TRANSPORTATION RESEARCH RECORD 1005

Highway Capacity, Traffic Characteristics, and Flow Theory

TRB
TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL
WASHINGTON, D.C 1985

Transportation Research Record 1005
Price $16.60
Editor: Elizabeth W. Kaplan
Compositor: Lucinda Reeder
Layout: Betty L. Hawkins

mode
1 highway transportation

subject area
55 traffic flow, capacity, and measurements

Transportation Research Board publications are available by ordering directly from TRB. They may also be obtained on a regular basis through organizational or individual affiliation with TRB; affiliates or library subscribers are eligible for substantial discounts. For further information, write to the Transportation Research Board, National Research Council, 2101 Constitution Avenue, N.W., Washington, D.C. 20418.

Printed in the United States of America

Library of Congress Cataloging in Publication Data
National Research Council. Transportation Research Board.
Highway capacity, traffic characteristics, and flow theory.

(Transportation research record; 1005)
1. Highway capacity—Congresses. 2. Traffic flow—Congresses. I. National Research Council (U.S.). Transportation Research Board. II. Series.
TE7.H5 no. 1005 380.5 s 85-13808
[HE336.H48] [388.3'14]
ISBN 0-309-03815-4 ISSN 0361-1981

Sponsorship of Transportation Research Record 1005

GROUP 3—OPERATION, SAFETY, AND MAINTENANCE OF TRANSPORTATION FACILITIES
D. E. Orne, Michigan Department of Transportation, Chairman

Committee on Highway Capacity and Quality of Service
Carlton C. Robinson, Highway Users Federation, chairman
Charles W. Dale, Federal Highway Administration, secretary
Donald S. Berry, Robert C. Blumenthal, James B. Borden, Fred W. Bowser, V. F. Hurdle, James H. Kell, Frank J. Koepke, Jerry Kraft, Walter H. Kraft, Joel P. Leisch, Adolf D. May, Jr., William R. McShane, Carroll J. Messer, Guido Radelat, Huber M. Shaver, Jr., Alexander Werner, Robert H. Wortman

Committee on Traffic Flow Theory and Characteristics
John J. Haynes, University of Texas at Arlington, chairman
Edmund A. Hodgkins, Consultant, secretary
Patrick J. Athol, E. Ryerson Case, Kenneth W. Crowley, Peter Davies, John W. Erdman, Nathan H. Gartner, Richard L. Hollinger, Matthew J. Huber, Edward Lieberman, Feng-Bor Lin, C. John MacGowan, Hani S. Mahmassani, Patrick T. McCoy, Carroll J. Messer, Panos G. Michalopoulos, Harold J. Payne, A. Essam Radwan, Paul Ross, Robert M. Shanteau, Steven R. Shapiro, Sam Yagar

David K. Witheford, Transportation Research Board staff

Sponsorship is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 1984.

NOTICE: The Transportation Research Board does not endorse products or manufacturers. Trade and manufacturers' names appear in this Record because they are considered essential to its object.

Contents

Addresses of Authors

Allen, Brian L., Department of Civil Engineering, McMaster University, Hamilton, Ontario L8S 4L7 Canada
Chang, Edmond Chin-Ping, Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843
Chang, Gang-Len, Department of Civil Engineering, University of Texas at Austin, Austin, Tex. 78712
Clark, J. Edwin, Department of Civil Engineering, Clemson University, Clemson, S.C. 29631
Croft, Frank M., Jr., Department of Engineering Graphics, The Ohio State University, Columbus, Ohio 43210
Golenberg, Marvin, SG Associates, Inc., 316 Stuart Street, Boston, Mass. 02116
Grote, Wulf, Division of Design and Construction, Regional Transportation District, 1600 Blake Street, Denver, Colo. 80202
Gunter, Margot A., Department of Civil Engineering, McMaster University, Hamilton, Ontario L8S 4L7 Canada
Gynnerstedt, G., Swedish National Road and Traffic Institute, 581 01 Linkoping, Sweden
Hall, Fred L., Department of Civil Engineering, McMaster University, Hamilton, Ontario L8S 4L7 Canada
Herman, Robert, Department of Civil Engineering and Center for Studies in Statistical Mechanics, University of Texas at Austin, Austin, Tex. 78712
Howard, Jane, SG Associates, Inc., 316 Stuart Street, Boston, Mass. 02116
Kalevela, Sylvester A. F., Engineering Department, Dar-es-Salam, Tanzania
Levinson, Herbert S., Department of Civil Engineering, University of Connecticut, Storrs, Conn. 06268
Lin, Feng-Bor, Department of Civil and Environmental Engineering, Clarkson University, Potsdam, N.Y. 13676
Mahmassani, Hani S., Department of Civil Engineering, University of Texas at Austin, Austin, Tex. 78712
Marsden, Blair G., Texas State Department of Highways and Public Transportation (File D-18T), P.O. Box 5051, Austin, Tex. 78763
Messer, Carroll J., Transportation Institute, Texas A&M University, College Station, Tex. 77843
Palaniswamy, S. P., Indian Institute of Technology, Kanpur, India
Phull, Y. R., Central Road Research Institute, New Delhi, India
Radwan, A. Essam, Civil Engineering Department, Arizona State University, Tempe, Ariz. 85287
Roskin, Mark E., Comsis Corporation, 11501 Georgia Avenue, Wheaton, Md. 20902
Smith, Steven A., JHK & Associates, 4660 Kenmore Avenue, Alexandria, Va. 22304
Staniewicz, James M., New Haven Parking Authority, One Temple Street, New Haven, Conn. 06510-3288
Teply, S., Department of Civil Engineering, University of Alberta, Edmonton, Alberta T6G 2G7 Canada
Williams, James C., Department of Civil Engineering, University of Texas at Austin, Austin, Tex. 78712
Yu, Jason C., Department of Civil Engineering, University of Utah, Salt Lake City, Utah 84112

Callahan Tunnel Capacity Management

HERBERT S. LEVINSON, MARVIN GOLENBERG, and JANE HOWARD

ABSTRACT

The Callahan Tunnel capacity management actions that were implemented beginning May 1983 in conjunction with one-way inbound toll collection are described. Traffic operations and capacities for both "before" and "after" conditions are analyzed. The analyses of existing Callahan peak-hour traffic volumes, flows through the parallel Sumner Tunnel, and volumes in the four New York City tunnels provided a basis for estimating tunnel and tunnel-system capacity. The analyses suggested a tunnel capacity of 1,600 to 1,650 vehicles per lane per hour compared with average peak-hour system volumes of 1,450 vehicles per lane. Channelization of the tunnel approach, elimination of outbound toll collection, and realignment of the tunnel exit lanes were estimated to increase system throughput by about 350 to 400 vehicles per hour--up to 200 vehicles per lane. "After" studies conducted during June 1983 indicated that flow rates of more than 1,600 vehicles per lane were achieved. The plan is significant in another respect--speed of implementation. Improvement concepts were formulated during March and April 1983, and their implementation began during May of that year.

The opening of the Sumner Tunnel in 1934 as a two-lane, two-way facility created the first significant and direct automobile link between East Boston and North Shore communities and the city of Boston and communities lying to the south of the Charles River and Boston Harbor. Traffic demands steadily increased, reflecting the accessibility created by the tunnel, growth in North Shore communities, and growth of the Logan Airport complex. To meet this demand, harbor crossing capacity was increased in 1962 by building the parallel two-lane Callahan Tunnel. At that time, the Sumner Tunnel was converted to one-way westbound flow from East Boston to the Boston central business district (CBD).

