

Rouwendal, Jan

Conference Paper

Driver behavior and congestion on highways

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

Provided in Cooperation with:

European Regional Science Association (ERSA)

Suggested Citation: Rouwendal, Jan (1998) : Driver behavior and congestion on highways, 38th Congress of the European Regional Science Association: "Europe Quo Vadis? - Regional Questions at the Turn of the Century", 28 August - 1 September 1998, Vienna, Austria, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:

<https://hdl.handle.net/10419/113622>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Driver behavior and congestion on highways.

Paper prepared for ERS98, Vienna

June 1998

Jan Rouwendal
Department of Economics and Management
P.O. Box 8060
6700 DA Wageningen
The Netherlands

Abstract

The paper discusses empirical evidence of highway congestion and observes that a number of seemingly important characteristics of this phenomenon are not explained by conventional economic models. It shows how to extend such models in order to incorporate these phenomena. The following results are derived.

If drivers reduce speed in order to allow traffic from a ramp to enter, a jam will be the result if a traffic flow exceeds capacity.

If there are queues in front of a point where traffic from a ramp merges with that on a road, full utilization of the capacity will usually not be reached. Hence there must be a decrease in inflow to a value strictly below the capacity of the bottleneck to let the queues disappear. This implies that there will be hysteresis.

Traffic that has passed the bottleneck will not necessarily return to a free flow situation immediately.

Random disturbances of traffic flow (e.g. because of a near accident) may cause a prolonged period of congestion because they may evoke a jam and a decrease in the traffic flow.

The paper concludes with a discussion of the implications for traffic control.

1. Introduction

Economic analysis of traffic problems focuses on the difference between user equilibrium and system optimum. The difference between both concepts is caused by an externality: the fact that one driver uses a road segment increases the travel times of other drivers on that road segment, at least for high traffic densities. Two types of models are employed by economists. The first makes use of a speed-flow function and gives essentially a static analysis of a congested road. This analysis dates back to Pigou [1920] and Knight [1924]. The second branch of literature originates from Vickrey [1969] and gives a structural analysis of the development and disappearance of a queue in front of a bottleneck. Both types of models aim at an analysis of the main characteristics of road congestion and give a deliberately simplified analysis. Nevertheless, one would expect that both types of models can be used for empirical work. The first type of model has indeed been regularly used for this purpose (see, e.g. Keeler and Small [1977]). However, its essentially static nature makes it difficult, if not impossible to use as a description of the inherently dynamic process of traffic congestion during peak periods.

At first sight, the second type of models seems to be much better equipped for the empirical task. A very brief description of a period of congestion based on that model is the following. Initially traffic proceeds at a constant high speed and a density that is low and the bottleneck's capacity is not reached. Then the density of traffic increases, capacity is exceeded, and from that moment on a queue develops in front of the bottleneck. The queue keeps growing as long as the traffic (in)flow exceeds the capacity of the bottleneck. However, the density of traffic will reach a peak and thereafter decrease. When it becomes so low that traffic flow does no longer exceed the capacity of the bottleneck, the length of the queue starts to decrease until it finally disappears completely. The model gives picture of congestion in front of a bottleneck that seems *a priori* plausible and is certainly testable. However, empirical work with the model seems to be lacking despite extensive theoretical investigations that point to its relevance (see e.g. Arnott, de Palma and Lindsey [1994]).

Nevertheless, much information congestion is to be collected. In a recent empirical study of traffic congestion on highways Kerner and Rehborn [1997] summarize empirical evidence from an extensive observations of congestion on a German highway. These authors studied the behavior of traffic in the neighborhood of a point where a ramp reaches a highway. This is a situation where traffic from $n+1$ lanes has to merge to n lanes for some $n \geq 1$, and this can clearly be regarded as a bottleneck. Kerner and Rehborn interpreted their data on the basis of a theory derived from physics and summarize the evidence as follows:

The (..) data allow us to conclude the following:

- (i) *There are two different types of phase transitions from in average basically similar initial states of free flow: (a) either to synchronized flow or (b) to jams. (...) Because of the growth of either a deterministic localized perturbation or a random perturbation one of these two types of phase transitions occurs in a localized region of the highway. The deterministic perturbation is caused, for example, by a deterministic peak in the flux of vehicles squeezing on to the highway from an on-ramp. The random localized perturbation whose growth leads to phase transitions can also be realized on a highway section without on-ramps or bottlenecks.*
- (ii) *A phase transition from free flow to synchronized flow caused by a peak in the flux of vehicles squeezing on to a highway from an on-ramp causes two deterministic processes: (a) a wave of induced transitions from free flow to synchronized flow in the upstream direction, and (b) a gradual spatial transition from synchronized flow to free flow in the downstream direction. After the phase transition has occurred the synchronized flow on the highway can further be self-maintained for several hours. This holds true even if the flux of vehicles on the highway upstream from the on-ramp become noticeably lower than the respective variables before the phase transition. (p. 4033)*

