The Two-Capacity Phenomenon: Some Theoretical Issues

JAMES H. BANKS

Theoretical issues related to two phenomena observed in a study of four freeway bottlenecks in San Diego are addressed. It was observed that flow immediately downstream of the bottlenecks decreased by a small amount when it broke down and that flow breakdown appeared to be triggered by speed instability. Most of the flow decrease could be attributed to the increase in vehicle passage time that occurs when speeds decrease, and most of the San Diego data are compatible with the linear car-following model of Chandler et al. as extended by Bexelius, although a number of questions about the validity and applicability of this model remain.

A previous paper reported on a study of flow processes in the vicinity of a high-volume freeway bottleneck in San Diego, in which detailed detector data were analyzed and compared with videotapes of traffic flow (*1*). Evidence was found that supported the hypothesis that capacities at this bottleneck decrease when queues form. In addition, it was found that queues form upstream of the merge point, despite extremely high merge rates and volumes downstream of the bottleneck. Subsequently, three additional bottlenecks on San Diego freeways were studied as a part of the same research project. These additional case studies are described in detail in the project final report (*2*); the results and their implications for ramp metering are discussed in this paper.

The two-capacity hypothesis was confirmed at all four locations. The test of the two-capacity hypothesis that was developed by Banks (*1*) was an extension of work by Hurdle and Datta (*3*) and Persaud (*4*). It was assumed that individual lanes might have separate capacities, and concluded that if even one lane is found in which the mean flow rate is less after the queue forms, or in which the highest short-term volume counts are less frequent after queue formation, this would confirm the hypothesis. This does not necessarily imply that flows across all lanes will decrease if the hypothesis as stated here is true, because the distribution of traffic across the lanes may shift.

Of the four sites, one was on a slight downgrade, and the other three were on upgrades ranging from 2 to 6 percent. Also, in two cases, there were apparently critical horizontal curves. Roadway widths ranged from two to five lanes in one direction; in all cases, lane and shoulder widths were standard. In three cases, the maximum volume per lane occurred just downstream of an on-ramp; in the other it was just upstream of a heavily used off-ramp. All sites were in the vicinity of metered on-ramps; the extensiveness and effectiveness of upstream metering varied considerably, however. Further details concerning site conditions were provided by Banks (*1,2*, and in the companion paper in this Record).

The study methodology involved comparison of videotapes of traffic flow with detailed analyses of 30-sec detector data. The most important analyses included comparison of flows averaged over 12-min periods before and after queue formation, comparisons of distributions of 30-sec counts for similar periods aggregated over all days studied, and linear regressions of counts versus time, which were used to determine whether flows were increasing significantly just before breakdown. Videotapes were used for supplemental counts and to confirm that queues were not backing up into the sections from downstream.

At all four sites, the left lane was the most heavily used, both before and after queue formation. On a majority of the days studied at each site, flows in this lane decreased when queues formed; when evaluated by the sign test, the number of decreases was statistically significant in three of the four cases. Flows across all lanes decreased significantly in one case; in the other three cases, there was no significant change. In two cases, this appeared to be because flows were increasing significantly just prior to queue formation, so that the prequeue average flow understated the flow at the time of queue formation; in the other case, the decrease in flow in the left lane was not statistically significant, and the most likely explanation is that the flow process itself was different from that at the other sites. In addition, when counts for 12-min periods before and after queue formation were aggregated over all days studied, the highest counts were less frequent after queue formation. This was true at all sites both for left lane counts and for counts averaged across all lanes. The decreases in flow did vary considerably from site to site, however.

In addition, it was found that in every case in which the typical point of flow breakdown could be seen, it was somewhere other than at the merge or diverge point that would have been identified as critical by Chapter 5 of the *Highway Capacity Manual* (*5*); this was in spite of the fact that the merge or diverge rates in question were far in excess of the supposed capacities of merge or diverge points.

In the course of the study, it was possible to observe the flow breakdown process on many occasions. These observations, coupled with the major findings cited above, raise several theoretical issues. Among these are the interrelated questions of what caused flow to break down and why flow decreased when it broke down. The flow breakdown process as observed at these sites is described, and these issues are discussed.

Civil Engineering Department, San Diego State University, San Diego, Calif. 92182.

OBSERVED PROCESS OF FLOW BREAKDOWN

The empirical literature describing phenomena related to flow breakdown at freeway bottlenecks is not very extensive. The flow breakdown process is also described by Persaud (*4*), who described the initiation of queueing at a lane drop in Toronto; Newman (*6*), who described flow breakdown at a merge location in Los Angeles; and Edie and Foote (*7,8*), who described flow processes in a New York tunnel. In addition, Forbes and Simpson (*9*) described driver and vehicle responses in freeway deceleration waves (including the initiation of such waves) on the basis of trajectories derived from aerial photographs; similar work has also been carried out by Trieterer and Myers (*10*). Of this work, that which is most relevant to the present study is Edie and Foote (*8*). Although the bottleneck they described was different in many ways from those considered here, the process of flow breakdown was similar.