BACKGROUND

Peak-hour traffic across the harbor continued to increase to the point where demand exceeded capacity because of toll collection, tunnel geometry, and the configuration and traffic control of access roads. This resulted in a lengthened peak period, increased congestion on streets adjacent to the tunnel approaches, and increased cross-harbor travel times. Resulting impacts were different geographically. Sumner Tunnel impacts were primarily felt on the East Boston side, including impacts on movement from Logan Airport to the Boston CBD and other areas south and west of Boston Harbor. Callahan Tunnel impacts were predominantly in the Boston CBD on surface streets and the Central Artery, often significantly affecting nontunnel traffic flows on these facilities. In addition, the Callahan Tunnel limited the reliability and quality of vehicle access to the Logan Airport complex.

Various agencies had long recognized that, without improvements in cross-harbor capacity, congestion problems would become increasingly severe. In response to these concerns, several studies were made of ways to increase cross-harbor capacity ranging from long-term major capital actions (third harbor crossing) to low-cost, short-term transportation systems management (TSM) approaches. The management action that received the greatest attention was to collect tolls only in the inbound direction (Sumner Tunnel) in coordination with similar one-way inbound toll operations on the Tobin Bridge (Mystic River Bridge). Underlying objectives were to (a) increase tunnel throughput, (b) reduce the amount and duration of queueing at the tunnel approach, and (c) reduce the tunnel journey time.

A report prepared by the Central Transportation Planning Staff in 1980 identified those factors that inhibit maximum flow potential and that should be alleviated through the one-way toll operation and TSM actions. A subsequent Massachusetts Port Authority staff report on Callahan Tunnel operations (March 1982) estimated that as much as a 14 percent increase in peak-hour throughput traffic volume is possible with elimination of toll collection, removal of the Callahan Tunnel toll booths, and traffic management improvements on the tunnel approach and within the tunnel.

SCOPE AND APPROACH

The Callahan Tunnel capacity management actions that were implemented beginning in May 1983, in conjunction with one-way inbound toll collections, are described. The preimplementation conditions (March, April 1983) are reviewed; the likely effects of improved traffic operations are analyzed; and before and after traffic operations are compared. More complete discussions of the feasibility and follow-up analyses are presented elsewhere (1,2).

Data obtained from the Massachusetts Turnpike Authority (MTA), Massachusetts Port Authority, and other public agencies were supplemented by videotape monitoring of peak-hour traffic on the tunnel approach and the toll plaza.

The analyses considered the Callahan Tunnel system as containing three interrelated operating sections: the approach area to the west portal, the tunnel itself, and the toll area from the east portal through the toll booth. The aim was to determine the maximum capacity for each of the three sections, identify the points of minimum capacity, and suggest corrective actions to increase system throughput.

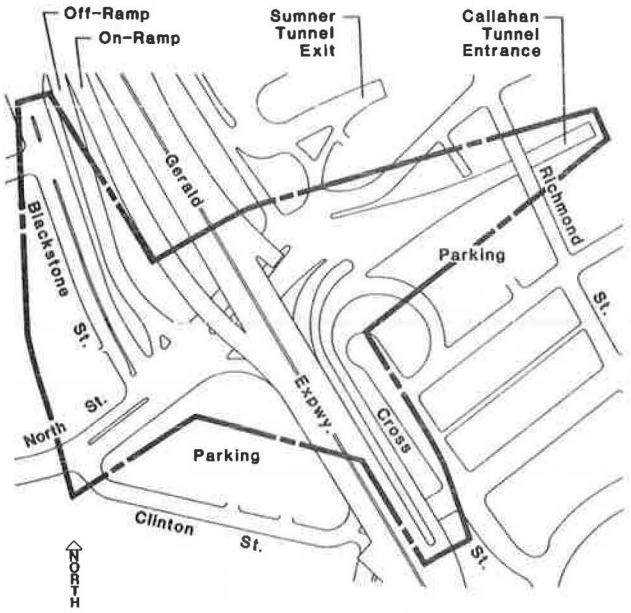


FIGURE 1 Study area boundaries—Callahan Tunnel approach area in downtown Boston.

INITIAL PHYSICAL AND TRAFFIC CONDITIONS

Physical and traffic characteristics of the Callahan Tunnel and its environs before implementation in May 1983 are discussed in the following subsections.

Physical and Geometric Features

The Callahan Tunnel contains two lanes 10 ft 6 in. wide, with a 1-ft lateral clearance. There are three vertical sections: west portal to end of descent, end of descent to beginning of ascent, and beginning of ascent (3.25 percent, 2,300-ft upgrade) to east portal.

The approach area in downtown Boston is shown in Figure 1. It extends from the intersection of Blackstone Street and North Street to the west to the tunnel portal and includes the Central Artery northbound off-ramp, the Surface Artery northbound lanes, and the North Street lanes that feed the tunnel. Before implementation, eight lanes of traffic merged into the two tunnel lanes as follows:

- Northbound Artery off-ramp--24 ft operating as two traffic lanes,
- Northbound Surface Artery--one 17-ft roadway operating as two traffic lanes, and
- North Street--a 43-ft roadway operating as three or four 10- or 11-ft lanes (including tunnel-bound traffic from North Street, Blackstone Street, and the Southbound Central Artery off-ramp).

The tunnel portal walls and the Central Artery defined the limits of the physical area within which actions could be developed. The preimplementation physical and operational features and constraints within the approach area are shown in Figure 2.

The east tunnel portal and its environs, as of March-April 1983, are shown in Figure 3. The toll plaza contained seven booths, five manual and two automatic, before toll collection was discontinued. Design of the plaza area was constrained by the presence of seven Sumner Tunnel westbound toll booths to the north and the portal walls to the south.

The distance from the East Boston portal of the Callahan Tunnel to the center of the toll plaza, measured along the right edge of the pavement, was approximately 710 ft. This section of the alignment consisted of a curve 337.43 ft long having a radius of 2,000 ft and a curve 183.73 ft long with a radius of 330 ft connected by a 9-ft tangent section. Less than 200 ft beyond the toll booths, Route 1A, a three-lane roadway, continues north with a branch to the airport. South of and adjacent to these two roadways are Havre Street and Porter Street, two local one-way streets that were relocated and, in part, removed to accommodate the toll plaza.

Traffic traveling through the Callahan Tunnel would continue through the toll plaza and onto Route 1A or turn sharply to the right and onto Porter Street. Havre Street traffic could turn right onto Porter Street or continue onto 1A. Havre Street is the only access to Route 1A for the surrounding area. The intersection between the tunnel traffic turning onto Porter Street and the Havre Street traffic turning onto Route 1A was controlled by a stop sign on Havre Street.

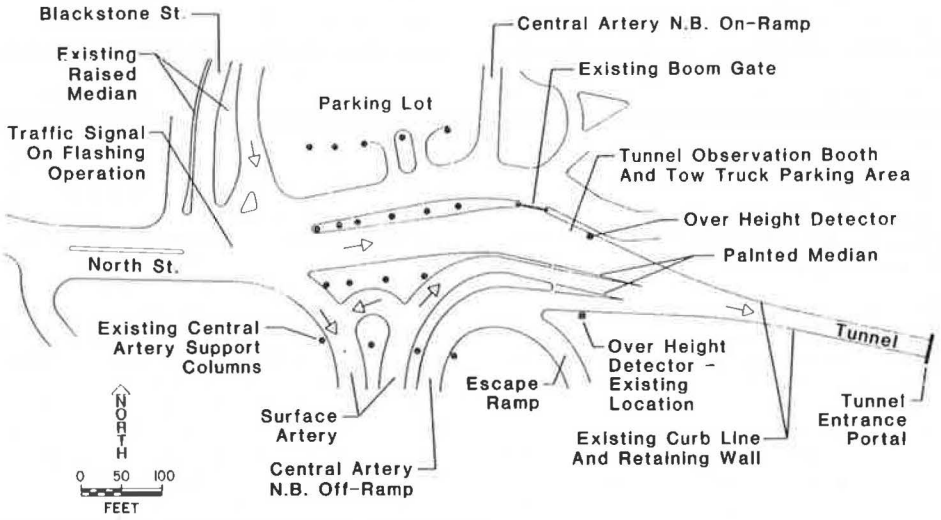


FIGURE 2 Approach area conditions, April 1983.