Even though the difference in terminology complicates the interpretation by economists somewhat, it is clear that these data do not confirm the picture that follows from the bottleneck model. That model predicts ‘the wave of induced transitions from free flow to synchronized flow in the upstream direction’ but most of the other characteristics seem to be lacking. The picture that emerges from the empirical research has the following characteristics that are lacking in the bottleneck model:

- a) The occurrence of a jam. The bottleneck model predicts only a switch from free flow to synchronized flow, but seems unable to show how and why a perturbation is able to cause a switch to a jam.
- b) Hysteresis. In the bottleneck model a return to a situation where the inflow of traffic is lower than the capacity of the bottleneck is sufficient for a return to the initial situation without queuing, whereas in reality a temporary increase in the flow is able to cause congestion for several hours.
- c) The gradual transition from synchronized to free flow in the downstream direction.
- d) The occurrence of random perturbations that lead to congestion.

We will refer to these characteristics as some *stylized facts* of highway congestion and the purpose of this paper is to analyse what changes have to be introduced into the bottleneck model in order to enable it to explain the stylized facts as a result of driver behavior.

Kerner and Rehborn interpret the data from the perspective of their theory, which is derived from physics. Such an approach has proved to be useful, witness the well known earlier work

on kinematic waves (Lighthill and Whitham [1955]) and kinetic theory of vehicular traffic (Prigogine and Herman [1971]) and may be able to give a better description of the data than a theory based on behavior. However, it is not self evident, to say the least, that a physical interpretation of traffic data should be preferred to a behavioral one. Application of theories developed in physics to traffic phenomena runs the risk developing an explanation in which driver behavior plays only an implicit role. This may result in overlooking possibilities to improve traffic conditions because a mechanistic model of driver behavior is used. For this reason, as well as for others,¹ it remains useful to develop models of traffic, which are based on explicit models of driver *behavior*. In the sections that follow some steps in the direction of such a behavioral model will be made. It will be shown that a temporary increase in the intensity of traffic that enters a road from a ramp (a perturbation) may evoke reactions of drivers that cause a sudden (phase) transition from free flow to traffic jam, and that the change from synchronized to free flow will usually result in underutilisation of the bottleneck's capacity, which gives rise to hysteresis.

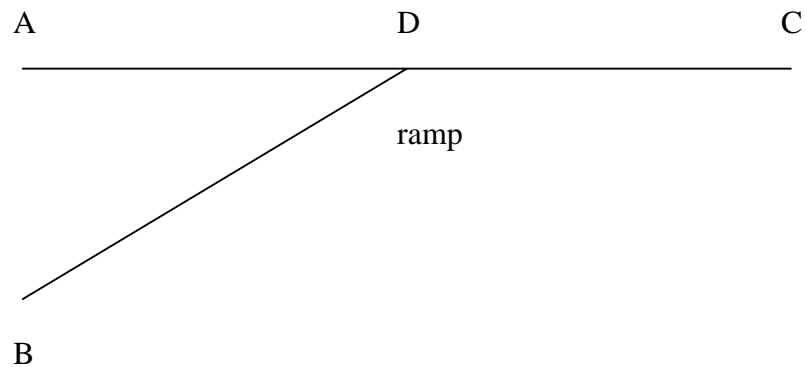
The paper is organized as follows. In section 2 a basic model is developed for a situation in which traffic from two lanes has to merge. The model is similar to the conventional bottleneck model. In section 3 we change our assumptions about driver behavior in the vicinity of the point where the traffic flows merge and show that plausible behavior leads to a switch from free flow to a jam. The jam results in a queue on both lanes. In section 4 we look at the way in which the traffic flows merge and speed up after these queues have been formed. We find that this process will as a rule lead to a traffic flow through the bottleneck that is lower than its capacity. Section 5 contains a short discussion of what happens when randomness and heterogeneity of drivers are introduced into the model. Section 6 concludes.

2. The basic model

2.1 The situation to be analyzed

Throughout most of the paper we will study two traffic flows that have to merge at a particular point. The situation may be interpreted as referring to two origins, say A and B, and one destination, say C. The roads from A and B to C are separate initially, but join at a point D. At that point traffic originating from B enters the road from A to C via a ramp. We will focus attention on what happens around point D and will often refer to 'the road' as the road from A to C and to the ramp as the last segment of the separate part of the road from B to C. The network is pictured in Figure 1. We assume until section 5 that traffic moving to C has one lane available both on the separate and on the joint parts of the network.

Figure 1 The situation to be analyzed



2.2 Assumptions

In this section the basic model of the paper is developed. The model is similar to the Vickrey-type bottleneck model of Arnott, de Palma and Lindsey [1990, 1993]. The main difference between that model and the one studied here is the situation to be studied. Arnott c.s. focus on a bottleneck that arises on a road when capacity decreases at a certain point on that road, whereas here attention is concentrated on what happens when traffic from two road has to merge. However, the difference between the two situations should not be exaggerated since a decrease in capacity of a road is often caused by a decrease in the number of available lanes. Arnott et al. do not describe the details of the bottleneck.