When it was possible to see what happened, the process of flow breakdown appeared to be similar at all the San Diego sites. This was in spite of considerable differences in site characteristics. Three of the four bottlenecks involved wide freeways (four to five lanes in one direction) of modern design; of these, two were on extended upgrades (of 2 and 3 percent, respectively) and the third was on a downgrade, but featured a possibly critical horizontal curve (with a 2,000-foot radius). The fourth bottleneck involved two lanes in one direction and featured much more restrictive geometry, including an extended upgrade, portions of which were as steep as 6 percent, and a 600-ft horizontal curve. In all cases, percentages of heavy vehicles were low (around 2.0 to 4.5 percent). Because the point of flow breakdown was difficult to see in the case involving a downgrade, most of the description here was derived from the upgrades; however, when it was visible, the sequence of events on the downgrade was similar.

This sequence normally began with the arrival of a vehicle traveling somewhat slower than the average speed of vehicles in the lane in question. Because flows and densities were great enough to impede passing, dense platoons of vehicles collected behind these slow-moving vehicles. Eventually, speeds in the platoon became unstable, with speeds of vehicles at the upstream end dropping below that of the leader; once the speed at the upstream end of the platoon dropped below a certain value, the instability appeared to escalate, and the usual outcome was that several vehicles would stop. Meanwhile, vehicles in adjacent lanes also reacted to the decrease in speed, so that eventually speeds in all lanes approached zero.

Once this happened, a shock wave would form and move upstream. These waves consisted of a small core of closely spaced vehicles that were stopped or nearly stopped, a zone of deceleration immediately upstream, and a zone of acceleration immediately downstream. Deceleration and acceleration, from zero to about 30 mph, took place rapidly; but acceleration above 30 mph occurred somewhat more gradually.

In many cases, secondary waves were observed to form repeatedly in the accelerating flow downstream from the primary wave. This process resulted in a flow pattern upstream of the bottleneck consisting of brief periods in which there was rapid deceleration followed by rapid acceleration as shock waves passed upstream, followed by rather longer periods in which speeds were nearly constant (but below the speed before flow breakdown) or gradually increasing. At two of the sites, multiple waves were common, so that the nearest wave never got far enough upstream for speeds in the vicinity of the bottleneck to recover completely. At the other two sites, on the other hand, isolated waves sometimes did occur, and speeds did recover.

THEORETICAL BACKGROUND

Two main features of the flow processes are explained. The first feature is the decrease in flow that occurs at flow breakdown; a related phenomenon is the structure of the shock waves (i.e., the dense core with deceleration and acceleration zones up- and downstream). Both of these features are primarily related to what might be called the "mechanics" of speed and flow relationships. The second major feature is the speed instability that seems to trigger flow breakdown; associated with this is the tendency observed at some of the sites for secondary shock waves to form in the accelerating flow downstream of the primary wave. These appear to be related to driver behavior; specifically, to car-following behavior.

One of these issues has a considerable history in flow theory literature; the other does not. The two-capacity hypothesis itself has a fairly long history, but most of the literature discussing it never goes into the question of why the phenomenon should occur. One reason for this is that most of the early discussions of it related it to so-called "two-regime" or "dual-mode" traffic flow theories (*11–14*). In these cases, it was noted that there appeared to be discontinuities in macroscopic data relating flows to speeds or concentrations, and it was assumed that these might indicate different capacities for congested and uncongested flow. One of the points of departure for this work was an awareness that these gaps in the data might not have any such implications. In particular, if the data were taken upstream of the bottleneck, a drop in flow is to be expected when the shock wave at the upstream end of the queue moves past the observer (*15–17*). The only previous study in which evidence that the phenomenon occurs is combined with an attempt to explain it appears to be Edie and Foote (*8*). The explanation they give is sketchy, but it serves as an important point of departure for that proposed here.

Meanwhile, the structure of the shock waves and the fact that distinct waves form repeatedly in flow upstream of fixed bottlenecks have long been known, although these facts have often been ignored in discussions of shock wave movement. Technically, a shock wave is a discontinuity between flow regions with dissimilar flows and densities. In much of the literature related to shock waves, there is a tendency to suppose that in queues upstream of fixed bottlenecks, there is a single shock wave between two relatively homogeneous flow regions: the high-density flow of the queue itself and the low-density flow approaching the queue from upstream. In contrast to this, qualitative descriptions of congested flow have long emphasized its instability and the pattern of repeated shock waves (sometimes referred to as the "accordion effect"). There have even been attempts to quantify the phenomenon (*18,19*) and to explain it (*20*).

