 temporarily). However, this is not proven formally in Verhoef (2001a).

When we thus start with an initial stationary state 0 and next study the consequences of a structural change in the arrival rate of new users to $\rho^1 \neq \rho^0$, the following stationary states or in some cases steady states (where a stationary state prevails on the road but a vertical queue before the entrance is growing at a constant rate) can then be shown to be approached asymptotically (proofs are provided in Verhoef, 2001a):

1. *State 0 is non-hypercongested,* $\rho^1 < \rho^0 < cap$: the new equilibrium is the *non-hypercongested* stationary state with $F^1 = \rho^1$.
2. *State 0 is non-hypercongested,* $\rho^0 < \rho^1 < cap$: the new equilibrium is the *non-hypercongested* stationary state with $F^1 = \rho^1$.
3. *State 0 is hypercongested,* $\rho^1 < \rho^0 < cap$: the new equilibrium is the *non-hypercongested* stationary state with $F^1 = \rho^1$. Hence: hypercongestion will dissolve after a reduction in the arrival rate of new users at the entrance. An intuitive explanation is that for a move to the hypercongested stationary state with $F^1 = \rho^1$, the speed would be required to *decrease* after a reduction in the arrival rate, which would require the distances between cars to *decrease* below δ^0. However, the reduction in the arrival rate means that distances will instead

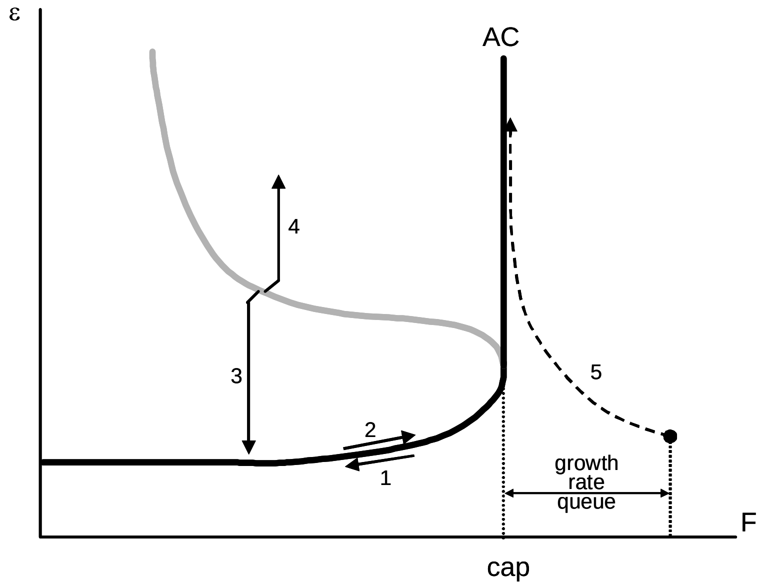
increase above δ^0 (for the first drivers especially near the entrance, but finally along the entire road), so that speeds will *increase*.

4. *State 0 is hypercongested,* $\rho^0 < \rho^1 < cap$: the new equilibrium is a steady state in which a queue is growing before the road's entrance at a rate equal to $\rho^1 - \rho^0$, and a flow $F^0 = \rho^0$ will sustain on the road as a *hypercongested stationary state* – which, however, does not involve constant hypercongested speeds S^0 as in the initial state, but where instead each driver accelerates throughout his trip once having left the queue, and approaches the hypercongested speed S^0 asymptotically from below. For a move to the hypercongested stationary state with a flow $F^1 = \rho^1$, the speed on the road would now be required to *increase* after an increase in the arrival rate, which would require the distances between cars to *increase* above δ^0. However, the increase in the arrival rate means that these distances will instead *decrease* below δ^0; in particular near the entrance. Speeds near the entrance will therefore *decrease* (but because every individual driver accelerates, the original flow level F^0 will sustain; compare footnote 4) and a queue will be formed before the road's entrance.
5. *State 0 is non-hypercongested,* $\rho^0 < cap < \rho^1$: the new equilibrium is a steady state with a flow $F^1 = cap$ on the road, and in addition a queue will be growing before the road's entrance at a rate equal to $\rho^1 - cap$.

These 5 points together imply that there is no transitional path that would lead to any of the hypercongested equilibria suggested by the conventional model. The candidate equilibria on the upper branch of the AC-curve in Figure 1 can therefore be ignored for market analyses, as these are dynamically unstable. Instead, the results suggest that the true 'supply curve' coincides with the lower branch of the AC-curve for non-hypercongested speeds, and would at *cap* not bend backwards, but instead rise vertically (representing queuing costs). This is illustrated in Figure 3, in which the arrows correspond to the five foregoing results, and the bold curve gives the 'true' AC-curve, henceforth defined as the curve showing average travel time costs as a function of traffic flow in dynamically stable stationary or steady states only.

Only one dynamically stable candidate equilibrium for an inverse demand function such as E in Figure 1 (not redrawn in Figure 3) remains, namely the non-hypercongested one. And for an inverse demand function such as E' in Figure 1, the resulting equilibrium would be a stationary state equilibrium with a queue of constant length before the road's entrance, implying constant queuing costs equal to the vertical distance between $AC(cap)$ and $E'(cap)$ in Figure 1, and with an arrival rate ρ equal to the flow on the road, $F = cap$ – after a transitional phase as described under point 5 has taken place. (With elastic demand, a queue will not be growing indefinitely long, but the arrival rate of new users at the entrance will fall as the total trip cost including queuing cost rises during the transitional phase).

The dynamic extension of the standard static model just presented thus takes away much of the ambiguity present in the standard exposition as in Figure 1. For an inverse demand function that intersects the upward sloping branch of the AC-curve, the analysis shows that that this (unique) intersection will be the unique, stable market equilibrium. For an inverse demand function that does not have such an intersection, a quite different market equilibrium involving queuing is identified (providing queuing before the entrance is physically possible – otherwise, no steady state equilibrium will exist), in which the flow on the road obtains the maximum possible value. This equilibrium is not even suggested by the standard model (compare Figure 1). Also this equilibrium is unique and stable. For either case, no hypercongestion will occur on the road.



Note: For arrows 1 – 3, the tail corresponds to $\rho^0=F^0$ and the head to $\rho^1=F^1$.
For arrow 4, the tail corresponds to $\rho^0=F^0$ and the head to ρ^1 (but note that $F^1=F^0<\rho^1$ will hold).
For arrow 5, the tail corresponds to ρ^1 and the head to $F^1=cap$.

Figure 3. Schematic overview of the dynamic stability analysis for the conventional AC-curve, and the resulting 'true' AC-curve for dynamically stable stationary or steady states (in bold)

One might object that hypercongestion is often observed in reality, implying that the above analysis must be wrong. However, we have not yet explored whether traffic in the queue, before the road's entrance, would evolve under hypercongested conditions, once the up to now assumed vertical queuing technology would be replaced by car-following behaviour similar to what is up to now considered to apply on the road itself. This issue was investigated in Verhoef (2002), to which we turn now.