We start our analysis with a model with the following assumptions:

- 1 All cars have length a .
- 2 Drivers have a preferred speed v , which is chosen whenever possible.
- 3 Drivers want to maintain a distance of at least d with (the rear of) the car in front of them.
- 4 Drivers may change their speed immediately from 0 to v and vice versa, if desired.

Assumptions 2, 3 and 4 imply that a car stops whenever the distance to a car in front of it is less than or equal to d . Whenever the distance with the car in front is larger than d , the car is

given speed v by the driver. These assumptions are clearly extreme, but they serve as a starting point only and will be relaxed later on.

Consider traffic driving on a one-lane road. All drivers choose speed d and maintain a distance of at least d to the car in front of them. This means that there can be at most $1/(a+d)$ cars per unit of distance. The number of cars that passes an arbitrary point along the road is therefore at most equal to $c=v/(a+d)$. Clearly c is the capacity of the road under the assumptions made. The intensity of traffic is the number of cars that actually passes a point along the road. The intensity of traffic is at most equal to the capacity of the road.

For simplicity we study regular traffic flows, i.e. flows in which the distance between subsequent cars is equal to some value $d + \epsilon$, $\epsilon \geq 0$. If $\epsilon > 0$ there is a free flow of traffic, for $\epsilon = 0$ there is a synchronized flow. (Clearly, the expression ‘free flow’ has lost some of its usual content in the present simplified setting.) If $\epsilon > 0$, the intensity or flux of traffic is less than the capacity.

2.3 Low densities

We consider the situation shown in Figure 1 and assume that all links have one lane. Let b_1 denote the intensity of traffic on the road (from A to C), and b_2 that on the ramp (i.e. the final segment of the separate part of the road from B to C). We will first consider the case where the sum of both intensities is at most equal to capacity.

Assume that traffic on the road and traffic that enters from the ramp merges in the following way. Driver 1 approaches the road from the ramp. Driver 0 is the last one who passed point D before driver 1 arrives there. At the moment when driver 1 arrives at point D, the distance between driver 0 and driver 1 is at least equal to d , or smaller than d . In the former case, driver 1 enters the road with speed v . In the latter case, driver 1 stops at point D, waits until the distance with driver 0 becomes equal to d , and then enters the road with speed v . From point D onwards, drivers 0 and 1 will maintain a distance d and drive at speed v .

In situations where a driver from the road and a driver from the ramp reach point D at the same time, the priority for entering alternates. Although following this rule exactly will be difficult in practice when such occasions occur now and then, it is realistic under conditions of traffic congestion. (This assumption is important only in situations of traffic congestion.) Driver 1 does not take care of traffic that comes behind him. Let driver 2 be the one who is the first to pass point D behind driver 1. At the moment when driver 1 enters the road the distance between him and driver 2 may be at least equal to d , or smaller than d . In the former case driver 2 experiences no congestion. In the latter case driver 2 will stop at the moment when driver 1 enters the road, wait until the distance between them becomes equal to d and

then resume speed v . Driver 1 and 2 will from that moment on maintain distance d , while both driving at speed v .

Although some driver have to wait for each other for a short period of time, there will be no queuing, as long as the intensities b_1 and b_2 add up to a number that is at most equal to c . The reason is simple: under the assumptions made the capacity of the road segment DC can be used completely and there is sufficient space in between subsequent drivers on the road segment AD to let all traffic from segment BD enter the road. Although some drivers may have to wait at point D, none of them will have to do so for more than $(d+a)/v$ seconds. In the limiting case where b_1+b_2 is exactly equal to c , the capacity of the road segment DC will be completely used.

(The merging procedure implies that the traffic flow will in general not be regular anymore from point j onwards, but this is of no importance.)

2.4 Effects of a higher density

Now consider the situation in which the traffic flow that enters from the ramp is larger than b_2 for a (short) period of time. Denote the intensity of traffic on the ramp during this time interval as b_3 and assume that $b_1+b_3>c$.ⁱⁱ (The situation in which traffic on the road increases temporarily while that on the ramp is constant is completely symmetric and need not be discussed.) It will suffice to consider the illustrative case in which b_1 equals b_3 , because generalization to other cases is easy. Recall that flows on the ramp and on the road are both regular. This means that every $1/b_1$ seconds there arrives a driver from the road and one from the ramp. Let driver 1 be the first one to pass point j from the ramp after the intensity of traffic on the ramp has increased. Driver 1 may have to wait for a short period of time in order to let the distance with driver 0 (who is defined in the same way as above) become equal to d . Then he enters the road, and he will force driver 2, who is already on the road, to wait for some time until the distance between driver 1 and 2 becomes equal to d . To see this, observe that b_1 is greater than $c/2$ (since both intensities are equal and their sum exceeds c), which implies that the distance between driver 0 and driver 2 before point j was smaller than $2*d+a$. If the minimum distance between driver 0 and 1 and driver 1 and 2 becomes equal to d , the distance between drivers 0 and 2 will be equal to $2*d+a$, and this means that this distance has to be increased. But this can only occur (under the maintained assumptions) if driver 2 waits for some time.