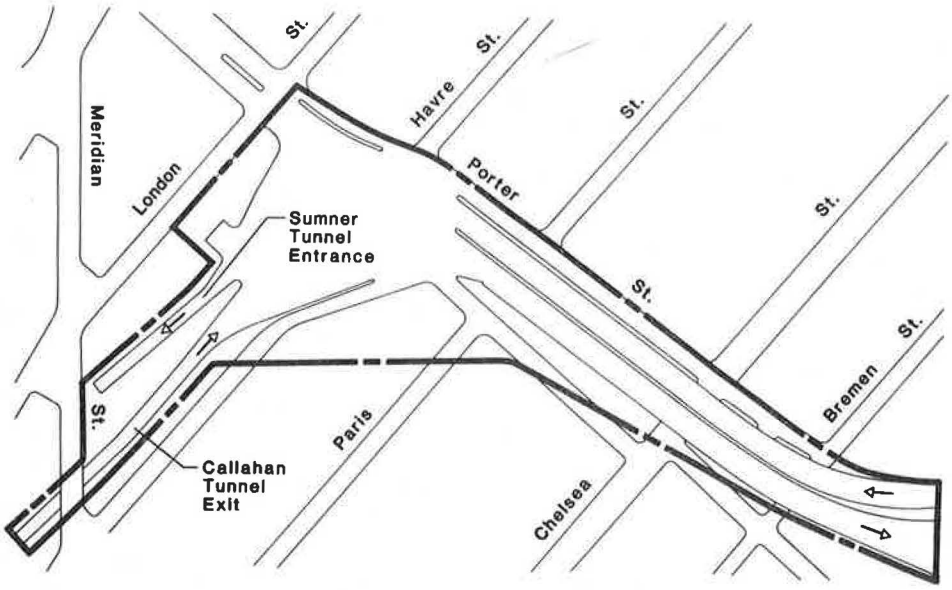


FIGURE 3 Study area boundaries—Callahan tunnel toll plaza area in East Boston.

"Before" Traffic Volumes

Table 1 gives average day and afternoon peak-hour volumes for the Callahan Tunnel taken from 1982 Massachusetts Turnpike Authority hourly counts. These "before" counts show an average daily traffic volume of 38,667 and an average afternoon peak-hour volume of 2,805 vehicles. Annual average daily traffic was 38,049 in 1982. Trucks and buses averaged 2.4 percent of the total peak-hour flow.

Table 2 gives peak-hour volume characteristics for October 1982 taken from toll station counts provided by the Massachusetts Turnpike Authority. The average volume was 2,794 vehicles, or 1,347 per lane if traffic were equally distributed; peak-hour tunnel traffic exceeded 3,000 vehicles for less than 10 percent of the time.

TABLE 1 Average Daily and Peak-Hour Tunnel Volumes, Callahan Tunnel, 1982

Day	Date	24-Hr Volume	Peak-Hour Volume (4:00 p.m.-5:00 p.m.)
Monday	9/13/82	39,353	2,797
Tuesday	9/14/82	40,782	2,886
Monday	9/20/82	38,935	2,662
Tuesday	9/21/82	36,958	2,845
Wednesday	9/22/82	38,281	2,769[a]
Monday	9/27/82	37,329	2,723
Tuesday	9/28/82	38,155	2,871
Wednesday	9/29/82	39,540	2,927
Average (8 days)		38,667	2,805
1982 Annual average daily traffic		38,049	

[a] 3:00 p.m. to 4:00 p.m.

TABLE 2 Analysis of Callahan Tunnel Peak-Hour Volumes, October 1982

	Total Vehicles	Vehicles per Lane
Maximum	3,059	1,530
90 percent	2,993	1,497
85 percent	2,975	1,488
75 percent	2,937	1,468
50 percent	2,850	1,425
Average volume	2,794	1,397
Standard deviation	213	107

Table 3 gives April 1983 4:45 p.m. to 5:45 p.m. peak-hour volumes, based on field surveys and volume analysis, entering the tunnel by direction of approach. Volumes approximated 3,000 vehicles per hour (vph) of which about 2 percent were trucks or buses. Flows were slightly higher than those indicated by the Turnpike Authority data, which were recorded on an hourly basis.

TABLE 3 Distribution of Callahan Tunnel Approach Volumes, March 1983 (4:45 p.m. to 5:45 p.m. peak hour)

Source	Volume	Percentage of Total Traffic
From the south	1,849	60.9
Central Artery	998	32.9
Surface Artery	851	28.0
From North Street	1,185	39.1
North St. (eastbound)	218	7.2
Blackstone St. (southbound)	352	11.6
Central Artery (southbound)	615	20.3
Total	3,034	100

The tunnel approach volumes were found to be unevenly distributed between the approaches from the south (Artery northbound off-ramp and Surface Artery) and the approaches from the north and west (Artery southbound off-ramp, Blackstone Street, and North Street). Approximately 61 percent of the traffic came from the four northbound lanes and 39 percent from the three or four North Street lanes. (North Street operated as three lanes for most of the 3:00 p.m. to 6:00 p.m. time period when data were collected, but the videotapes recorded the emergence and disappearance of a fourth lane throughout the peak hour in response to congestion in the other three lanes.)

TUNNEL CAPACITY ANALYSIS

The capacity of a tunnel system is determined by the capacities provided at the tunnel approach (entrance), within the tunnel itself (the upgrade), or at the tunnel exit (toll plaza for the Callahan Tun-

nel) (see Figure 4). The goal is to equalize the capacity of these three points to maximize tunnel throughput and to avoid creating "shock waves."

If the volume entering the tunnel and reaching the foot of the upgrade exceeds the capacity of the system, shock waves can develop that, in turn, reduce the capacity of the system. These waves also can be triggered by inadequate capacity at the exit point. Actions at the approach area and exit portal, then, must be developed to ensure stable traffic flows at the optimum density, speed, and speed variance that are necessary to achieve the maximum throughput capacity that can be developed for the tunnel section (Figure 5).

The capacity analysis first examined the capacity of the tunnel itself, then the exit portal and toll plaza, and finally the approach area to determine the capacity of each section and the effects of each section on tunnel-system capacity. The analysis was based on a review and analysis of previous reports, MTA toll station records, and tunnel experience in the New York metropolitan area. The analysis reflects the results of field reconnaissance investigations, videotape studies, and meetings with Port Authority of New York and New Jersey personnel.

Tunnel Sections

Freeway and tunnel flows in other cities provided a basis for assessing the capacity of Callahan Tunnel:

- Peak-hour direction volumes on 23 urban freeways in larger U.S. cities ranged from 1,400 to 2,000 vehicles per lane per hour. The average was 1,730, and standard deviation was about 190.
- Reported tunnel volumes in the New York metropolitan area averaged 1,270 vehicles per hour. These flows reflect the metering effects of the street systems and toll plazas as well as the vehicle mix. Trucks, for example, comprised about 25 to 30 percent of the peak-hour peak-direction flow in the Holland Tunnel where the 90 percentile volume averaged 1,300 vehicles per hour per lane.
- The New York City Department of Transportation estimated that maximum potential tunnel-system capacity was 1,300 to 1,415 passenger vehicles per lane per hour. The Port Authority of New York and New Jersey's Tunnels and Bridges Department considers 1,350 to 1,400 passenger car units per hour a realistic maximum potential capacity; it estimates the theoretical (but unreachable) capacity based on

KEY
1. ENTRANCE
2. DOWNGRADE
3. LEVEL
4. UPGRADE
5. EXIT

NOTES:
- TOLL PLAZAS MAY BE AT POSITION 1 (LINCOLN TUNNEL (E.B.) / SUMNER TUNNEL (W.B.) OR AT POSITION 5 (CALLAHAN TUNNEL (E.B.)
- TUNNEL SYSTEM CAPACITY IS DETERMINED BY MINIMUM CAPACITY
- GOAL IS TO EQUALIZE POSITIONS 1 AND 4. HAVE POSITION 5 GREATER THAN 4

FIGURE 4 Components of tunnel capacity.