4. Peak congestion and a bottleneck: dynamic equilibria

Verhoef (2003) extended the above modelling framework into two directions. First, in order to study the question of whether hypercongestion can be generated in a model with the proposed car-following congestion technology once a downstream bottleneck exists, the extension considers the situation where the single road has a bottleneck halfway, comparable to the set-up of Mun (1999). This bottleneck is caused by a reduction in the number of lanes (from two to one; see Figure 4), where traffic on each lane behaves according to the same distance-speed relation as assumed for the constant-capacity (single lane) model described above. To prevent discrete changes in speed, it is assumed that traffic merging takes place gradually, over a road segment of positive length between two locations indicated with x_1 and x_2. Between these two points, a given user i to an increasing extent considers another diver i–1 (always coming from the different lane in a dynamic equilibrium; see Verhoef, 2003, for a proof) as his relevant leader, instead of the 'upstream leader' i–2. This smooth process was represented in the numerical model by redefining a driver's δ_i^t (for use in (1) and (2); t denotes time) as follows:

$$\delta_i^t = \begin{cases} x_{i-2}^t - x_i^t & \text{if } x_{i-1}^t < x_1 \\ w(t)\cdot\left(x_{i-2}^t - x_i^t\right) + (1-w(t))\cdot\left(x_{i-1}^t - x_i^t\right) & \text{if } x_1 \le x_{i-1}^t \le x_2 \\ x_{i-1}^t - x_i^t & \text{if } x_{i-1}^t > x_2 \end{cases} \tag{3}$$

$$\text{with: } \quad w(t) = 1 + 2\cdot\left(\frac{t - t_{i-1,1}}{t_{i-1,2} - t_{i-1,1}}\right)^3 - 3\cdot\left(\frac{t - t_{i-1,1}}{t_{i-1,2} - t_{i-1,1}}\right)^2$$

The function *w* thus defines the weights attached to drivers *i*–2 and *i*–1, which sum up to unity. An (otherwise arbitrary) functional specification for *w* was chosen that secures that the weight for driver *i*–2 falls continuously over time from 1 to 0 as driver *i*–1 proceeds from x_1 to x_2, and that *w* has a zero time derivative at the instants driver *i*–1 passes x_1 and x_2.

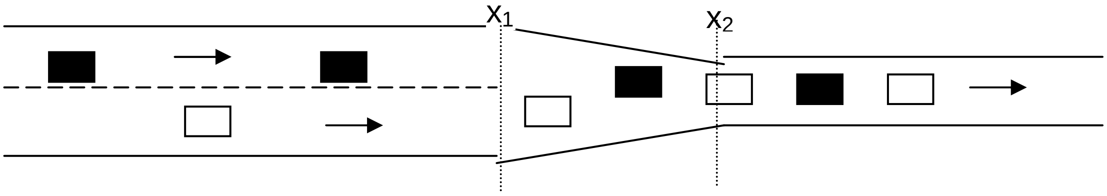


Figure 4. The spatial lay-out of the road network and the implied bottleneck

Second, to investigate whether hypercongestion, if generated, can occur as a (dynamic) equilibrium phenomenon, the demand side of the model is adjusted. Specifically, the analysis in Section 3 assumed inelastic ‘continuous’ demand (a given constant departure rate), which was changed exogenously to investigate dynamic stability of stationary state equilibria. Verhoef (2003) considers inelastic ‘peak’ demand: a given number of users want to use the road near some desired arrival time (elastic overall demand could easily be introduced, but was ignored to maintain comparability with prior dynamic economic equilibrium models, which typically assume inelastic overall demand). The departure times and hence the overall duration of the peak are endogenized, so that instantaneous demand is not inelastic. A dynamic equilibrium is obtained if, as in Vickrey (1969), the sum of travel time costs, schedule delay costs and tolls (if levied) is constant over time if travelling occurs, and higher otherwise (no user can improve total trip costs by choosing a different departure time; note the intuitive parallel with Wardrop’s (1952) first principle). As customary in this literature, the schedule delay cost function was assumed to be piecewise linear for the numerical model, with slopes $-\beta$ for early arrivals and γ for late arrivals.

A main question studied with the model is whether hypercongestion can and/or will be generated as a dynamic equilibrium phenomenon by the model.[5] The answer is yes. More precisely, provided demand is large enough relative to the capacity of the downstream (single lane) road segment, so that the peak lasts sufficiently long, it can be inferred that hypercongestion will arise with certainty. However, hypercongestion will arise on the upstream (large capacity) road segment only. On the downstream (small capacity) road segment, in contrast, the flow will asymptotically approach the maximum flow (for that segment), cap^{ds}, and speeds will be constant.

If we compare this downstream segment with the single road in the previous section, the latter finding is consistent with result 5 from the stability analysis. Indeed, the departure rate will exceed cap^{ds} from some moment onwards. This is needed to maintain the equilibrium

[5] The other main question concerns optimal pricing, which will not be addressed here.

condition that travel times should increase over time, to match for the falling schedule delay costs for arrivals before the desired arrival time. For the upstream segment, with a capacity $cap^{us} = 2{\cdot}cap^{ds}$, the fact that the inflow exceeds cap^{ds} whereas the outflow asymptotically approaches cap^{ds} (from below) implies that the average distances and hence speeds between vehicles must be falling over time. This will continue until the inflow into the upstream segment, as well as the flow at each point along that segment, would become equal to cap^{ds}, too. But with a departure rate exceeding cap^{ds}, this could occur only if a queue would be formed before the upstream segment's entrance. And as demonstrated in Verhoef (2001), a stationary state on the upstream segment with a flow (cap^{ds}) below capacity ($2{\cdot}cap^{ds}$) and with queuing before the road's entrance can only exist if that stationary state is hypercongested.[6] In other words: speeds will be falling on the upstream road segment until a hypercongested state is reached, provided demand is sufficiently large and hence the peak lasts sufficiently long.

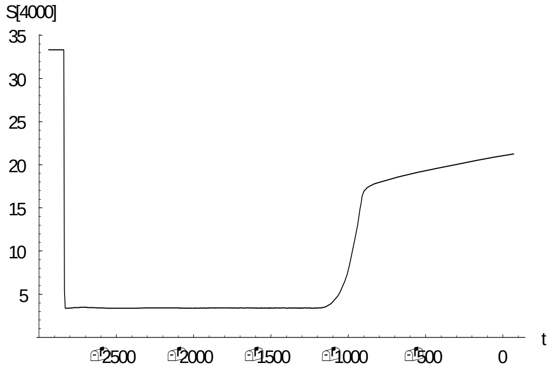


Figure 5. The clock-time speed function for driver i=4000 in the numerical simulation model for peak demand and with a bottleneck halfway the road

Figure 5 illustrates these results, by showing the clock-time speed function for driver i=4000 in the numerical simulation model for peak demand with the number of users N=5000, and with a bottleneck halfway the road (details of the model and its equilibrium are given in Verhoef, 2003). The speed on most of the upstream segment is indeed (near) the hypercongested speed consistent with $½{\cdot}cap$ according to Figure 2-II (the high speed over the first meters is caused by the fact that driver 4000 arrives slightly after the desired arrival time, so that hypercongestion is in fact already dissolving), while the speed on the downstream segment is near the value consistent with cap according to that same Figure 2-II.