If driver 2 has to wait for some time, the distance between him and driver 4, who is the first one behind him on the road before point D, will decrease. If driver 3 enters from the ramp between drivers 2 and 4, the implication will be that driver 4 has to wait longer in order to let

the distance with driver 3 become equal to d than driver 2 did in order to make his distance to driver 1 at least as large as d . It is possible that driver 3 was already waiting behind driver 1 before driver 1 could enter the road. In that case driver 3 may enter the road before, as well as after driver 2 depending on whose turn it is to have priority, but the implications for the time driver 4 has to wait are identical.

Denote the drivers that reach point D from the road subsequently as 0, 2, 4, et cetera. Driver 4 has to wait longer than driver 2 before he is able to pass point j. But driver 6 has to wait even longer since the distance between him and driver 4 has been shortened because of the waiting time of driver 4, and a car from the ramp will enter between them. Driver 8 has to wait still longer and in this way a queue develops on the road before point D. In exactly the same way a queue will develop on the ramp.

In the case considered here, where traffic intensities on the road and on the ramp are equal, the two queues will grow at with the speed.

What happens if the intensities of traffic on the road and on the ramp are different? As long as the sum of both intensities exceeds capacity, there must occur a queue on either the ramp or the road, or on both. If both b_1 and b_2 exceed $c/2$, queues will develop on the ramp and the road. Since vehicles from the road and from the ramp pass point D alternately as soon as the queuing has started, the queue on the road increases with $b_1 - c/2$ vehicles per second and that on the ramp with $b_2 - c/2$ vehicles per second. When one of the traffic intensities is below $c/2$, there will be queuing either on the ramp or on the road, but not on both. If $b_1 > c/2$ and $b_3 < c/2$, there will develop a queue on the road only. This queue increases with $b_1 + b_3 - c$ vehicles per second. The situation in which a queue develops on the ramp only is similar.

Traffic in the queue moves in a stop-and-go style because of the assumptions 2-4 made above. The average speed of traffic in the queue is equal to $c \cdot (d+a)/2$.

Note that the intensity of traffic from the bottleneck onwards will be equal to the capacity c of the road from the moment that b_2 switches to the higher value b_3 .

2.5 Effects of the return to the original density

After the intensity of traffic on the ramp returns to its original - lower - value b_2 the queue or queues will disappear. The following situations may be distinguished. If there is no queue on the ramp, b_3 was lower than $c/2$, and hence $b_2 < c/2$. When traffic intensity on the ramp decreases, traffic from the road passes point j with intensity $c - b_2$, which is greater than b_1 . This means that more cars leave the queue at point j than enter it on its tail. The queue will therefore decrease. It is easy to verify that it decreases with $c - b_2 - b_1$ cars per second.

If there is a queue on the ramp only, $c - b_1$ cars will continue to leave the queue, which will therefore decrease with $c - b_1 - b_2$ cars per second as soon as traffic intensity on the ramp has decreased.

If there are queues on the road and on the ramp, the queue on the ramp will start to decrease as soon as traffic intensity of the ramp decreases. The decrease will be equal to $c/2 - b_2$ cars per second. However, as long as the queue on the road has not completely disappeared, the queue on the road will continue to increase with $b_1 - c/2$ cars per second. If the queue on the ramp has disappeared, there will be $c - b_1$ cars that leave the queue on the road per second, instead of $c/2$ cars. Since $c - b_2$ is larger than b_1 , the number of cars that enters the queue per second, the queue will start to decrease and will finally disappear.

2.6 Evaluation

The model developed in this section is very close to the bottleneck models. A queue develops on one or both of the road segments AD and BD as soon as the sum of the traffic flows exceeds the capacity c of the road segment DC, which is the bottleneck. The queue disappear again when the sum of the traffic flows on the road segments AD and BD falls below c again. The model is unsatisfactory when judged on its ability to explain the stylized facts of congestion on highways listed in section 2. First of all, there is no prediction of a jam at the point where ramp and road join after the intensity of traffic on the ramp increases. True, the cars have to wait before being able to pass point D, but at that point there occurs a steady flow of traffic at speed v and intensity c . Second there is no hysteresis. A reduction of the intensity of traffic to its original value is always sufficient to let traffic return to the original situation without congestion. Third, there is no prediction of a gradual transition from synchronized to free flow downstream from D. Fourth, there are no random perturbation that may cause congestion.

On the positive side, it should be noted that the model is able to explain switches from free flow to synchronized flow upstream from point D on the road, on the ramp or on both as a result of temporarily increased density of traffic on the ramp, which is undoubtedly one of the most important empirical characteristics of highway congestion. It is also able to predict situation in which a queue develops on the road, but not on the ramp, or vice versa. In the following sections it will be shown that relatively straightforward extensions of the basic model developed here will suffice to introduce the desired additional characteristics.