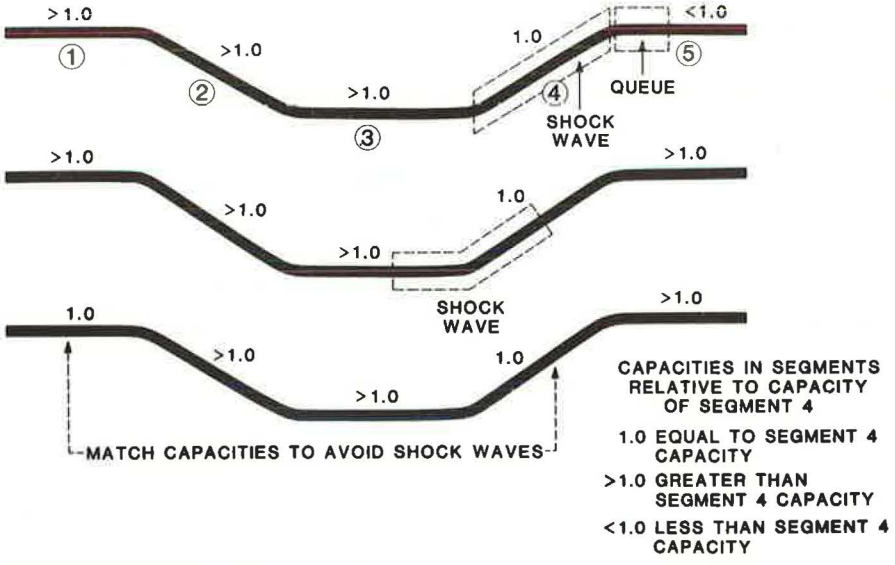


FIGURE 5 Matching section capacities to avoid shock waves.

TABLE 4 Comparative Tunnel Capacity Estimates (vehicles/lane/hour)

	Tunnel Only	System	
		In	Out
New York City			
Brooklyn-Battery		1,390[a]	1,310[a]
Queens Midtown		1,390[a]	1,415[a]
Lincoln		1,320[a]	1,350[a]
Holland		1,300[a]	1,415[a]
New York Port Authority			
Theoretical maximum (effects of lane width and clearance)	1660[a]		
Practical		1,350-1,400[a]	
Boston			
Callahan			
Observed		1,400-1,500	
Transportation Systems Center (1970)	1,550	1,520	
Interim Materials on Highway Capacity (TRB Circ. 212)	1,435-1,485		
Sumner		1,600	
Suggested value for Callahan (April 1983)	1,550-1,650	1,450-1,500	

[a]Passenger car equivalents.

lane widths and lateral clearances at 1,660 vehicles per hour.

• A 1970 study conducted by the Transportation Systems Center estimated a minimum capacity of 1,550 vehicles per lane at the beginning of the upgrade in the Callahan Tunnel.

A summary of the various capacity estimates is given in Table 4. On the basis of these estimates the capacity of the Callahan Tunnel before improvements was estimated at 1,600 to 1,650 vehicles per lane per hour. This compares with maximum peak-hour system volumes of 1,450 to 1,500 vehicles per lane per hour. The differences between tunnel capacity and actual system volume resulted from the turbulence and constraints at entry and exit points. The tunnel entrance and exit, not the tunnel itself, limited the capacity of the system.

Approach Area

The tunnel approach area was characterized by nearly equal flows merging with each other. There were, however, backups on various approaches, suggesting that the approach road system and weaving areas limited system capacity. Accordingly, weaving vehicles were tracked by videotape to determine their effects on approach traffic flow.

A summary of weaving volumes is shown in Figure 6. There were four key findings:

• The two Central Artery lanes and the southerly Surface Artery lane contributed 96 percent of the traffic in the south tunnel lane, with only 4 percent coming from the other four or five lanes. Similarly, the North Street lanes plus the northerly Surface Artery lane contributed 95 percent of the traffic in the north tunnel lane, with only 5 percent coming from the remaining three lanes.

• Weaving traffic thus constituted 4.5 percent of total approach traffic.

• The highest weaving volumes came from the two Surface Artery lanes (11.3 percent of Surface Artery total). However, equal Surface Artery lane volumes fed each tunnel lane. The Surface Artery lanes, which offer the most choice between tunnel lanes, served as a "load balancer" for the system.

• There was practically no weaving from either of the two outer lanes.

The number of weaves was translated into approach area capacity reductions by assigning seconds of delay for each lane crossed by a weaving vehicle and

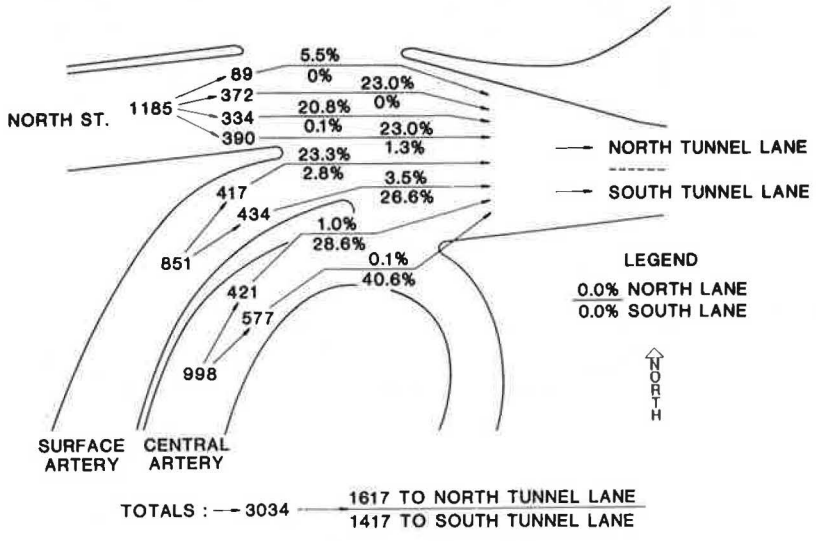


FIGURE 6 Summary of approach area weaving analysis.

translating the total delay into a capacity loss. The delay values were based on actual delays measured from the videotapes. Delays usually resulted from weaving vehicles waiting for gaps during which to enter the next lane. Delays were particularly noticeable for trucks and buses, which required a longer gap. The resulting average delays for each weaving vehicle are given in Table 5.

The capacity loss in vehicles per hour was estimated from the following formula:

$$Cr = D \times F = WTF$$

TABLE 5 Calculation of Approach Area Weaving Delay by Lane

	Cars Weaving from Lane (south to north)	No. of Lanes[a] Crossed	Total Delay (sec)
Central Artery	A	3	25
	B	2	15
Surface Artery	C	1	5
	D	1	5
North Street	E	2	15
	F	3	25
	G	4	30
	H	5	35

[a]Includes lane entered.

where

D = total seconds of lane delay = W x T;
F = flow rate, expressed in vehicles per lane per second;
Cr = capacity reduction in vehicles per hour; and
T = seconds of delay per weave.

Table 6 gives seconds of delay converted into capacity losses in vehicles for four flow rates ranging from 2,900 to 3,350 vehicles per hour, or 0.11 to 0.133 vehicle per lane per second. The reduction in capacity caused by weaving delays ranged from 120 to 150 vehicles per hour depending on the flow rates through the tunnel.