In summary, the homogeneous-road model from Section 3 showed that hypercongested equilibria are dynamically unstable and can therefore be ignored in market analyses, while the 'bottleneck-road' model from this section predicts that hypercongestion will occur with certainty as a dynamic equilibrium phenomenon on the upstream road segment. These conclusions are by no means incompatible, since the homogeneous road in fact corresponds with the downstream segment of the bottleneck road in the present section (both have no downstream bottleneck), which too will not have hypercongestion. Likewise, the upstream segment of the bottleneck-road corresponds with the vertical queue in the homogeneous-road model (both have a downstream bottleneck). The bottleneck-road model thus show that queuing will take place under hypercongested conditions; and more precisely: at the hypercongested speed that is (for the upstream segment) consistent with a flow equal to cap^{ds}.

[6] An easy way to see this is to observe that when constructing an initial state with a queue and a non-hypercongested stationary state with a flow below capacity, regardless of the arrival rate at the back of the queue, the inflow into the road will immediately increase to a value above the assumed initial non-hypercongested flow on the road.

Which are the implications of the present bottleneck-road set-up for the shape of the AC-curve for the bottleneck-road, compared to that shown in Figure 3? Recall that we defined the AC-curve as representing possible stationary or steady state equilibria only. These exist only for downstream flows up to a level cap^{ds}. At that level, the AC-curve will not bend backwards, but again rises vertically, as in Figure 3. For this vertical branch, the only difference with the homogeneous-road model is that the vertical distance between the minimum possible $AC(cap^{ds})$ and any equilibrium value of $AC(cap^{ds})$ (at the intersection with an inverse demand defined over flow in a stationary state) would not reflect waiting costs incurred in a vertical queue. Now it reflects travel costs on the upstream segment above the minimum possible upstream costs for a flow cap^{ds} – a part of which will generally be travelled under hypercongested conditions. Interestingly, therefore, although a part of the trip will be driven under hypercongested conditions, the AC-curve is still not backward-bending in those equilibria – in sharp contrast with what might be inferred from the conventional analysis. This reflects that, starting in such a stationary state equilibrium and raising the departure rate ρ above cap^{ds} would not cause the flow on the downstream segment to drop, but would only lead to larger (and growing) travel times on the upstream segment, and a larger and growing portion of that segment to be travelled at a hypercongested speed.

The shape of the AC-curve for stationary equilibrium flows below cap^{ds} will also be comparable to that shown in Figure 3. However, the relation with the shape of the speed-flow function will be slightly more complicated than in Figure 3. The total travel time would then namely no longer correspond to a single speed, because the equilibrium speed will be higher on the upstream than on the downstream segment. The AC-curve for $F<cap^{ds}$ therefore reflects the average, non-constant speed over the full trip (as well as the length of the full road and the value of time).

An AC-curve as defined above is of theoretical interest, because of the fundamental difference with the conventional AC-curve of Figure 1 which also considers stationary traffic conditions only. It is however not of direct use for an empirical testing of the model's predictions, because stationary state conditions will seldom apply in reality, and are particularly unlikely for the range of most interest: the vertical branch of the AC-curve. The question rises whether a curve similar to the AC-curve would be predicted by the model also for non-stationary traffic conditions. As we will see in the next section, this appears to be the case if some mild restrictions are placed on the time pattern of the departure rate that causes traffic conditions to be non-stationary – restrictions which are consistent with standard dynamic equilibrium departure patterns as also applying in Verhoef (2002). Another shortcoming of the AC-curve when used in the context of dynamic equilibrium models is that it does not reflect schedule delay costs and – related to this – is not directly related to dynamic equilibrium cost levels. This second shortcoming will be addressed in Section 6, when addressing 'reduced-form' cost functions for dynamic equilibrium models.

5. Comparing the predictions with observed traffic data

5.1. Formulating testable hypotheses

Because stationary traffic conditions seldom apply in reality, especially not for the vertical branch of the proposed AC-curve, it is important to formulate testable hypotheses based on the above models that go beyond these models' behaviour in stationary traffic conditions. Because one could make many assumptions on possible time-varying patterns of departure rates that would give rise to non-stationary traffic conditions, it is useful to restrict attention to the typical case of traffic congestion and queuing. This involve the situation where traffic is initially relatively light and the departure rate ρ^D has a level well below cap^{ds}, then rises monotonically to a level well above cap^{ds}, stays above cap^{ds} sufficiently long, and next falls again monotonically to a level well below cap^{ds}. For the resulting non-stationary traffic conditions, the models just discussed would then predict that the following stylized facts will be observed on a road that has a higher upstream than downstream capacity:

H1. The arrival rate (or outflow of the road) ρ^A (at the exit of the downstream segment) will approach the downstream capacity cap^{ds} and stay close to it at least as long ρ^D has not yet fallen below cap^{ds}. As a result of the predicted induced time-varying queuing on the upstream segment, a parametric plot of travel times $tt(t^A)$ (t^A denotes the arrival time) as a function of $\rho^A(t^A)$ will have a shape similar (but not identical) to $AC(F)$ in Figure 3. In particular, $tt(\rho^A)$ – which will be the short-hand notation for this parametric plot – will not bend backwards, but will rise vertically at cap^{ds}.[7]

H2. Along the downstream road segment, traffic flow will approach its capacity cap^{ds}, at the constant speed consistent with this flow, at least as long ρ^D has not yet fallen below cap^{ds}. Observations of combinations of instantaneous speed and flow levels along the downstream road segment will therefore include only non-hypercongested situations, and observations up to the flow equal to cap^{ds}.

H3. On the upstream road segment, queuing before the downstream segment's entrance will take the form of driving under hypercongested conditions, at a flow equal to cap^{ds} (compare Section 4). Observations of combinations of instantaneous speed and flow levels along the upstream road segment will therefore cover both hypercongested and non-hypercongested situations, and do not necessarily include observations up to the flow equal to the upstream capacity cap^{us}.

In this section, these hypotheses will be verified on the basis of a simulation using the model described in Section 4 above, and on the basis of empirical observations. Sections 5.2 and 5.3 first provide the main backgrounds of both sources.

5.2. The simulation model

The supply side of the simulation model is fully defined by equations (2) and (3), with the remaining parameters set as follows: x_1 = 9 000 (the location of the beginning of the bottleneck in Figure 4), x_2 = 11 000 (the ending of the bottleneck), and X = 20 000 (the length of the road). The locations used for reporting upstream and downstream instantaneous traffic conditions are denoted x_U = 8 000 and x_D = 12 000, respectively.

Instead of using the equilibrium departure rate obtained under the piece-wise linear schedule delay costs used in Verhoef (2003), the exogenous time-pattern depicted in Figure 7 (upper panel) was used. This would be an equilibrium pattern when the schedule delay cost function had a shape given by $K - vot{\cdot}tt(t^A)$, with $tt(t^A)$ as depicted in Figure 8 (upper panel) and K a constant. The reason for using this departure rate pattern was to increase comparability with the shape of that observed in the empirical data (Figure 7, lower panel). Departures occur over a time frame of 4 000 seconds, during which 3506 users make a trip.