3. More realistic assumptions about changes in speed

3.1 General

The assumption that drivers will either have speed 0 or speed v is obviously unrealistic. It seems more appropriate to take into account the desire of drivers to avoid sudden changes in speed, as well as their inability to realize such changes due to technological constraints on brakes and acceleration power.

What would such alternative assumptions imply for the above model? Consider first the situation of free flow traffic. All that drivers have to do in that case is to adapt their positions through speed changes in such a way that the minimum distance remains equal to d . Drivers on the ramp who anticipate entering just behind a car on the road if they maintain speed v will throttle back somewhat in order to enter the road with a distance to that car equal to d , and speed up again as soon as they enter the road. Drivers on the road who anticipate that a car will enter from the ramp just in front of them will also decrease their speed temporarily in order to create enough space. This adjustment process takes place without the large speed differences assumed in the model of the above paragraphs, but with the same results.

Now consider what happens if the capacity of the road is exceeded by the sum of the traffic intensities. Drivers on the ramp and on the road will reduce their speed in order to create enough space between them. Since speed cannot be immediately increased, or since drivers do not want to realize a sudden increase in speed, the speed at which traffic passes point D will decrease. This implies that the number of cars that pass point D will be lower than c . The capacity of the road at D will therefore not be used completely, and this worsens the congestion problem.

This effect may occur to a limited extent even under conditions of free flow. However, when the capacity of the road is exceeded it takes on a qualitatively different form because cumulative effects will occur. Drivers reduce speed in order to create enough space between them and their predecessors. The reduction in speed should at least be large enough to make one's own speed smaller than that of the predecessor, who may himself have reduced speed already. If capacity of the road is exceeded, speed adjustments are unable to create sufficient space between all subsequent cars. All subsequent cars that approach the point where road and ramp join each other have to reduce speed more than their predecessors did. The ultimate result will be that traffic stops completely.

3.2 An example

To make this more clear, consider the case with a regular traffic flow on the road. The distance between subsequent cars is $d+\epsilon$, with $\epsilon < a+d$. The intensity b_3 of traffic that enters

from the ramp is such that between every pair of subsequent cars on the road a car from the ramp should enter. For simplicity, assume that the intensity b_2 equals 0. The first driver that enters from the ramp is driver 1. He enters the road in front of driver 2, who reduces his speed temporarily in order to create enough space between him and driver 1. If he anticipates driver 1 to enter the road, driver 2 may have already created enough space when this happens to resume speed v at point D. However, the distance between driver 2 and driver 4, who is the first driver on the road segment AD behind driver 2, will have been reduced by the temporary reduction in speed of driver 2. A car, that of driver 3, will enter the road from the ramp between drivers 2 and 4. In order to keep the distance with that car at least equal to d , driver 4 should have anticipated the bottleneck at point D earlier than driver 2 did, or should reduce his speed more than driver 2 did, or both. Similarly, driver 6 should have anticipated the bottleneck earlier than driver 4 did, or reduce his speed more than driver 4 did, or both. Et cetera. Since earlier anticipation of the bottleneck becomes increasingly difficult and therefore improbable for drivers with higher even numbers, speed reductions will ultimately have to do the job. However, speed reductions are insufficient, since there is simply not enough space between the cars on road segment AD to let all cars from the ramp enter. Moreover, speed reductions worsen the problem since they imply that the number of cars that passes the bottleneck is smaller than c . Speed will have to be reduced more and more, until it reaches its lower bound: 0. This means that a jam occurs.

The analysis for other cases, in which the intensities of traffic on the road and on the ramp differ, is similar and leads to the same conclusion. If capacity is exceeded, speed reductions will be insufficient to keep traffic flowing and there will occur a traffic jam.

3.3 Conclusion

In this section the result of more realistic assumptions about speed changes, which are based on the hypothesis that drivers wish to avoid, or are unable to realize, rapid changes, have been studied. It has been concluded that such alternative assumptions lead to the prediction that an increase in traffic intensity on the ramp will lead to a complete standstill of traffic if capacity is exceeded. This means that an important stylized fact of congestion on highways has now been introduced in our model: a transition from free flow to jam as a result of a deterministic perturbation in the form of increased traffic intensity on the ramp.

The occurrence of the traffic jam at D will, of course, have consequences for traffic upstream from D. If the standstill continues, there will be a transition to jam also on points along that part of the highway. However, experience suggests that traffic speeds up again soon after the initial standstill, but at a lower speed than v . The emergence of a queue implies a transition

from free flow to synchronized flow upstream point D on the ramp or on the road or on both. However, the basic model already predicted this phenomenon.

4. Hysteresis

4.1 General

One can imagine that traffic speeds up after the initial standstill just like the carriages of a train do. Then all drivers keep the mutual distance with their predecessors equal to d while increasing speed. However, experience demonstrates that distances between cars increase during the process of speeding up. In the situation studied here, this phenomenon is necessary in order to enable the flows from the road segments AD and CD to merge into one flow on the road segment DB. Acceleration from the low speed in the queue to a higher speed after passing the queue interacts with the merging of the two flows of traffic into one. Experience suggests that there will emerge a stationary process once traffic has recovered from the initial jam. In this section that process will be studied.