In a merging situation, traffic volume in a lane next to a wall must be greater than volumes merging in an adjacent lane. Otherwise, the lane along the wall is "pinned" against the wall with a total loss in capacity. Analysis of the approach area videotapes showed that the north lane of traffic entering the tunnel presented a flow problem that related to a "weak wall." Figure 6 shows that the southernmost approach lane carried the highest volume into the tunnel--577 peak-hour vehicles. This steady stream of traffic fed almost completely into the south tunnel lane and contributed 41 percent of the volume of the south lane. To the north, however, the northerly North Street lane carried 372 vehicles, and this flow was disturbed by the intermittent formation of the eighth lane to the north. Because these flows were both lower and perturbed in nature, vehicles proceeding to the north tended to be pinned against the portal wall as vehicles from the other lanes merged toward the tunnel. Each pinned vehicle would create delays as it negotiated its way back into the stream from a dead stop. Although the effects on capacity of the weak wall were not quantified, strengthening the flow in the northerly lane would minimize stops and improve the flow rate.

Toll Plaza Operations

Turbulence in the toll plaza environs, and imbalanced use of both toll plazas and tunnel lanes, resulted in a capacity loss of about 250 vehicles per hour at the tunnel exit. This estimate was based on a detailed analysis of videotapes that provided traffic flows through each of the service toll booths at 5-min intervals.

The counts were then factored to hourly volumes to establish maximum flow rates (i.e., toll booth capacity). Toll Booth 11, which handled the highest volumes, achieved a maximum hourly service rate of 552 vehicles and an average rate of 511. If all seven toll booths had been used evenly, their total capacity would have approximated 3,580 vehicles; this is nearly 300 vehicles per hour more than the capacity of the tunnel itself.

However, both the videotape analysis and counts taken at the exit portal and at the toll booths revealed an uneven use of the seven toll booths (Figure 7, based on videotape counts by SG Associates, Inc., May 8, 1983, and Table 7):

- First, although lane usage entering the tunnel was almost exactly equal, at the exit portal 54 percent of the traffic used the north lane and 46 percent used the south lane, indicating that crossovers to the left were occurring within the tunnel in response to the trapping effect of the right-lane queues.
- Maximum peak-hour flow rate for Booth 1, based on 5-min volumes, was 348, as opposed to 516 for Booth 13 and 540 for Booth 11.
- Actual peak volumes for the two right booths were 974 as opposed to 726 in the two left lanes.
- Total volume increased in the 5:00 p.m. to 6:00 p.m. hour by 190 over the 4:00 p.m. to 5:00 p.m. hour. Of this increase, 133 vehicles, or 70 percent, used the left three booths, increasing booth volume in Booths 1, 3, and 5 by 12 percent as

TABLE 6 Hourly Approach Area Capacity Loss Resulting from Weaving Vehicles Under Slow-Speed and Forced-Flow Conditions

	Lane	No. of Vehicles Weaving	Delay per Weave (sec)	Total Lane Delay (sec)	Capacity Loss in Vehicles at Flow Rate			
					410 (.11 veh/sec)	430 (.12 veh/sec)	470 (.13 veh/sec)	480 (.133 veh/sec)
Central Artery	A	2	25	50	6	6	7	7
	B	16	15	240	26	29	31	32
Surface Artery	C	57	5	285	31	34	37	38
	D	40	5	200	22	24	26	27
North Street	E	18	15	270	30	32	35	36
	F	2	25	50	6	6	7	7
	G	0	–	–	–	–	–	–
	H	0	–	–	–	–	–	–
Total				1,075	121	131	143	147

Note: Dashes = not applicable because the number of weaving vehicles was zero.

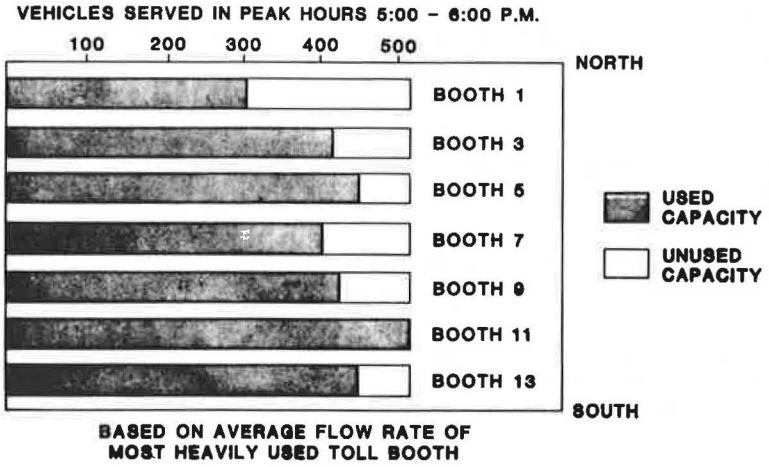


FIGURE 7 Unused toll booth capacity under booth usage patterns of April 1983.

TABLE 7 Toll Booth Usage from Videotape Record 4:00 p.m. to 5:00 p.m. and 5:00 p.m. to 6:00 p.m., April 1983

		4:00 p.m.-5:00 p.m.		5:00 p.m.-6:00 p.m.	
	Booth	No.	Percentage	No.	Percentage
North	1	283	9.9	308	10.0
	3	378	13.1	422	13.8
	5	412	14.3	476	15.5
Subtotal (1, 3, 5)		1,073	37.3	1,206	39.3
South	7	382	13.2	406	13.2
	9	444	15.4	455	14.9
	11	499	17.4	511	16.7
	13	478	16.7	488	15.9
Subtotal (9, 11, 13)		1,421	49.5	1,454	47.5
Total		2,876		3,066	

opposed to a 6 percent rise in Booth 7 and only a 2 percent rise in Booths 9, 11, and 13. The larger increase in Booths 1, 3, and 5 clearly illustrates the potential capacity gain of more evenly distributed booth usage.

The imbalanced toll booth use caused queues at the three south booths to extend farther and farther back during the peak hours. "Escape" to the left was prohibited. This caused both tunnel lanes, but especially the south lane, to slow down in the tunnel upgrade section. The slowdown in the south lane encouraged crossovers to the north lane within the tunnel--an illegal movement. Further, the tendency toward the right booth and the "trapping" effect reduced the throughput potential of the south tunnel lane.

The result of this queueing on tunnel-system capacity is apparent from the differences in lane volumes at the tunnel exit portal. Between 4:00 p.m. and 5:00 p.m. the north tunnel lane carried 1,526 vehicles, as opposed to 1,252 in the south lane. The difference between the two lane volumes represented the loss in capacity to the geometry of the toll plaza and driver behavior--about 235 to 275 vehicles per hour.

It was this phenomenon, not the toll booths per se, that limited tunnel capacity. The toll booths themselves could handle more than 3,500 cars in the peak hour, or 1,750 per lane; yet the uneven distribution of booth usage, the resultant queueing, and the trapping of vehicles in the south lane reduced capacity by approximately 250 vehicles in the peak hour.

Summary

Adding the capacity reductions caused by conflicts at the entrance and exit portals gave an approximation of the throughput that could be attained if these problems were eliminated:

Toll plaza geometry and driver behavior	250 vehicles/hour
Approach area weaving	120-150 vehicles/hour
Total capacity reduction	370-400 vehicles/hour

When this capacity reduction is added to the March-April 1983 p.m. peak-hour volumes of 2,800 to 3,000 vehicles, a volume of 3,170 to 3,400 vehicles per hour results. This improved volume represents the tunnel-system flows that could be achieved on a sustained basis if improvement actions were implemented--a 13 percent improvement over existing peak-hour conditions. It compares with the estimated in-tunnel capacity of 3,200 to 3,300 vehicles per hour. The estimated capacity gain would be more evident in the south tunnel lane and less pronounced in the north lane, which was effectively at capacity.