On the other hand, there is an extensive literature related to the relationship between driver behavior and flow stability. The most important work along these lines is a series of studies of car-following processes published by researchers at General Motors Research Laboratory in the late 1950s and early 1960s. This work involved mathematical models relating the acceleration of a particular vehicle to the speed difference and distance spacing between it and the preceding vehicle. The most important work in this series was conducted by Chandler et al. (*21*) and Herman et al. (*22*), who propose the so-called "linear" model and explored its stability characteristics; Gazis et al. (*23*), who proposed the reciprocal-spacing model and established a link between the microscopic car-following models and macroscopic flow models; and Gazis et al. (*24*), who generalized the earlier particular models into a family. Other important early work in this area was conducted by Newell (*25*) and Edie (*14*). A somewhat later effort by Bexelius (*26*) extended this work to consider cases in which drivers respond to the speeds and spacings of more than one vehicle ahead of them.

THE FLOW DECREASE

A key feature of the flow breakdown process, as observed at the San Diego bottlenecks, was that it appeared to be triggered by unstable speeds. This speed instability resulted in a brief but drastic reduction of speed at the point the shock wave began, in which a few vehicles either stopped or came near to stopping. There are at least two ways in which such a speed disturbance could affect flow, as measured immediately downstream.

First, whenever two vehicles are traveling at different speeds, both their time and distance separations must be changing. Because the flow rate is the reciprocal of the average time headway, any drastic change in speed such as that described earlier should result in a change in flow. Specifically, at the beginning of such a speed disturbance there should be a decrease in flow just downstream as the last vehicles not involved in the speed decrease pull away from the first vehicles involved in it.

Second, the time headway consists of the time it takes a vehicle to pass a point (which is referred to as the "passage time") and the time gap between the vehicle's front bumper and the rear bumper of the preceding vehicle. Passage time is a function of vehicle length and speed only and must increase as speed decreases. Unless time gaps decrease without limit (which is implausible), speed decreases must eventually increase the average time headway and thus decrease the flow rate.

Note that these two mechanisms are essentially different. The first affects flow in front of the first vehicles to slow down, and tends to increase the time gaps ahead of them. The second affects flow behind the vehicles slowing down but may not involve increases in time gaps. The second of these mechanisms is the one identified by Edie and Foote (*8*) as crucial to the development of shock waves in the tunnel they studied. This mechanism will also be shown to be the more important of the two phenomena in explaining the flow decreases at the San Diego bottlenecks.

SPEED DIFFERENCES BETWEEN VEHICLES

First consider the way flow is affected by speed differences between vehicles at different points in the traffic stream. Figure 1 shows the trajectories of four hypothetical vehicles as they pass two fixed points (Stations A and B) that are assumed to be entirely up- and downstream of a shock wave. These points are separated by a distance Δx. It is assumed that there are no on- or off-ramps in the vicinity and that each pair of trajectories is separated by a fixed number ΔN of vehicles at each point in the traffic stream. The diagram can be applied to situations in which passing occurs by adopting the device suggested by Makigami et al. (*27*): whenever passing occurs, the vehicles are renumbered and the trajectories rebound. The wave is assumed to begin after the passage of the first vehicle and to dissipate after the passage of the last one. The dashed lines indicate the average speeds between A and B for trajectories 2 and 3.

Now consider the relationship between changes in speed between different vehicles in the traffic stream and changes in flow over time and space. Because it simplifies the algebra, the relationship is actually defined in terms of the reciprocals of speed and flow. The reciprocal of flow is the average time headway; let $h_A = 1/q_A$ and $h_B = 1/q_B$ be the headways at A and B, respectively, where q_A and q_B are the flows. Following the convention of Vaughan et al. (*28*), the reciprocal of speed is referred to as "tardity" and is designated by Λ; Λ_1 refers to the reciprocal of the average speed of Vehicle 1 between points A and B.

From the definition of headway, $h_A = t_A/\Delta N$ and $h_B = t_B/\Delta N$, where t_A and t_B are the times separating some pair of trajectories (say 1 and 2) at A and B, respectively. From the diagram,

$$t_B = t_A + \Delta x(\Lambda_2 - \Lambda_1) \tag{1}$$

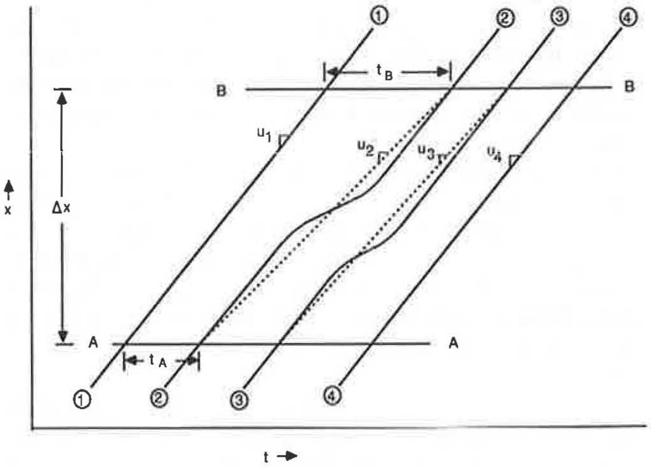

FIGURE 1 Hypothetical vehicle trajectories.