5.3. The empirical case study

For this study, empirical data were available for a study area north of Amsterdam, depicted in Figure 6. Unfortunately, the network is more complex than that used in simulation model. As in the simulation model, there is only one downstream segment: the southbound western part of the A10 (the Amsterdam ring road). There is however a 'composite' daily

[7] Whereas for a given 'bottleneck-road' with a given car-following equation, $AC(F)$ will be uniquely defined, $tt(\rho^A)$ will not be. Due to the path-dependency inherent in car-following models, the exact travel time that a driver arriving when the arrival rate equals ρ^A will have incurred will depend on the initial state and the full history of departure rates before his departure. But under the single-peaked departure rate patterns considered, the general shape of $tt(\rho^A)$ will always resemble that of $AC(F)$.

traffic jam of drivers who want to enter this segment, located on the northern segment of the A10 (westbound) and on the A8 (southbound), with an additional spill-back to the A7 (southbound) when congestion reaches its peak.

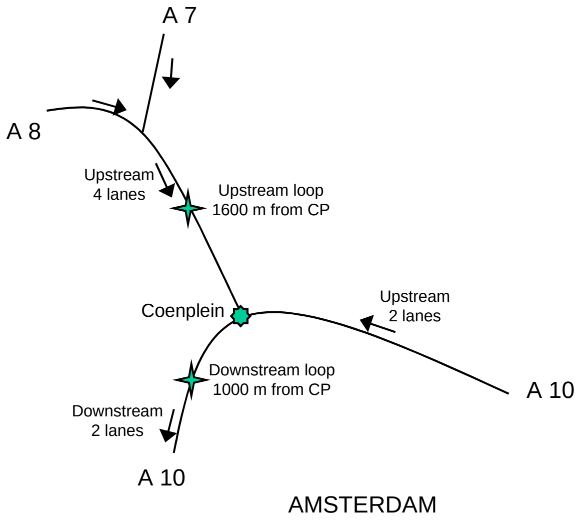


Figure 6. The study area for the empirical case study

The traffic data used for this analysis consist of data which were derived from loop detector data, and that were pre-processed by a group of engineers at a Dutch consultancy bureau OC&C (the original data were unfortunately not available for the present research). They have produced figures that are averaged over all ‘normal’ working days (*i.e.*, excluding holiday periods) in the year 2000. This has the advantage of giving a rather general picture by reducing dependency on day-specific circumstances, but – as we will see shortly – the disadvantage of introducing some conceptual problems due to the procedure of averaging over working days. The data consist of clock-time dependent measures (averaged for 15 minutes intervals) for:

1. The inflow into the network studied aiming for the downstream link considered ($\rho^D(t)$; see Figure 7, lower panel), aggregated over the three upstream links, and measured at loop detectors that were chosen such that these are upstream of the queue’s tail even when it reaches its greatest length. This inflow was derived from total flow at these loop detectors by correcting for the share of traffic that, once arrived at the Coenplein, does not choose the downstream link considered (the A10 southbound) but instead goes east on the A10.
2. The outflow from the network ($\rho^A(t)$; see Figure 7, lower panel), measured at a detector some 2 km south of the Coenplein.
3. Traffic flow ($F(t)$) at an upstream (U) and at a downstream (D) detector (both indicated in Figure 6); see Figure 9, lower panel.
4. Traffic speed ($S(t)$) at the same upstream and a downstream detectors; see Figure 9, lower panel.

The data in categories 2–4 above were obtained by a simple averaging of observations over all normal working days. The first variable required additional computations, performed by OC&C, based on a comparison of cumulative flows at the points of measuring $\rho^D(t)$ and at the three exit links of the junction ‘Coenplein’. This allowed for the correction of total flows at the points of entrance that was necessary to identify that part of the flow heading for the downstream link considered.

In what follows, no distinction will be made between travel times experienced in the three ‘sub-queues’ that can be distinguished, for the simple reason that the available data do not allow making this distinction. These travel times (Figure 7, lower panel) were calculated by a comparison between cumulative inflows $\rho^D(t)$ and outflows $\rho^A(t)$, while using that the weighted average of distances between the inflow and the outflow detectors is 11 km (5.5 minutes at a free-flow travel speed of 120 km/h, as applying in the beginning of the peak). As a result, the travel times reported are a weighted average for the three sub-queues. This simplification does however not affect the conclusions with respect to the issues of central interest in this paper; in particular the general shape of the $tt(\rho^A)$ function.

A final issue is whether in the case study, it is indeed the downstream link (on the A10) that is the bottleneck, as in the simulation model. It could of course be the case that the ‘true’ bottleneck is the junction ‘Coenplein’ itself. The evidence in favour of the downstream link being the true bottleneck is that from its observed speed-flow function in Figure 9 (lower panel), one can infer that the Coenplein does not seem to limit its inflow below its capacity. That is: a maximum flow of some 2150 vehicles per lane per hour is observed, at a speed of around 65 km/hr. This corresponds to capacities and corresponding speeds found elsewhere (*e.g.* Small, 1992; and Smith, Hall and Montgomery, 1996).

5.4. Comparing the results

Figure 7 shows the time patterns of ρ^D, ρ^A, and $tt(t^A)$ for the simulation model (upper panel) and the empirical case study (lower panel). As explained in Section 5.2, the exogenous departure rate pattern for the simulation model (upper panel) was set such that it is single-peaked, with a maximum well above the downstream segment’s capacity cap^{ds}. As a result, traffic conditions follow the same pattern as in Verhoef (2002) and described in detail in Section 4 above. Note in particular that the arrival rate ρ^A approaches cap^{ds} (= 0.965).[8] The clock-time speed functions for drivers arriving near t = 3500 (not shown graphically) are also similar to what is shown in Figure 5, and hypercongested queuing occurs on the upstream link (see also Figure 9 below). Finally, travel times as a function of arrival time, $tt(t^A)$, show the expected single-peaked pattern.

The empirical patterns in the lower panel of Figure 7 are roughly similar. (A conceptually minor difference with the simulation model is that after the peak, demand does not fall to zero.) Also here, ρ^D is single-peaked, exceeds ρ^A during the beginning of the peak and next falls below it. The latter remains flat for a significant part of the peak. The joint result is that $tt(t^A)$ has the expected single-peaked shape. This function $tt(t^A)$ was determined by assuming – consistent with the observations shown in Figure 9 – that the first driver that can be tracked (departing at 5:07:30) travels the 11 km considered at a free-flow speed of 120 km/hr. His arrival time can then be determined as 5:13:00. The travel times for subsequent drivers were next calculated as the difference between the clock times at which cumulative inflows (after 5:07:30) and outflows (after 5:13:00) have reached the same value (the required continuous versions of both $\rho^D(t)$ and $\rho^A(t)$ were constructed by non-linear interpolation of the 24 observations available for both between 5:00 and 11:00). The last step was to assign these travel times to their instant of arrival, t^A.