REVISED CONDITIONS

The operations and capacity analysis of existing conditions provided the basis for traffic management actions after May 1983. The various improvements to the tunnel entry and exit points were designed to (a) maximize Callahan Tunnel throughput and (b) im-

prove cross-harbor travel times. Accordingly, the following specific design principles were keyed to each of those sections of the tunnel road system:

1. Approach area
   - Weaving in the approach area should be reduced or eliminated, leading to a capacity increase of 405 percent;
   - The tunnel approach area should be kept filled with traffic to ensure maximum flow into the tunnel, yet queues should not extend to the local streets or main-line expressway (Figure 8);
   - Approach flows should be balanced to achieve equal (maximum) volume in each tunnel lane and hence maximum total tunnel volume;
   - Trucks and buses should be allowed to use either tunnel lane (necessary to eliminate weaving in the approach area);
   - Earlier warning for oversize vehicles should be provided to facilitate escape before entry into the approach area;
   - The weak wall of traffic in the north lane should be strengthened to ensure two strong flows on the outside; and
   - Conflicts at the North and Blackstone intersection should be reduced to allow a more regular flow of traffic from North Street and the southbound artery ramp into the tunnel.
2. Tunnel area
   - Speed reductions on the upgrade within the tunnel should be minimized; and

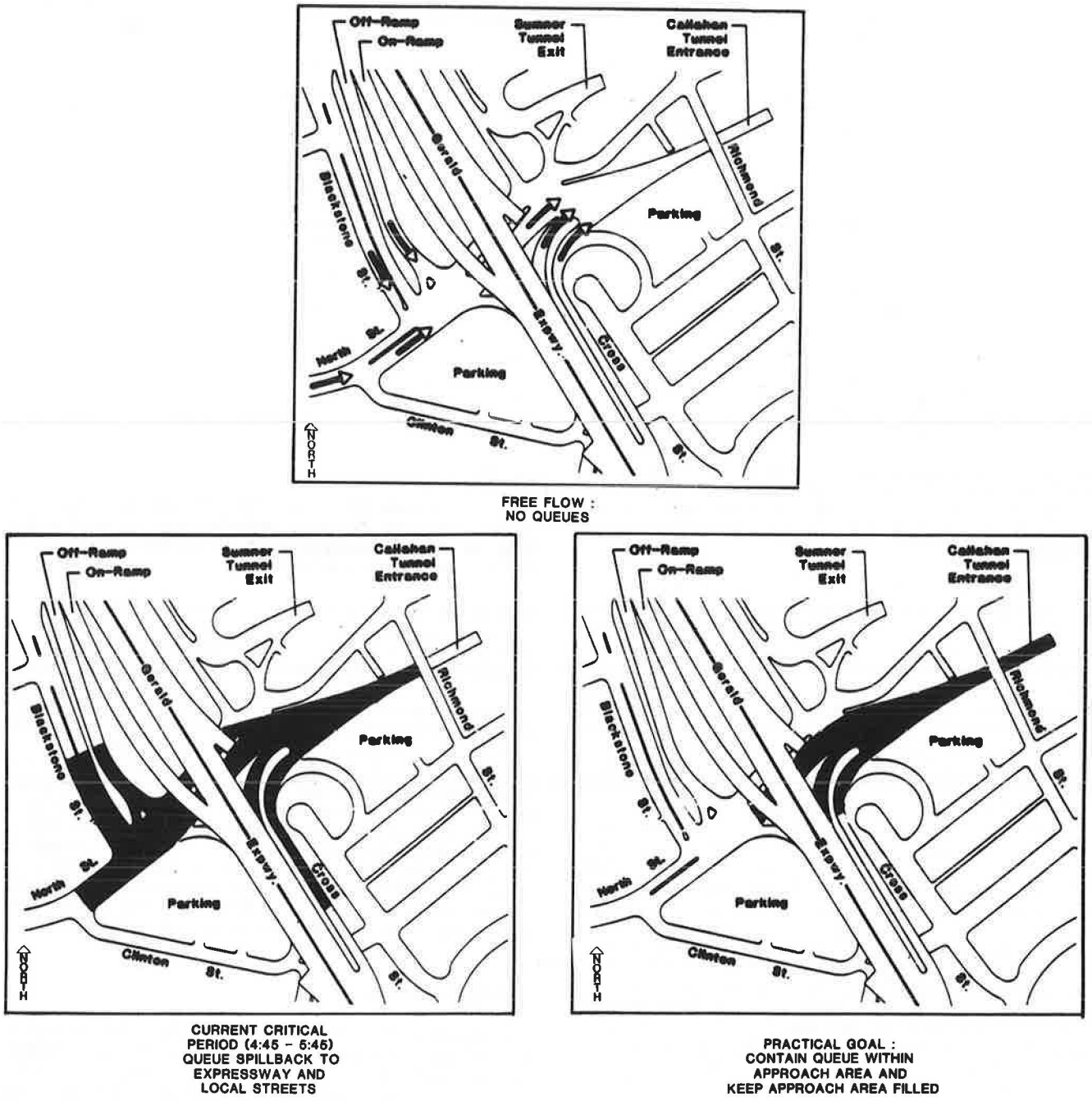


FIGURE 8 Goal of approach area TSM actions.

- Speed variations throughout the tunnel should be minimized; maximum volume throughput is achieved at an optimum speed with minimal speed variation.

3. Exit portal and toll booth area
   - The capacity restraint resulting from uneven use of the toll booths should be eliminated through channelization in the exit area or by elimination of toll collection, or both;
   - As vehicles exit the toll booth area, conflicts between those headed for local streets and those headed for McLellan Highway should be reduced;
   - Adequate maneuvering room must be provided for tow trucks; and
   - Adequate provision must be made for emergency vehicles in case of an airport-related or other disaster.

On the basis of these guidelines, various improvement options were developed and reviewed by participating public agencies. This led to three basic actions that were implemented beginning May 2, 1983:

- One-way toll collection inbound on the Sumner Tunnel and elimination of outbound toll payment on the Callahan Tunnel;
- Realignment of the tunnel exit road, made possible by eliminating the toll booths and collection (Figure 9); and
- Channelization of the tunnel approach roads, first by barrels and then by permanent construction to better funnel flow and reduce weaving movements (Figure 10).

RESULTS

Traffic volumes and patterns on the approach to the tunnel were obtained from videotape analysis for conditions "after" the one-way toll collection and temporary approach channelization were placed in effect. Table 8, based on field and videotape counts done by SG Associates in 1983, and Figure 11 present comparisons of the before and after patterns:

- The p.m. peak-hour traffic volumes for the after conditions totaled 3,238 vehicles compared with 3,034 before the improvements were made. This

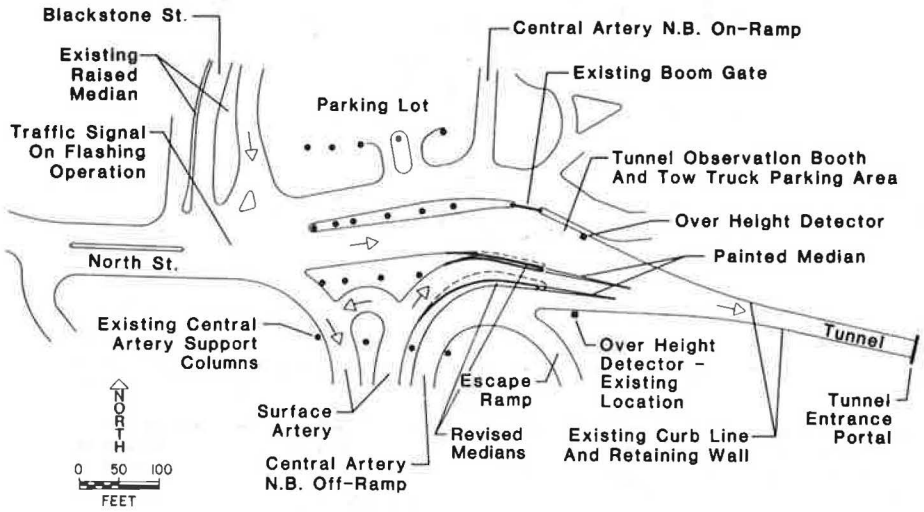


FIGURE 9 Alignment plan for Callahan and Sumner tunnels—exit road.