from which

$$h_B = h_A + (\Delta x/\Delta N)\Delta\Lambda \qquad (2)$$

or

$$\Delta h/\Delta x = \Delta\Lambda/\Delta N \qquad (3)$$

Equation 3 may appear in a more familiar light if an alternative derivation is considered. Figure 2 shows the cumulative number of vehicles passing points A and B as a function of time, beginning with the arrival of some particular vehicle in each case. The horizontal dimension of the graph is the time it takes a given vehicle to travel from A to B, and the vertical difference between the cumulative arrival curves is the number of vehicles stored between A and B at any time. The average slopes of the cumulative arrival curves are the average flow rates q_A and q_B. As Makigami et al. (*27*) point out, Figures 1 and 2 are actually equivalent representations of the traffic stream.

Consider Vehicles 1 and 2, separated by ΔN vehicles, where points A and B are once again separated by a distance Δx. The two vehicles travel from A to B in times $\Delta x\Lambda_1$ and $\Delta x\Lambda_2$, respectively, so from the diagram

$$\Delta x\Lambda_2 = \Delta x\Lambda_1 + \Delta N(1/q_B - 1/q_A) \qquad (4)$$

or

$$\Delta x\Delta\Lambda = \Delta N(h_B - h_A) \qquad (5)$$

Equation 3 follows from replacing $h_B - h_A$ by Δh and rearranging.

Equation 3 relates a change in flow over distance to a difference in average speed between two trajectories. In and of itself, it does not necessarily predict a decrease in flow at the downstream location when the speed drops; it only predicts a decrease in flow at B relative to that at A. As such (as is clear from the derivation from Figure 2), it might represent no more than the decrease in average speed across the section between A and B that occurs as a queue builds up in the section because flow at A exceeds the capacity of some point downstream.

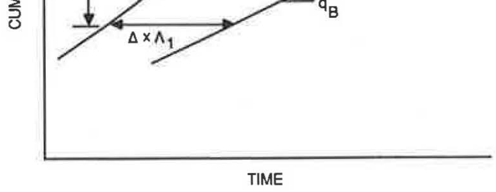

FIGURE 2 Hypothetical cumulative arrival curves.

On the other hand, consider the case in which there is an isolated speed disturbance of the sort described earlier (that is, no secondary shock waves form in the accelerating flow) and suppose that flow was steady state, so that in the absence of the disturbance, there would be only random variations in speeds and flows between points A and B. In that case, if a nonrandom difference in flow (caused by the speed disturbance) should occur between A and B, it should result in a nonrandom change in flow over time at B. Moreover, in this case the queue is only the core of the speed disturbance, and quickly reaches a stable length. However, at the beginning of the disturbance there is a decrease in speed, which creates a decrease in flow. If flow immediately downstream from the point of the disturbance is measured over a short enough time interval, it should be possible to detect this decrease.

However, given the data available in this study, it is almost impossible to detect any such effect. The minimum-count interval is 30 sec; the flows in question are on the order of 18 to 20 vehicles per lane per 30 sec; and the decreases in speed do not take place instantly: usually they take 1 min or more. The consequence is that $\Delta\Lambda/\Delta N$ is a rather small number, and the change in flow that would result from it gets lost in the random variation in the 30-sec counts.

INCREASE IN PASSAGE TIME

The theory outlined presents a second problem. As can be seen from Equation 3 and Figure 1, the decrease in flow should persist only so long as there is a change in average speed over the section. In the case of an isolated wave in steady-state flow, this change in speed should occur fairly rapidly; once it is accomplished, $\Delta\Lambda/\Delta N = 0$, there is no change in flow between A and B, and the flow rate should recover. In fact, the decrease in flow tended to persist until speeds recovered at the location in question. This suggests that most of the flow decrease was the result of a direct relationship between speed and flow, such as would result from increased passage times.

Figure 3 shows the effect of speed changes on passage times and, hence, time headways. Two successive vehicle trajectories are indicated. Front and rear trajectories are shown for each vehicle, with the shaded area representing the space occupied by the vehicle itself. Dashed lines are used, as in Figure 1, to show the average speeds of the fronts of the two vehicles between different points. In the diagram, it is assumed that time gaps do not vary with speed; given this assumption, although flow must decrease as speeds decrease, a subsequent increase in speed leads to an increase in flow.