[8] The arrival rate ρ^A stays close to cap^{ds} nearly to the end of the simulation, and then suddenly drops to a level approximately equal to the corresponding departure rate ρ^D. This is the result of a feature that was called ‘footloose queuing’, and discussed in greater depth, in Verhoef (2002): the drop in speeds that drivers experience due to the existence of the bottleneck dissolves while propagating downstream when ρ^D falls below cap^{ds}.

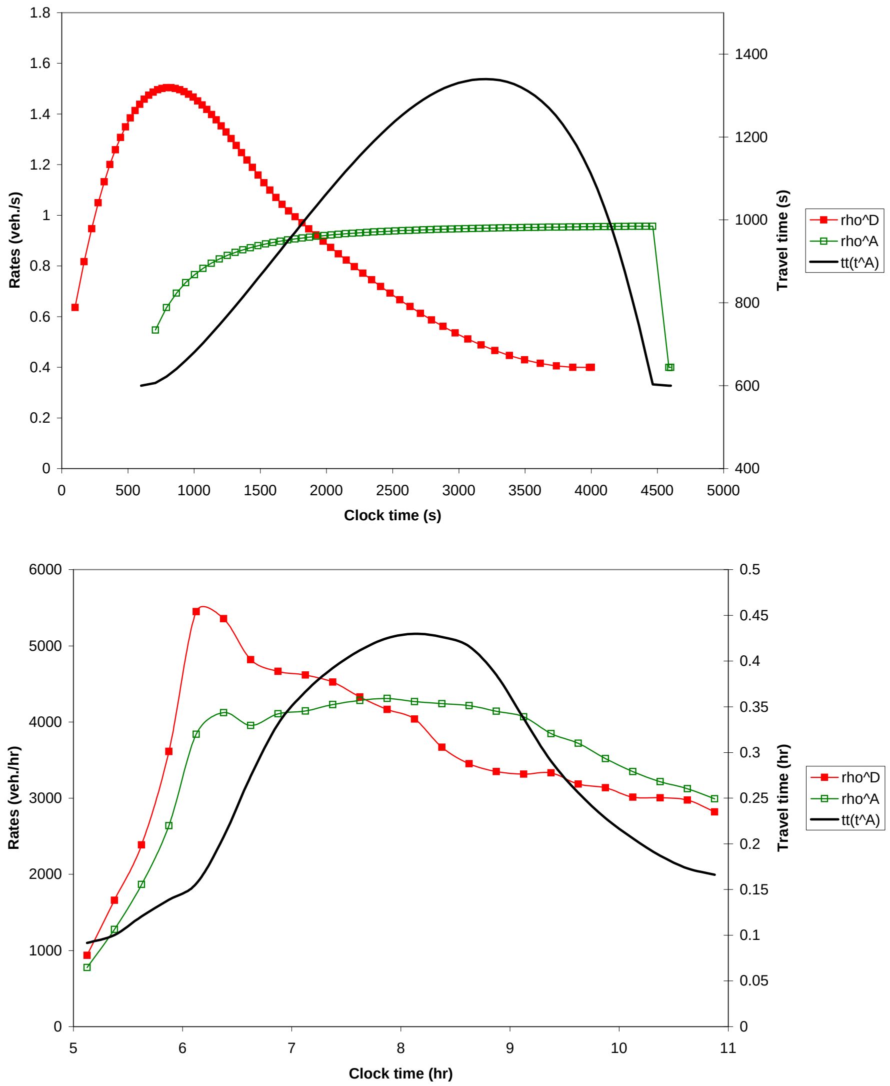


Figure 7. Inflows, outflows and travel times for the simulation model (upper panel) and the empirical case study (lower panel)

The $tt(\rho^A)$ functions shown in Figure 8 follow as the parametric plot of $tt(t^A)$ and $\rho^A(t^A)$ from Figure 7. A similar parametric plot can be constructed for departures, by combining $tt(t^D)$ and $\rho^D(t^D)$ (with t^D denoting departure time) – where $tt(t^D)$ is not included in Figure 7. For ease of reference, the resulting plots $tt(\rho^D)$ are also included in Figure 8.

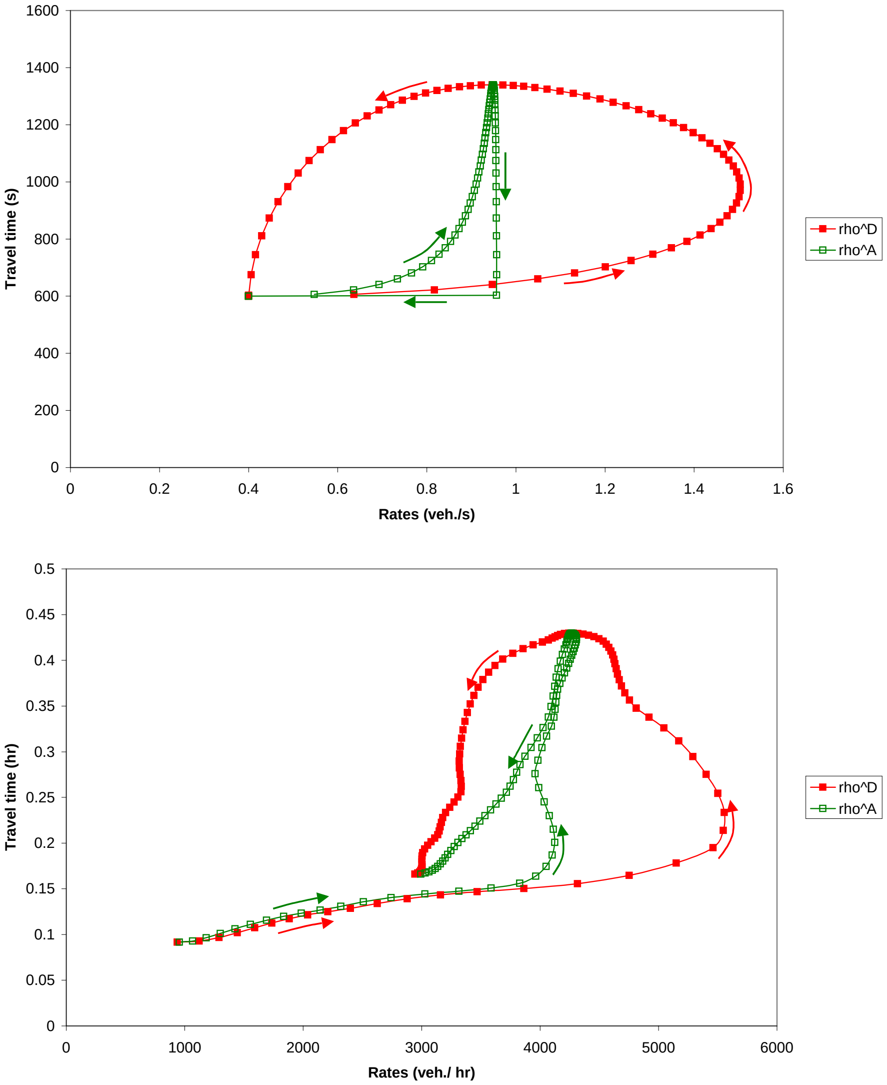


Figure 8. Travel times by departure and arrival rate for the simulation model (upper panel) and the empirical case study (lower panel); arrows indicate development over time