In the present context hysteresis is identified with a situation in which the capacity of road segment DC will not be used completely after a queue has developed in front of point D on road segment AD or on road segment BD or on both. We will focus on the latter situation. Such an under-utilization of the capacity of DC implies that the inflow of traffic from A and C at the tails of the queue has to be *strictly lower* than c in order to remove the queues. Although a queue will develop only if inflow from A and C exceeds c , it has to fall back to some number $c^* < c$ in order to remove the queue.

Three situations in which the capacity of road segment DC is not fully used can be distinguished: (i) it is possible that the cars that pass have only the minimal distance d between them, but have a speed that is lower than v , (ii) the cars that pass point D have speed v , but the distance in between them exceeds d and (iii) speed is lower than v and distance in between exceeds d . In the first case, the number of cars that pass point D is equal to the ratio of the actual speed, which is smaller than v , and the distance $(a+d)$. This ratio will therefore be smaller than c , which was defined as $v/(a+d)$. In the second case the number of cars that pass point D is equal to the ratio of v and (the actual distance between subsequent cars plus a). Since the denominator is larger than $(a+d)$, this ratio will be smaller than c . In the third case the numerator will be smaller than v and the denominator larger than $(a+d)$, hence the ratio will also be smaller than c .

4.2 Stationary situations

In this subsection we study the way in which traffic behaves around point D after a queue has developed on both road segments AD and BD. It is assumed that traffic has recovered from the initial jam we study the switch from synchronized to free flow.

In the queues the minimum distance d between the cars is relevant. At point D, traffic merges and this means that, under our assumptions, enough space should be created between two cars that are behind each other in a queue to let one car from the other queue enter in between.

This can only happen if the leader moves faster than the follower for some time. If the speed of traffic in the queue is constant, say v^c , then it means that the leading car has to accelerate before point D. Let D' be the point on the road where acceleration starts and assume for simplicity that acceleration is a constant, k .

If it takes t seconds for the leading car (car 0) to move from D' to D, its distance with the following car (car 2) will have increased by $kt^2/2$ meters if the following car keeps speed v^c . We should therefore have:

$$k t^2 / 2 \geq a + d \quad \mathbf{1}$$

and from this equation we solve for t :

$$t \geq [2 (a + d) / k]^{1/2}. \quad \mathbf{2}$$

The process will be stationary if (and only if), each car in the queue accelerates at the same point D' . This means that car 2 has to reach point D' in the t seconds it takes car 0 to move from D' to D. In t seconds car 2 will have passed $v^c t$ meters. Since car 2 was immediately behind car 0, we should have:

$$v^c t = a + d. \quad \mathbf{3}$$

Equations (1) (or (2)) and (3) should hold simultaneously. Solving from equations (2) and (3) we derive for v^c :

$$v^c \leq [(a + d) k / 2]^{1/2}. \quad \mathbf{4}$$

Now observe that the capacity of the road segment DB will not be completely used if (1) holds as a strict inequality. Hence there will be hysteresis in that case.

If (1) holds as an equality, (4) will also hold as an equality. The speed that will be reached at point D is equal to $v^* = v^c + kt$ and can be computed as:

$$v^* = 3 [(a+d) k / 2]^{1/2}. \quad 5$$

There is no a priori reason why v^* should be equal to the preferred speed v . We can therefore distinguish 3 cases:

- a) $v^* < v$. There is hysteresis: situation (i) of the preceding subsection is relevant.
- b) $v^* = v$. The capacity of road segment DB is completely used. There is no hysteresis.
- c) $v^* > v$. Then (1) cannot hold as an equality, and we also have hysteresis: situation (ii) of the preceding subsection is relevant.

It follows from this analysis that hysteresis will be present except in the special situation in which v^* is exactly equal to v . Since there is no a priori reason why this should be so, hysteresis will be the rule.

Before concluding this subsection, we should note that the analysis of traffic on the ramp is exactly equal to that on the road as long as there is a queue on both. The analysis becomes somewhat different if there is a queue on the road, but not on the ramp. The merging process will then be somewhat more complicated because two or more cars from road segment AD may pass point D immediately behind each other before a car from the ramp enters in between. We do not provide a formal analysis for this situation, but note that the lack of regularity introduces friction into the merging process that makes it hard to reach the maximum flow c at point D. Hysteresis will therefore also be the rule in this situation. A similar remark can be made for the analogous situation in which there is a queue on the ramp, but not on the road.

4.3 Which case is relevant?

The empirical evidence summarized in section 1 suggests that there will be a gradual return to free flow traffic after point D has been passed. The analysis presented in the previous subsection clearly allows the possibility that speed the v^* at which traffic passes point D is lower than v and associates this with hysteresis. This suggests that situation (i) is relevant in practice.