The alternative, of course, is that the time gaps also increase as speed decreases. In this case, it turns out to be fairly easy to use the data available to distinguish between these two possibilities. Data include flows and occupancies. The average time headway is the reciprocal of the flow; meanwhile, because occupancy represents the aggregate time during which vehicles are over the detector, one minus occupancy represents the aggregate time during which vehicles are not present,

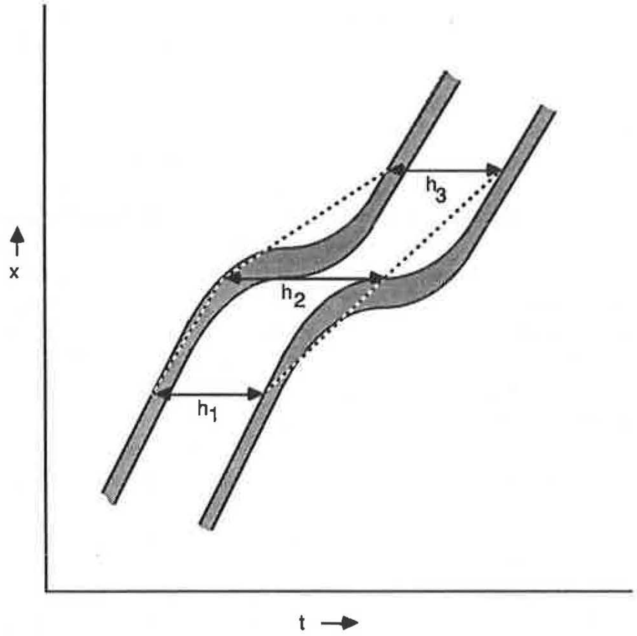

FIGURE 3 Effect of passage time on time headway.

and if this is divided by the flow, the result is the average time gap.

If occupancy were literally measured at a point, this would be strictly true; in fact, the detector length is significant relative to the average vehicle length. In San Diego, the physical length of the detectors is 10 ft; however, their effective length is somewhat less (because of differences in physical and electrical lengths, scanning rates, sensitivities, and the like). Based on the effective lengths that make speed calculations come out right (Caltrans uses 24.75 ft for the effective length of the vehicle plus the detector in San Diego) and a guess as to the actual mean vehicle length (somewhere between 17 and 20 ft), the average effective length of the detector is on the order of 5 to 8 ft.

The actual calculation of the average gap then is

$$g = (1 - H)/q + d/u \tag{6}$$

where

g = time gap,
H = occupancy,
d = effective length of the detector, and
u = average speed.

But because

$$u = [q(L + d)]/H \tag{7}$$

where L is the average vehicle length, Equation 6 can be simplified to

$$g = \{1 - [HL/(L + d)]\}/q \tag{8}$$

This implies a correction to H, which varies only with the average vehicle length, and not with speed. Even if the ratio of $L/(L + d)$ assumed is incorrect, Equation 8 correctly distinguishes increases or decreases in gaps when two time intervals are compared, unless there are significant changes in the average vehicle length.

Time headways and time gaps were calculated for the left lane for 12-min periods before and after flow breakdown at each of the bottlenecks studied. At three of the four, there appeared to be no significant change in the time gap when periods before and after flow breakdown were compared. When data for individual days were compared, sometimes the average gap increased and sometimes it decreased, but the total number of increases and decreases was roughly equal, and the average difference for all days considered was near zero (between −0.03 sec and +0.03 sec). At the fourth site, gaps decreased in 15 out of 17 cases and the average decrease was 0.13 sec, or about 10 percent.

From this, it may be concluded that the tendency for time headways to decrease when flow broke down at these sites resulted from the increase in passage time rather than increased time gaps. It may also be concluded that drivers were at or near their minimum acceptable time gaps before flow broke down at three of the four sites. Thus it appears that the reduced flow downstream from these bottlenecks is largely the result of the increase in passage time that occurs when speeds decrease.

SHOCK WAVE STRUCTURE

In the preceding section, it was pointed out that when time gaps are insensitive to speed, changes in speed imply changes in flow because time headways are merely a function of the passage time. Also, it was found that the relationship is valid for both acceleration and deceleration. It can be further demonstrated that shock waves of the sort observed at the San Diego bottlenecks imply a similar relationship between flow and speed.

Recall that these waves consisted of a small dense core, in which vehicles were stopped or nearly stopped, with a zone of deceleration upstream and a zone of acceleration downstream. All features of the wave move upstream over time. The equation for shock wave speed, originally introduced by Lighthill and Withem (*29*), is based on the conservation of vehicles. In its discrete form, it is

$$u_w = \Delta q/\Delta k \tag{9}$$

where

u_w = speed of the wave,
Δq = change in flow across the wave, and
Δk = change in density across the wave.

In order for the wave to move upstream, the signs of Δq and Δk must be opposite.

In the case being considered, there are actually two shock waves: one between the low-density traffic upstream of the disturbance and the higher-density core, and another between the core and the lower-density traffic downstream. In order for the wave to move upstream, there must be a decrease in flow between the traffic upstream and the core and an increase in flow between the core and the traffic downstream. From the argument in the preceding section, the wave motion might represent no more than the effect of the acceleration and deceleration, although in some cases time gaps might also vary across the wave.