The patterns for the simulation model and the empirical case study are again similar. That is, as predicted in Hypothesis H1 in Section 5.1, $tt(\rho^A)$ rises nearly vertically when ρ^A approaches a certain level. For the simulation model, this is no surprise, because it is fully consistent with the theoretical model from which the hypothesis was derived. It is, in addition, easily checked for the simulation model that the 'certain level' of ρ^A for which $tt(\rho^A)$ appears to have a vertical asymptote corresponds to cap^{ds} (0.965 veh./s). The lower panel of Figure 8 shows that a similar patterns holds for the empirical case study for arrival rates around ρ^A = 4300. As mentioned, for a two-lane highway this rate corresponds to generally accepted capacities. A conservative conclusion would thus be that the empirical observations do not reject H1.

An important difference between the simulation results and the empirical observations arises towards the end of the period considered. The arrival rate for the numerical model keeps growing over time nearly up to the last drivers arriving, also when the travel time is already falling over time. For the empirical data, the arrival rate already appears to fall significantly when the travel time has come down to some 0.3 hours (18 minutes), a duration for which the results for the early part of the peak suggest that queued traffic should still occur – if only because the implied average speed over the 11 km considered is still only 37 km/hr.

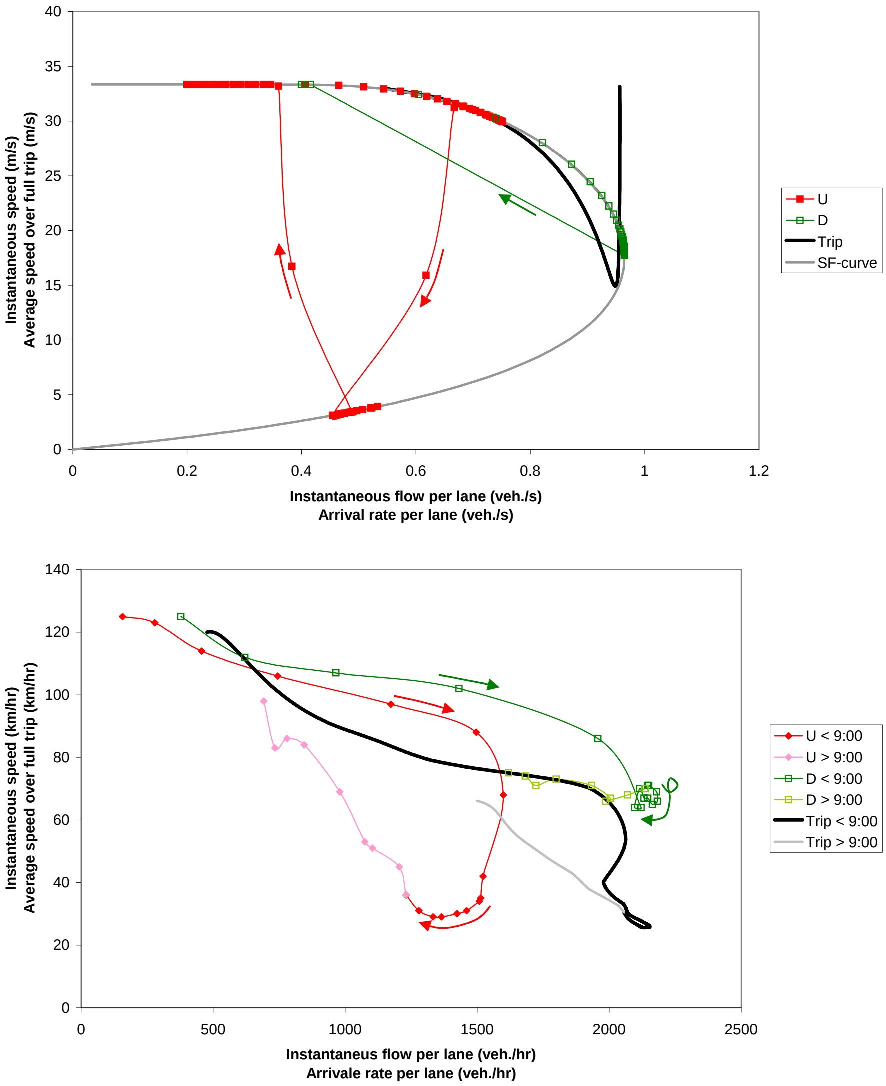


Figure 9. Observed speed-flow combinations on upstream (U) and downstream (D) road segments, and average speeds over full trips by arrival rate, for the simulation model (upper panel) and the empirical case study (lower panel); arrows roughly indicate development over time

The most plausible explanation for this unexpected result is that the observations over this part of the peak provide weighted average (over working days) arrival rates and travel times that represent two distinctly different situations, namely one in which the queue still exists and the arrival rate still equals around 4000 veh./hr or more, and one in which the queue has already dissolved and the arrival rate has dropped to, for example, 3000 veh./hr. For later observations (when moving south-west along the $tt(\rho^A)$ curve, the latter state receives an increasing weight, which explains the shape of the $tt(\rho^A)$ curve over this range. It is especially here that the nature of the data (averaged over all ‘normal’ working days) may give a biased representation of what will occur during a single morning peak. This is due to the sharp drop in the arrival rate once the queue has dissolved (compare also the upper panel of Figure 8). For this reason, the empirical observations pertaining to clock times after around 9:00 should be treated with suspicion, and will therefore be printed in a lighter shade in the final speed-flow diagrams in Figure 9, for which these observation become particularly unusual.

Verification of Hypotheses H2 and H3 requires consultation of speed-flow diagrams. Figure 9 shows these for the upstream (*U*) and downstream (*D*) observation points, with flow normalized with respect to the number of traffic lanes, and again with the simulation results in the upper panel and the empirical data in the lower one. For the former, the 'true' speed-flow function is known (see Figure 2-II), and is added in grey. Note that upstream traffic flow in the empirical data also includes vehicles that are heading for other downstream links than the one considered above: the flow data in Figure 9 represent total flow at the observation points.

H2 refers to speed-flow observations on the downstream road segment. Both the simulation results and the empirical data confirm the hypothesis. No observations of hypercongestion occur on these segments in either panel. Next, for the simulation model, it is easily verified that the flow indeed approaches cap^{ds} (= 0.965 veh./s), and the speed the level associated with this maximum flow (= 17.5 m/s). The relatively large number of observations near this point shows that these conditions apply over a relatively long time span – after which the flow drops rapidly and the speed rises rapidly. For the empirical data, no objective measure of capacity is available. However, the shape of the speed-flow function near the large concentration of observations around a flow of around 2150 veh./hr and a speed of around 65 – 70 km/hr suggest that this is indeed near the road's capacity. As mentioned, these values also correspond closely to prior evidence from other sources.