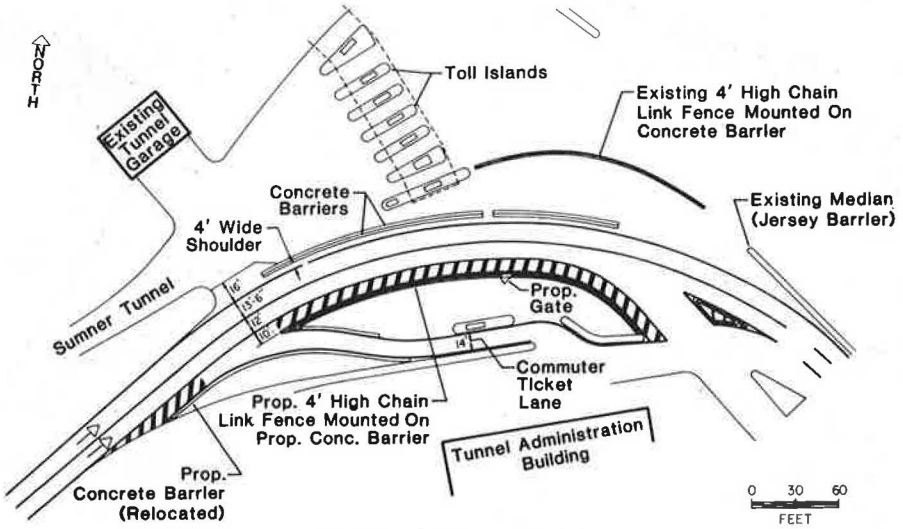
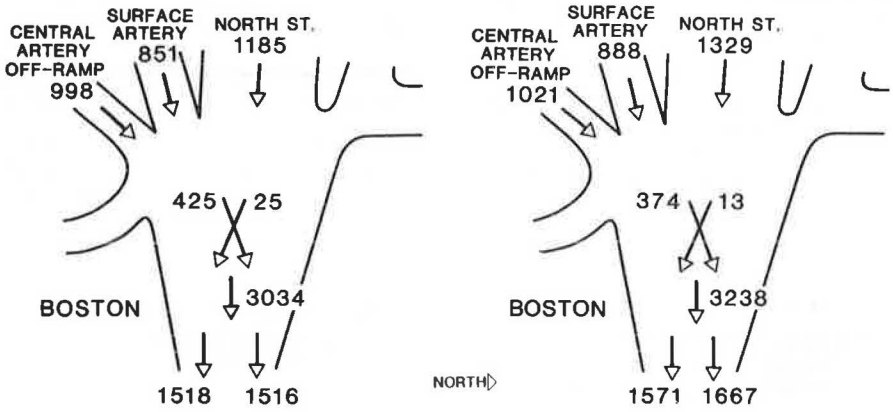


FIGURE 10 Alignment plan for Callahan and Sumner tunnels—approach roads.

TABLE 8 Summary Distribution of Callahan Tunnel Approach Volumes Before and After One-Way Toll Experiment and Approach Area Channelization

Source	4:45 p.m.-5:45 p.m. Preexperiment Peak Hour (3/4/83)		5:00 p.m.-6:00 p.m. Postexperiment Peak Hour (6/10/83)		Difference	
	Volume	Percentage of Total Traffic	Volume	Percentage of Total Traffic	Volume	Percentage
Central Artery	998	32.9	1,021	31.6	23	2.3
Surface Artery	851	28.0	888	27.4	37	4.3
Subtotal from the south	1,849	60.9	1,909	59.0	61	3.3
From North Street	1,185	39.1	1,329	41.0	144	12.2
Total	3,034	100	3,238	100	204	6.7



Before 3/4/83
P.M. Peak Hour 4:45 - 5:45 P.M.

After 6/10/83
P.M. Peak Hour 5:00 - 6:00 P.M.

FIGURE 11 Callahan Tunnel traffic volumes.

suggests a volume increase of 200 vehicles. The after peak volume falls within the 3,200 to 3,300 capacity range anticipated for the tunnel system.

- The crossover or weaving volumes dropped from 450 to 387 vehicles, even though there was no change in the approach distributions; they remained 60 percent south and 40 percent north.
- Volumes recorded at the tunnel exit in East Boston increased from 3,120 to 3,280 vehicles.

Thus, the operational changes appear to have improved Callahan Tunnel performance. In addition to the capacity gains, there was a reduction of delays and turbulence at both the entry and the exit, and average speeds through the tunnel improved.

The improved speeds have resulted in fewer breakdowns (due to stoppages) and lower utility costs for tunnel fans (due to the "piston effect" of moving traffic and reduced CO emissions).

SUMMARY AND INTERPRETATION

The analyses of existing tunnel-system volumes and operations suggested a tunnel capacity of 1,600 to 1,650 vehicles per lane per hour compared with existing peak-hour system volumes of 1,450 per lane. The tunnel capacities exceed those obtained by traditional capacity procedures--1,435 to 1,485 vehicles per lane per hour.

Channelization of the tunnel approach, elimination of outbound toll collection, and realignment of the tunnel exit lanes were estimated to increase system throughput to about 1,650 vehicles per lane. Sample "after" studies indicate that this flow rate is actually achieved with reduced delay throughout the tunnel system.

Perhaps even more significant was the rapid implementation of improvements--too often lacking in management actions. Proposed concepts were developed in March and April 1983 and implemented during May 1983 by the various state agencies.

ACKNOWLEDGMENTS

The authors wish to acknowledge the many public agency officials who guided the study effort. Special thanks are expressed to Mel Crain of the Massachusetts Turnpike Authority, Norman Garamelli of the Massachusetts Port Authority, and Matthew Coogan, Assistant Secretary of Transportation for the Commonwealth of Massachusetts.

REFERENCES

1. H.S. Levinson. Callahan Travel TSM One-Way Toll Project. SG Associates; Massachusetts Port Authority and Massachusetts Turnpike Authority, Boston, May 1983.
2. Callahan Tunnel TSM Study: Traffic Volumes Before and After Initiation of One-Way Toll Experiment and Approach Area Channelization. SG Associates, Boston, Mass., July 12, 1983.

Publication of this paper sponsored by Committee on Highway Capacity and Quality of Service.

Quantitative Measure of Levels of Service

FRANK M. CROFT, JR., and J. EDWIN CLARK

ABSTRACT

This study was undertaken to accomplish two specific objectives pertaining to the measurement of traffic flow quality: (a) to demonstrate that the internal energy of the traffic stream as determined by the parameter acceleration noise is related to the traffic flow parameters of speed, volume, and density and (b) to demonstrate that acceleration noise can be used as a quantitative measure of the level of service for traffic flow on an urban freeway. Traffic flow data at different levels of service were collected from sections of Interstate 65 in Louisville, Kentucky, using a "floating car" equipped with a Greenshields Traffic Analyzer. The traffic analyzer data were later reduced using an interactive computer program to yield values of acceleration noise. The relationships between acceleration noise and the traffic flow parameters of speed, volume, and density were analyzed using scatter diagrams and plots of the mean values of each parameter for each level of service. The results of F-tests between the variables showed the relationships to be significant and the values of R^2 indicated a high degree of correlation. As a result of this study, it was concluded that acceleration noise due to traffic interaction is a quantitative measure of the level of service for traffic flow. This conclusion is based on the relationships between acceleration noise and the traffic flow parameters of speed, volume, and density. Respectively, these relationships are linear, exponential, and quadratic in nature. Furthermore, acceleration noise increased by a factor of 12.64 as level of service deteriorated from A to E, compared to speed, which decreased by a factor of 1.47, and compared to volume and density, which increased by factors of 2.89 and 4.29, respectively, for the same changes in levels of service.