SPEED INSTABILITY AND CAR-FOLLOWING

Flow breakdown at the San Diego bottlenecks appears to have resulted from instability in speeds. Car-following models predict that this sort of instability in speed will develop under certain circumstances and may be useful in explaining why it occurs. Models of the type described by Payne (*14*) and others (*21–26*) possess two key features that can be verified by means of average time gap data, such as that calculated for the San Diego bottlenecks.

The first of these features is that each microscopic car-following model implies a macroscopic model relating speed to flow or density (*22–24*). These relationships, in turn, can be expressed in terms of the relationship between speeds and time gaps.

At three of the four San Diego bottlenecks, time gaps appear to have been unaffected by flow breakdown. Macroscopic flow and concentration data from one of these sites were previously found to imply constant average time gaps throughout the range of congested flow (including incident queues that were considerably denser than anything included in the present study) (*17*); unpublished data at other San Diego bottlenecks appear to be similar. Such constant average time gaps in congested flow are consistent with the linear model of Chandler et al. (*21*).

This model states that the acceleration of the trailing vehicle at time $t + \Delta$ (where Δ is the reaction time) is a linear function of the difference in speed between the lead vehicle and the trailing vehicle. Mathematically,

$$a_2(t + \Delta) = \lambda[u_1(t) - u_2(t)] \tag{10}$$

where a_2 is the acceleration of the trailing vehicle, u_1 and u_2 are the speeds of the leading and trailing vehicle, and λ is a constant sensitivity factor. The equivalent macroscopic model for equilibrium flow conditions may be obtained by ignoring the reaction time, integrating, and substituting the appropriate boundary conditions (*22*). Integration results in

$$u = \lambda(x_1 - x_2) + C \tag{11}$$

where u is average speed and x_1 and x_2 are the positions of the two vehicles; that is, speed is a linear function of the average distance separation. Setting $u = 0$ for $x_1 - x_2 = x_0$, where x_0 is some minimum distance separation,

$$u = \lambda(x_1 - x_2 - x_0) \tag{12}$$

If x_0 is assumed equal to the average vehicle length, solving for λ and taking the reciprocal implies

$$1/\lambda = [(x_1 - x_2) - L]/u = g \tag{13}$$

where g, as before, is the time gap. Even if drivers are assumed to allow some constant distance buffer, so $x_0 = L + c$, it is still true that

$$u = (1/g)(x_1 - x_2 - L) \tag{14}$$

so that the linear car-following model implies constant time gaps for the range of flow conditions to which it applies.

Clearly, this model only applies to congested flow; if distance separations are allowed to increase without limit, it predicts infinite speed. It was the original assumption of Chandler et al. (*21*) that it only applied once distance separations declined to a certain critical value; if flow is to break down, this value must place the model at or beyond its stability limit at the point at which interaction begins.

The stability criterion itself is the second point at which the model can be compared with the data. In the case of the linear model, it is

$$\lambda < \frac{1}{2}\Delta \tag{15}$$

Given the interpretation of λ developed earlier, this relation implies that drivers must maintain time gaps that are at least twice their reaction times.

In the case of the San Diego bottlenecks, speed instability did develop, but it was fairly rare, compared with the total number of vehicle platoons observed in high-volume uncongested flow. However, at some of the locations, the probability of speed instability appeared to be somewhat higher in the accelerating flow downstream of the first wave. From this, one might conclude that the actual flow was relatively stable (at least before initial breakdown). The probability of instability for any given vehicle pair or even any large platoon was small but cumulatively significant; meanwhile, the probability of a collision (the sort of instability most often discussed in the car-following literature) was almost infinitesimal compared with the total number of vehicle interactions.

When the average time gaps computed for the San Diego bottlenecks are compared with reaction times commonly reported, the linear model in its unmodified form is not nearly stable enough. When an average vehicle length of 17 ft was assumed, average time gaps in the left lane for individual 12-min intervals ranged from 1.00 to 1.73 sec. Averages over all days for different bottlenecks ranged from 1.1 to 1.4 sec. When an average vehicle length of 20 ft was assumed, the corresponding estimates decreased by about 0.04 sec.

Given that flow appeared to be at least marginally stable, this would imply average reaction times of 0.5 to 0.9 sec. The experimental work reported by Chandler et al. (*21*), which was used in the calibration of the various models developed at General Motors, suggested an average value for the reaction time as high as 1.5 sec. Hurlbert (*30*) indicated that the median of experimentally measured reaction times was 0.66 sec when the stimulus was expected and 0.9 sec when it was not. In either case, it is hard to reconcile the relatively high level of stability observed with the stability characteristics of the model.