Note that the final observations (after 9:00) for the empirical speed-flow function, printed in a lighter shade, seem to fall off the speed-flow function suggested by the earlier observations. This is consistent with the earlier hypothesis that these observations are biased due to the averaging over working days. This hypothesis is supported by the rather close correspondence between the pattern for these observations and the interpolating line, for the simulation results, between the final observation near capacity and the subsequent first observation that is well below capacity.

Hypothesis H3 refers to observations on the upstream segment. The occurrence of both hypercongested and non-hypercongested observations on these segments is easily verified. Both diagrams show how the flow as measured at the observation points *U* first increases over the non-hypercongested range of the speed-flow function. Then the speed drops rapidly to a hypercongested level. The accompanying fall in flow is consistent with the notion that the flow in the queue will be approximately equal to that on the downstream road segment, while the flow upstream of the queue will exceed that level when the queue is growing. Similarly, when the queue dissolves and the speed rises back to a non-hypercongested level, the observed flow will drop: a shrinking queue requires the flow in the queue to exceed the flow upstream of the queue.

The empirical observations after 9:00 again follow the same pattern as the interpolating line for the simulation model. Again, the most plausible explanation is that these empirical observations are in fact weighted averages of observations on – or near – the upper and lower branches of the speed-flow function for this upper road segment. (Similarly, the two empirical observations on the nearly vertical branch – between the non-hypercongested and hypercongested range – may be largely attributable to time averaging, rather than corresponding to actual traffic conditions). Note, however, that also in the simulation results, there are two observations that clearly fall off the speed-flow function. These observations illustrate the point made in footnote 4, that when cars accelerate or decelerate, flow is not identically equal to the product of speed and density.

H3 further stipulates that the hypercongested flow on the upstream segment be equal to downstream capacity cap^{ds}. For the simulation results, it is clear that this is indeed (approximately) the case: the observations are near a flow of $½·cap^{ds}$ (= 0.48); the variations around this value are due to transitional dynamics. For the empirical data, this claim cannot be verified for two reasons: first, no objective measure for the downstream capacity is available; and secondly, the upstream link feeds into multiple downstream links, so that in fact multiple downstream capacities would have to be considered.

The final statement in H3 says that for the upstream segment, no observations up to its capacity need be found. Again, in the upper panel this is easily seen to be the case. But also the empirical data can be interpreted this way. Specifically, the fact that traffic flows have been normalized with respect to the number of traffic lanes allows a comparison between the suggested speed-flow functions for the upstream and downstream road segment. The non-hypercongested observations up to a flow of 1500 veh./hr suggest that these speed-flow functions may well be nearly the same. As for the simulation results, the upstream and downstream data would then give overlapping non-hypercongested observations up to a certain flow level, while the upstream road segment adds observations from the hypercongested branch, and the downstream road segment higher-flow observations from the non-hypercongested branch. It does not take a lot of imagination to pencil in, in the lower panel of Figure 9, a shared single speed-flow function that applies for both road segments (as shown in grey for the simulation model in the upper panel of Figure 9).

Note that this also offers a simple explanation for the pattern often found in empirical speed-flow scatter plots entailing observations from the hypercongested branch. These often show a rather flat normally congested branch followed by a sharp drop towards the hypercongested range. The fact that hypercongested observations are included means that a downstream bottleneck must be present. But this, in turn, implies that speed-flow observations near the upstream road segment's capacity are unlikely to be made, because a queue will arise even before the normally congested flow reaches capacity. As soon as the queue reaches the observation point, the observed hypercongested flow will be approximately equal to the downstream bottleneck's capacity. The result will be a censored data set, showing a 'blunter' speed-flow curve than what would be found than in absence of the bottleneck.

Finally, in both panels of Figure 9 a thick curve labelled 'trip' is included. This curve shows the average speed over the full trip (*i.e.* between the points of in- and outflow) as a function of the arrival rate per lane at the point of outflow. It is the inverse of the $tt(\rho^A)$ curves shown in Figure 8, corrected for the length of the trip and the number of lanes on the downstream segment. These 'average-speed – arrival-rate' curves do not overlap the instantaneous speed-flow functions, but instead again show how an average speed (over the full trip) in the hypercongested range of speed levels would correspond with an arrival rate close to the downstream road segment's capacity. This difference underlines the pitfalls in deriving average cost functions from speed-flow functions in the conventional way. Moreover, the curves deviate already well before this arrival rate is reached. This is in the first place due to the inclusion of queuing time losses in the average-speed – arrival-rate curve, which begin to occur already before the arrival rate approaches the downstream capacity closely. Secondly, with a non-constant capacity, even for stationary state traffic would the average speed, over full trips, as a function of the arrival rate differ from the instantaneous speed consistent with the corresponding flow.

6. Reduced-form cost functions for dynamic equilibrium models

Up until this point, we have considered cost functions that relate travel costs to traffic flows or to arrival rates; *i.e.*, output measures that are normalized with respect to the time dimension. Another type of cost function arises when we relate travel costs to the total number of users *N*, not normalized with respect to time. It is of interest to briefly address this type of cost function in the context of the proposed car-following model, especially because these functions allow us to also include schedule delay costs – apart from travel time costs – which are an important component of total travel costs for dynamic equilibrium models.

As explained above (and also in greater detail in Verhoef, 1999), cost functions with quantities rather than flows or rates of drivers are not meaningful to consider when modelling stationary state traffic. But for dynamic equilibrium models of traffic congestion, with endogenous departure times and peak duration, it is useful to consider such cost functions, as they show the impact on total travel costs when demand increases. One could call these 'reduced-form cost functions', to reflect that these functions depict dynamic equilibrium cost levels as a function of total demand over the full peak, without directly conveying information on the dynamic patterns of travel time costs and schedule delay costs inside that peak, which underlie the realization of these cost levels.

As shown for example by Arnott, De Palma and Lindsey (1998), the total travel cost (*i.e.*, the sum of travel delay and schedule delay costs over all users) for the simplest version of Vickrey's (1969) bottleneck model is quadratic in the number of users during the full peak, which renders the corresponding average and marginal cost functions linear. Their reduced form character is illustrated by two of their features. One is that different cost functions apply when optimal (or second-best) tolling is in place, compared to the no-toll cost functions (see equations (4.10a) versus (4.12) in Arnott, De Palma and Lindsey, 1998). The other is that unlike what would be the case for a static model with the same linear cost functions, the difference between average and marginal 'reduced-form costs' is not a measure for the optimal toll – which should be time-dependent.