It is possible to elaborate this conjecture somewhat if it is noted that the analysis predicts that v^* is equal to $3v^c$ (cf. Equations 4 and 5). Although the situation analyzed by Kerner and

Rehborn [1997] is clearly more complicated than the one considered here, some comparisons can be made. Their figures 1(b) and 2(h) show that during the congested period speed in the queues falls back to 30 km/h. However, figure 1(b) refers to the left lane (of three) where speed may have been somewhat higher than on the other lanes, and figure 2(h) refers to a position that may be so close to the point where traffic merges that the cars are already accelerating. This implies that v^c may be lower than 30 km/h. The detector that is placed at the point where traffic from the ramp merges with that from (the right lane of) the road (detector D3) shows speeds just above 60 km/h during the congested period on the right lane and speeds on the middle and left lane that are close to 70 km/h on average. This implies a ratio of the speeds at the bottleneck and in the queue which is lower than predicted by the equations of subsections 4.2, but the difference may well have to do with complications that are important in practice, but assumed away in the simplified analysis presented there.

4.4 Discussion

In this section it has been shown on the basis of some simple equations that hysteresis will be the rule if traffic from a road and a ramp have to merge and there is a queue on both road segments. This conclusion is based on assumptions about the minimum required distance between subsequent cars and the way cars accelerate. There may, of course, be other reasons that cause hysteresis. For instance, in situation (i), which appears to be the most relevant one, it is perfectly reasonable that safety considerations lead drivers to choose a larger distance to their predecessor than the minimum value d in the vicinity of point D. This would, of course, strengthen the case for hysteresis. Moreover, such additional factors would probably relax the strong prediction regarding the ratio of the speeds at the bottleneck and in the queue to some extent.

5. Additional generalizations

5.1 Irregular traffic

In this section we will discuss some further generalizations of the analysis carried out in the previous sections.

Until now we have assumed that traffic flows on AD and CD are regular, in the sense that the distances between subsequent cars are equal. This assumption is, of course, not realistic. It was only made for convenience and all the phenomena that have been studied will also occur under conditions of irregular traffic, and some should even be expected to be more likely to occur. For instance, if a stochastic process (e.g. Poisson) determined distances between subsequent cars, it is possible during a short interval of time many cars will approach point D

either from the ramp or from the road, or from both. Especially in the latter situation, the capacity of the road may be exceeded for a short period of time, which may nevertheless be sufficiently long to evoke congestion when combined with the hysteresis phenomenon. The deterministic perturbation that has been studied in the previous sections can arise randomly under conditions of irregular traffic.

5.1 More lanes

Until now we have studied highway congestion with a model that assumes only one lane roads. What will change in the analysis if there are more lanes? To answer this question, assume that the road from A to C has two lanes, while the road segment from B to D has only one lane. Traffic enters the road at D from a ramp that merges with the right lane. If a car enters from the ramp, cars on the road may either reduce speed to create enough space or move to the right lane, if there is enough space available. However, if the density of traffic on the road is large, this may be impossible, and then the same mechanism that created a jam in the situation with one lane will also cause a jam in this case.

When this happens, traffic from the right lane approaching D will try to move to the left lane, and this creates a similar situation on that left lane as when traffic from a ramp would directly enter that lane, including the associated mechanism for a jam. It seems therefore that the simplified situation studied in the previous section is also relevant for the situation in which there are more lanes.

5.2 Heterogeneous traffic

The assumptions that all cars have the same length, that all drivers want to maintain the same minimum distance to their predecessor and all have the same desired speed and acceleration behavior are of course all simplifications that one would prefer to relax. Although it seems possible to do so, the discussion in the previous sections has been focused deliberately on a very simplified situation in order to concentrate attention on the fundamentals. However, introducing heterogeneity of traffic seems to be necessary in order to cover the stylized fact that has not been considered yet.

Heterogeneous traffic would allow for the possibility of, for instance, a slowly moving truck on one of the lanes of a highway. If traffic densities and speeds on the other lanes allow drivers to do so, many will take over by means of a temporary move to another lane. But the moves between lanes of cars with differing speeds will cause coordination problems that are similar to those discussed above. The main difference will of course be that instead of a given point where two lanes merge, there will now be a bottleneck that moves. This should be expected to prevent a jam.

However, the situation becomes different if, for instance because of a near accident, one or more cars reduce speed fastly and unexpectedly to zero. Then one lane becomes temporarily unavailable for traffic and this necessitates traffic on that lane to merge with that on the other lanes (if present), with as a possible consequence the occurrence of a jam on the other lanes (and the associated possibility of additional real or near accidents). The temporary random disturbance caused by the near accident may nevertheless cause congestion for a prolonged period because of hysteresis, as was explained in the previous section.

6 Summary and conclusion

6.1 Driver behavior and the stylized facts

In section 1 of this paper the following stylized facts of traffic congestion that are not covered by conventional economic models have been listed:

- a) The occurrence of a jam.
- b) Hysteresis.
- c) A gradual transition from synchronized to free flow downstream the bottleneck.
- d) The occurrence of random perturbations that lead to congestion.