It turns out that difficulty in reconciling stability criteria with observed time headways at maximum volumes is a problem with most car-following models. It has previously been considered by Bexelius (*26*), who extended the linear model to allow for sensitivity to more than the first vehicle ahead. Bexelius (*26*) derived the stability criterion for a linear model in which the driver of the trailing vehicle reacts to the speeds of the two preceding vehicles. It can be shown that by this device, the form of the model is not affected (so that it still predicts constant time gaps in congested flow), but the stability is increased, allowing stability to be maintained with smaller gaps.

It appears, then, that the linear car-following model in its extended form is consistent with the time gap evidence at three of the San Diego sites. There remain, of course, several problems with it. First, all the car-following models in the literature appear to be oversimplified when considered as models of human behavior. This fact is particularly true of the linear model, and is one of the reasons the group that developed it moved on eventually to other models. Unfortunately, the mathematics involved in stability analysis of existing car-following models is already quite complicated, and there is little prospect that more realistic models would prove tractable.

Second, Trieterer and Myers (*10*) found a definite pattern in speeds and vehicle spacings when platoons of vehicles pass through shock waves. This pattern is somewhat more complicated than what should result from the linear model. They also proposed separate models for the acceleration and deceleration process and fit them to speed and density data. In neither case was the best-fit model the linear one.

Third, the model is deterministic, but the behavior in question is clearly stochastic, with wide ranges of random variation being typical of such variables as vehicle spacing. It remains to be shown that a stochastic version of the linear model would predict the same average behavior, particularly with regards to stability.

Fourth, observation of flow at the San Diego sites left the distinct impression that instability was more likely in accelerating flow downstream of the initial wave than in the high-volume flow immediately before breakdown. There is nothing in the linear model as developed so far that would predict this. It is possible that a stochastic model would shed some light on this point, because it appeared that although the mean of the distribution of time gaps was not affected significantly by flow breakdown, the variance may well have been (in particular, the large gaps in front of vehicles leading platoons tend to disappear).

Finally, at one site, time gaps did decrease when flow broke down. This might imply that a different model would be appropriate at this site. On the other hand, the key characteristic of the linear model is that time gaps not vary with speed in congested flow. At the first three sites, it also appears that time gaps were already at the minimum that drivers would tolerate before flow breakdown. It may be that under some circumstances, gaps in free flow do not reach the minimum drivers will tolerate before speed stability sets in and that in other cases they do. A similar suggestion was made by Wasielewski (*31*) in a study of time headway distributions.

CONCLUSION

Theoretical issues related to two phenomena observed in a study of four freeway bottlenecks in San Diego have been addressed. These issues were that flow immediately downstream of the bottlenecks tended to decrease by a small amount when it broke down, and that the breakdown process seemed to be triggered by speed instability. Most of the flow decrease appeared to be caused by the increase in vehicle passage time that occurred when speed decreased, and the increase in passage time was related to the structure of the shock waves that were observed. Most of the San Diego data were compatible with the linear car following model of Chandler et al. (*21*) as extended by Bexelius (*26*), although a number of questions about the validity and applicability of this model remain.

ACKNOWLEDGMENTS

This research was funded by the State of California Department of Transportation. Special thanks are due to Karel Shaffer and Michael Pellino.

REFERENCES

1. J. H. Banks. Flow Processes at a Freeway Bottleneck. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.
2. J. H. Banks. *Evaluation of the Two-Capacity Phenomenon as a Basis for Ramp Metering*. Final Report. Civil Engineering Report Series 9002, San Diego State University, San Diego, Calif., 1990.
3. V. F. Hurdle and P. K. Datta. Speeds and Flows on an Urban Freeway: Some Measurements and a Hypothesis. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 127–137.
4. B. N. Persaud. *Study of a Freeway Bottleneck to Explore Some Unresolved Traffic Flow Issues*. Ph.D. dissertation, Department of Civil Engineering, University of Toronto, Toronto, Canada, 1986.
5. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
6. L. Newman. Traffic Operation at Two Interchanges in California. In *Highway Research Record 167*, HRB, National Research Council, Washington, D.C., 1963, pp. 14–43.
7. L. C. Edie and R. S. Foote. Traffic Flow in Tunnels. *HRB Proc.*, Vol. 37, 1958, pp. 334–344.
8. L. C. Edie and R. S. Foote. Effect of Shock Waves on Tunnel Traffic Flow. *HRB Proc.*, Vol. 39, 1960, pp. 492–505.
9. T. W. Forbes and M. E. Simpson. Driver-and-Vehicle Response in Freeway Deceleration Waves. *Transportation Science*, Vol. 2, No. 1, 1968, pp. 77–104.
10. J. Trieterer and J. A. Myers. The Hysteresis Phenomenon in Traffic Flow. *Proc., 6th International Symposium on Transportation and Traffic Theory*, D. J. Buckley (ed.), American Elsevier, New York, 1974, pp. 13–38.
11. P. J. Athol and A. G. R. Bullen. Multiple Ramp Control for a Freeway Bottleneck. In *Highway Research Record 456*, HRB, National Research Council, Washington, D.C., 1973, pp. 50–54.
12. L. C. Edie. Car-Following and Steady-State Theory for Non-Congested Traffic. *Operations Research*, Vol. 9, No. 1, Baltimore, Md., 1961, pp. 61–76.
13. J. Drake, J. Shofer, and A. May. A Statistical Analysis of Speed-Density Hypotheses. In *Highway Research Record 154*, HRB, National Research Council, Washington, D.C., 1967, pp. 53–87.
14. H. J. Payne. Discontinuity in Equilibrium Freeway Traffic Flow. In *Transportation Research Record 971*, TRB, National Research Council, Washington, D.C., 1984, pp. 140–146.
15. F. L. Hall, B. L. Allen, and M. A. Gunter. Empirical Analysis of Freeway Flow-Density Relationships. *Transportation Research*, Vol. 20A, No. 3, Elmsford, N.Y., 1986, pp. 197–210.
16. F. L. Hall. An Interpretation of Speed-Flow-Concentration Relationships Using Catastrophe Theory. *Transportation Research*, Vol. 21A, No. 3, Elmsford, N.Y., 1987, pp. 191–201.
17. J. H. Banks. Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 53–60.