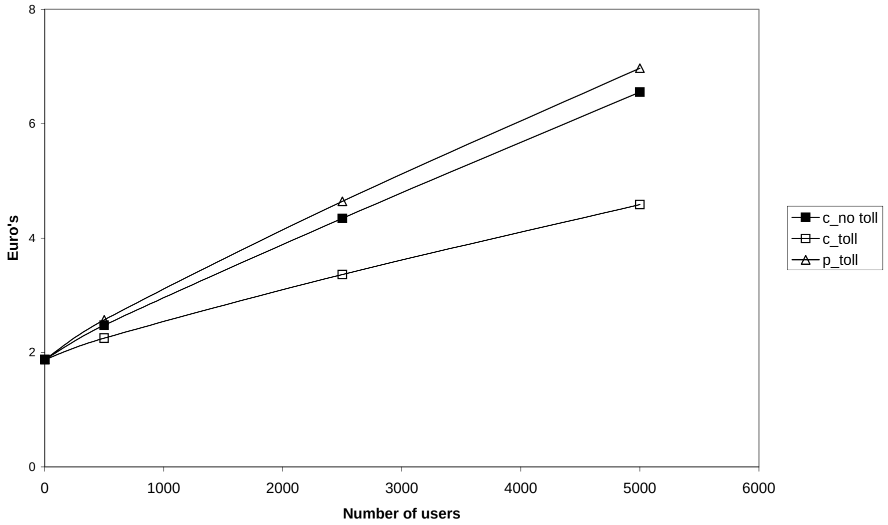


Figure 10. Reduced-form average cost functions for a dynamic equilibrium model based on car-following congestion technology

It is often implicitly or explicitly assumed that average travel costs are a convex, increasing function of the number of users during the peak. Vickrey's (1969) model shows that these functions may also be linear. It is interesting to see what sort of shape is predicted by the car-following model used above. Figure 10 shows three such curves. The sum of average travel time costs and schedule delay costs for no-toll equilibria with varying levels of demand is shown as $c_{no\ toll}$. The corresponding reduced-form average cost function with tolling is given by c_{toll}, while the optimal average trip price p_{toll} includes toll payments. The latter two apply when optimal time-varying tolling is in place (this in fact refers to 'approximately optimal tolling', as explained in detail in Verhoef, 2003). The cost functions shown were derived for the bottleneck road used in Verhoef (2003), which differs from the one used above only in that its total length X equals 30 000 instead of 20 000 meters. Furthermore, the demand-side parameters were defined as follows: the value of time α is equal to € 7.5 (the Dutch average), while the (constant) shadow prices for early and late arrivals were set at β = € 3.75 and γ = € 15, respectively. The curves shown in Figure 10 were constructed by interpolating observations for numbers of drivers $N = \{1, 500, 2500, 5000\}$.

A first noteworthy feature is that, although the cost functions discussed in previous sections (that relate average travel time costs to arrival rates) all had the intuitive convex shape (see Figures 3 and 9), the reduced-form cost functions (that relate average total costs to the total number of users) have a concave shape. This shape therefore deviates even further from the intuitive expectation of convex functions, than do the linear reduced-form cost functions from Vickrey's (1969) bottleneck model.

The intuition behind the concave shape is most easily given for the no-toll average cost function. Consider a dynamic no-toll equilibrium for N users, with N_E users arriving early and N_L=N–N_E users arriving late. Then imagine adding one user. This will make the peak last longer, and to start earlier and end later. But because of the assumed linearity of the schedule delay cost functions, the headways between the first N_E users must remain the same to maintain constancy of costs over these users, although each of them departs earlier. In other words, we can imagine the additional driver as being 'timed' at the desired arrival time, with each of the original N_E users arriving earlier and each of the original N_L users arriving later. The duration of the peak will therefore increase; approximately with the inverse of the arrival rate at the most desired arrival time t^*: $\rho^A(t^*)$. But since $\rho^A(t^*)$ is increasing in N, this increase in peak duration decreases with N. This means that the increase in schedule delay costs for the first and the last driver due to an increase in N decreases with N. Because these two drivers will drive at a free-flow speed, also the induced increase in their travel costs decreases with N. But because these travel costs are equalized over users in a dynamic equilibrium, the same holds for the equilibrium average costs. Hence the concave shape of $c_{no\ toll}$. The concavity of $c_{no\ toll}$ can be expected to vanish as $\rho^A(t^*)$ approaches cap^{ds}, which is why $c_{no\ toll}$ appears to become nearly linear when moving to the right in Figure 10.

From the perspective of this paper, probably the most important conclusion is that a convex speed-flow relation, and an associated convex travel time cost function $AC(F)$ for stationary states, may very well be consistent with a concave reduced-form average cost function (including both travel time and schedule delay costs) for dynamic equilibria.

A second noteworthy feature of Figure 10 is that the generalized price experienced by users (including tolls) increases when tolling is implemented (as in Chu, 1995). This contrasts with the results of the Vickrey (1969) model for identical users, where this price does not change due to tolling (tolls exactly replace no-toll travel delay costs). However, the relatively small discrepancy between $c_{no\ toll}$ and p_{toll} in Figure 10 shows that this price increase will be modest, while the relatively large difference between p_{toll} and c_{toll} reflects that substantial efficiency gains are to be expected from departure time changes (the only source of efficiency gains in Vickrey's (1969) model), rather than from trip suppression.

7. Conclusion

Both the theoretical model and the empirical data suggest that the average travel cost function for congested traffic will be markedly different from the backward-bending function known from the conventional analysis, derived directly from the speed-flow function. Instead of bending backwards, the average cost function will rise vertically: at the road's capacity for a homogeneous road with a queuing facility before its entrance, or at the road's downstream segment's capacity in case of a 'bottleneck road'. Hypercongestion is dynamically unstable for roads without a downstream bottleneck, but will certainly occur on road segments with a downstream bottleneck as a dynamic equilibrium phenomenon, if demand is sufficiently large. When such hypercongestion occurs, however, the average travel cost function is still not bending backwards. Finally, it was shown that reduced-form average cost functions, that relate the sum of average travel cost and average schedule delay costs to the number of users in a dynamic equilibrium, need not have the intuitive convex shape, but may very well be concave – despite the fact that the underlying speed-flow function may be convex.

References

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

De Meza, D., and Gould, J.R., 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.

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

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

Mun, S.-I., 1999. Peak-load pricing of a bottleneck with traffic jam, Journal of Urban Economics, 46 323-349.

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

Ohta, H., 2001. Probing a traffic congestion controversy: density and flow scrutinized, Journal of Regional Science, 41 659-680.

Ohta, H., 2001. Probing a traffic congestion controversy: response to comment, Journal of Regional Science, 41 695-699.

Pigou, A.C., 1920. Wealth and Welfare, Macmillan, London.

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

Smith, W.S., Hall, F.L., and Motgomery, F.O., 1996. Comparing speed-flow relationships for motorways with new data from the M6, Transportation Research 30A 89-101.

Verhoef, E.T., 2001a. An integrated dynamic model of road traffic congestion based on simple car-following theory: exploring hypercongestion, Journal of Urban Economics, 49 505-542.

Verhoef, E.T., 2001b. Probing a traffic congestion controversy: a comment, Journal of Regional Science, 41 681–694.

Verhoef, E.T. 2003. Inside the queue: hypercongestion and road pricing in a continuous time – continuous place model of traffic congestion, Journal of Urban Economics, 54 531-565.

Vickrey, W.S., 1969. Congestion theory and transport investment, American Economic Review, 59 251-260.

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

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