In the sections that followed all these empirical characteristics of highway congestion have been dealt with. In section 2 a basic model was developed that is similar to the bottleneck model in many respects and was used to analyze traffic from two lanes that had to merge at a certain point. The model is unable to explain the stylized facts. In section 3 the plausible assumption was made that drivers create space for cars that enter from a ramp by throttling back. It was shown that this leads to occurrence of a jam when the capacity of the bottleneck is exceeded by the sum of the two flows. In section 3 traffic behavior in the vicinity of the bottleneck was analyzed in congested situations. It was shown there that plausible assumptions about stationarity of the process by which drivers accelerate and the two flows merge lead to the conclusion that hysteresis should be expected to be the rule, rather than the exception. Moreover, one of the situations in which hysteresis occurs implies that the speed of traffic at the bottleneck is lower than the desired speed. This implies that downstream the bottleneck a further increase in speed will take place. This can be interpreted as a gradual transition from synchronized to free flow. In section 5 generalizations of the model to situations in which there are more lanes on a single road segment, traffic is irregular and preferred speeds of drivers may differ were shortly discussed. In such a more general setting where randomness plays an essential role, the occurrence of congestion also becomes a random phenomenon.

It may therefore be concluded that these stylized facts can be introduced into economic models so as to make these models more realistic. Of course, one should ask what gains should be expected from such an extension of these models, given that the essence of model building is to concentrate on the main features of the phenomenon under study while leaving out the less important ones. An answer to that question is provided in the next subsection.

6.2 The stylized facts and economic analysis

If one would build an economic model that took into account the stylized facts that have been dealt with in this paper, what would be the difference with the basic model developed in section 2, which is similar to the bottleneck model?

- a) The start of the congested period with a jam at the bottleneck would imply that for a short period of time the flow of traffic through the bottleneck will be very low. The queues upstream point D will initially grow much faster than is suggested by the basic model.
- b) The hysteresis associated with congestion implies that the queue will disappear at a slower rate than implied by the basic model. The queues will be longer and remain present for a longer period.
- c) Congestion will have implications for traffic downstream the bottleneck as well since it takes some time before a return to free flow conditions has been realized.
- d) The occurrence of congestion is co-determined by random events.

It may be argued that at least some of these additions will result only in minor modifications of conventional economic models. In particular, the gradual return to free flow mentioned under c) seems to be of less importance. Also the start of a congested period by a jam, mentioned under a) seems to call for a minor modification of the model only. However, the combination of hysteresis and randomness implies that an essentially unpredictable event such as the occurrence of a temporary large density of traffic on either the road or the ramp evokes a prolonged period of congestion results in a qualitative difference with the basic model. The basic model is essentially a deterministic one and the optimal tolls are derived from deterministic relations. These relations are best thought of as referring to average speeds, densities, flows et cetera. The tolls intend to keep these averages at their optimal values and to handle congestion in an efficient manner through that instrument. The analysis of the present paper suggests, on the other hand, that it might be as important to control the variation in traffic density in order to prevent a (random or deterministic) local perturbation to evoke cumulative effects. Tolls may be effective means to control average densities and flows of traffic, but they appear to be less effective as instruments to control variations in these variables.

Traffic engineers have often been reluctant in supporting the economists' advice for tolling. They seem to be more in favor of instruments that allow for localized traffic control. The closer look at the details of the highway congestion problem that has been provided by Kerner and Rehborn [1997] suggests that such instruments are important, if not necessary, supplements to road pricing.

Literature

- Arnott, R., A. De Palma and R. Lindsey [1990] Economics of a Bottleneck, Journal of Urban Economics, **27**, 111-130.
- [1993] A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand, American Economic Review, **83**, 161-179.
- Keeler, Th. E. and K.A. Small [1977] Optimal Peak-Load Pricing, Investment and Service Levels on Urban Expressways, Journal of Political Economy, **85**, 1-25.
- Kerner, B.S. and H. Rehborn [1997] Experimental Properties of Phase Transitions in Traffic Flow, Physical Review Letters, **79**, 4030-4033.
- Knight, F. [1924] Some Fallacies in the Interpretation of Social Costs, Quarterly Journal of Economics, **38**, 582-606.
- Lighthill, M.H. and G.B. Whitham [1955] On Kinematic Waves II: A Theory of Traffic Flow on Long Crowded Roads, Proceedings of the Royal Society, **A229**, 317- 345.
- Pigou, A.C. [1920] Wealth and Welfare, London, MacMillan.
- Prigogine, I. and R. Herman [1971] Kinetic Theory of Vehicular Traffic, Elsevier, New York.
- Vickrey, W.S. [1969] Congestion Theory and Transport Investment, American Economic Review, **59** p&p, 251-260.

Notes

ⁱ It may be argued, for instance, that a theory about driver behavior is inherently more satisfactory as an explanation of traffic phenomena even if it does not result in better (quantitative) predictions of traffic flow .

ⁱⁱ The case in which $b_1+b_3 \leq c$ is trivial.