18. H. S. Mika, J. B. Kreer, and L. S. Yuan. Dual-Mode Behavior of Freeway Traffic. In *Highway Research Record 279*, HRB, National Research Council, 1969, pp. 1–12.
19. T. Lam and R. Rothery. The Spectral Analysis of Speed Fluctuations on a Freeway. *Transportation Science*, Vol. 4, No. 3, 1970, pp. 293–310.
20. R. D. Kuhne. A Macroscopic Freeway Model for Dense Traffic—Stop-Start Waves and Incident Detection. *Proc. 9th International Symposium on Transportation and Traffic Theory*, J. Volmuller and R. Hamerslag (eds.), VNU Science Press, Utrecht, Netherlands, 1984, pp. 21–42.
21. R. E. Chandler, R. Herman, and E. W. Montroll. Traffic Dynamics: Studies in Car Following. *Operations Research*, Vol. 6, No. 2, Baltimore, Md., 1958, pp. 165–184.
22. R. Herman, E. W. Montroll, R. D. Potts, and R. W. Rothery. Traffic Dynamics: An Analysis of Stability in Car Following. *Operations Research*, Vol. 7, No. 1, Baltimore, Md., 1959, pp. 86–106.
23. D. C. Gazis, R. Herman, and R. B. Potts. Car Following Theory of Steady-State Flow. *Operations Research*, Vol. 7, No. 4, Baltimore, Md., 1959, pp. 499–505.
24. D. C. Gazis, R. Herman, and R. W. Rothery. Nonlinear Follow-the-Leader Models of Traffic Flow. *Operations Research*, Vol. 9, No. 4, Baltimore, Md., 1961, pp. 545–567.
25. G. F. Newell. Nonlinear Effects in the Dynamics of Car Following. *Operations Research*, Vol. 9, No. 2, Baltimore, Md., 1961, pp. 209–229.
26. S. Bexelius. An Extended Model for Car-Following. *Transportation Research*, Vol. 2, No. 1, Elmsford, N.Y., 1968, pp. 13–21.
27. Y. Makigami, G. F. Newell, and R. Rothery. Three-Dimensional Representation of Traffic Flow. *Transportation Science*, Vol. 5, No. 3, Baltimore, Md., 1971, pp. 320–313.
28. R. Vaughan, V. F. Hurdle, and E. Hauer. A Traffic Flow Model with Time-Dependent O–D Patterns. *Proc., 9th International Symposium on Transportation and Traffic Theory*, J. Volmuller and R. Hamerslag (eds.), VNU Science Press, Utrecht, Netherlands, 1984, pp. 155–178.
29. M. J. Lighthill and G. B. Witham. On Kinematic Waves: II. A Theory of Traffic Flow on Long Crowded Roads. *Proc., Royal Society, London*, Series A, Vol. 229, No. 1178, 1955, pp. 317–345.
30. S. Hurlbert. Driver and Pedestrian Characteristics. *Transportation and Traffic Engineering Handbook*, J. E. Baerwald (ed.), Prentice-Hall, Englewood Cliffs, N.J., 1976, pp. 38–72.
31. P. Wasielewski. An Integral Equation for the Semi-Poisson Headway Distribution Model. *Transportation Science*, Vol. 8, No. 3, 1974, pp. 237–247.

The contents of this paper reflect the views of the author, who is responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the state of California. This report does not constitute a standard, specification, or regulation.

Publication of this paper sponsored by Committee on Traffic Flow Theory and Characteristics.