Researchers and traffic engineers have long recognized the need for methods of comparing the effects of various design options and traffic controls on the quality of traffic flow and levels of service in a quantitative manner. Various parameters, indices, and models have been developed over the years in an effort to solve this problem, yet none has gained wide acceptance as a quantitative measure of traffic flow quality.

Travel time, the "old reliable" of traffic flow parameters, has been used practically every way possible to quantify traffic flow quality (1-3). But travel time is not a good indicator of operating conditions during a trip on a section of highway unless delays are exceedingly long or frequent.

Traffic flow has been compared to heat flow and a traffic flow model based on heat flow has been developed (4). This heat flow analogy assumes that a traffic lane is similar to a long, slender, insulated rod (controlled access and no opportunity for lane change). Heat flow through such a rod is described by a differential equation that relates the variables of temperature, distance, time, and the material properties of conductivity, specific heat, and density. Traffic flow variables are simply substituted for the appropriate heat flow variables in the differential equation and the solution yields the speed-density relationship of the traffic stream. This approach appears to be straightforward and easily understood. However, there is a problem with this analogy. The heat flow equation, when applied to traffic flow, implies that speed might be negative if the change in vehicular density increases with respect to distance. This situation is not realistic and therefore reduces the conceptual appeal of the model as a quantitative measure of traffic flow quality.

A considerable amount of work has been done with regard to fluid flow modeling of the traffic stream (4). A single lane of traffic offers a striking analogue to the flow of a compressible fluid, such as a gas, in a pipe with a constant area. Both systems consist of discrete particles: individual molecules in the case of the fluid and individual vehicles in the case of the traffic stream. In application of fluid flow models to traffic flow problems, a greater concern is implied with the overall statistical behavior of the traffic stream than with the interactions between vehicles because none of the parameters relate to the individual vehicle. Fluid models have certain shortcomings as quantitative measures of traffic flow quality because the sample size associated with traffic problems includes only a small portion of the particles or vehicles (4).

The heat flow model and fluid flow models involve systems that are governed by natural physical laws. The system associated with the traffic stream includes the driver, the road, and the vehicle. Also, traffic conditions, which are related to vehicle behavior and roadway geometry, are important factors associated with the traffic stream. None of these elements of the traffic stream are governed by the natural physical laws that are associated with heat flow or fluid flow. Although these models have been successful on a limited basis, they are not generally likely to reflect the actual traffic flow characteristics associated with the traffic stream.

A parameter that successfully measures the quality of traffic flow on a given highway facility in a quantitative manner must be sensitive to the three basic elements of the traffic stream: the driver, the road, and the vehicle. Also, such a parameter must be sensitive to traffic flow conditions and it

must measure the smoothness or the quality of flow within the traffic stream. Acceleration noise is a parameter that is sensitive to the three basic elements of the traffic stream and it is a measure of the quality of flow within the traffic stream. Acceleration noise is defined as the standard deviation of the individual accelerations and decelerations of a vehicle traveling over a section of highway.

The quality of traffic flow on a given highway facility affects many features of the traffic stream that are quantitative in nature. For example, vehicle operating costs are directly affected by the quality of traffic flow and are measured quantitatively. To determine the effect that the quality of traffic flow has on these features, a quantitative measure of traffic flow quality is needed. At present, the only parameter that is an accepted measure of traffic flow quality is level of service and it is qualitative not quantitative in nature. Therefore, it is not now possible to relate the quantitative features of the highway environment to traffic flow quality.

PURPOSE AND OBJECTIVES

The purpose of this study is to accomplish two specific objectives. These objectives are to demonstrate

- That acceleration noise is related to the traffic flow parameters of speed, volume, and density and
- That acceleration noise can be used as a quantitative measure of the level of service for traffic flow on an urban freeway.

DATA COLLECTION, ANALYSIS, AND RESULTS

Introduction

Data related to acceleration noise and level of service were collected for this study. Specific attention was given to weather and holiday limitations, highway section selection, use of the test vehicle and research equipment, and traffic count methods. The methods employed in the data-collection effort are discussed.

An attempt was made to identify and exclude those variables that were difficult, if not impossible, to control. Of course, the first of these variables to be considered was weather conditions. Adverse weather conditions can cause variation in traffic flow patterns. Data were not collected during periods of inclement weather. Furthermore, the pavement of each test section was dry and free from moisture before any data-collection runs were made. If adverse weather conditions occurred during a series of test runs, the runs were terminated and not resumed until the pavement was dry.

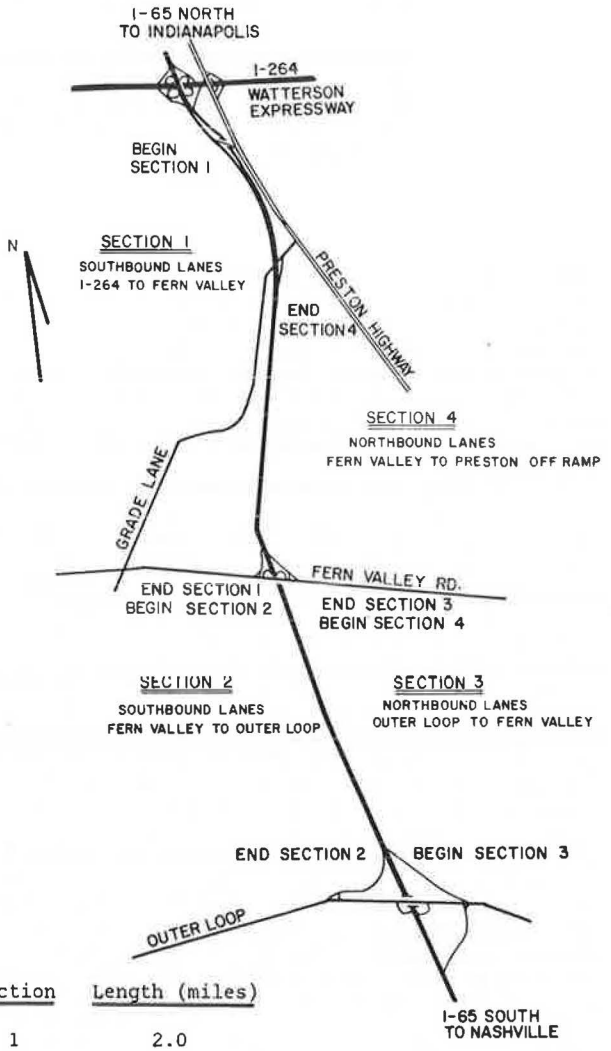
During holiday periods and weekends, the traffic flow characteristics on any highway can change dramatically compared to typical weekday traffic conditions. To eliminate any variance in the data attributable to this variable, data were only collected on weekdays excluding holidays.

Highway Test Section

Four sections of Interstate 65 in Louisville, Kentucky, were selected for study. I-65 is a four-lane urban freeway with 10-ft shoulders and a 20-ft grass median. The horizontal alignment is relatively straight, and the vertical alignment is relatively flat. Preliminary studies showed that a wide range of levels of service existed and that traffic volumes approach capacity during peak periods. These sections are shown in Figure 1. These freeway sections carry heavy traffic volumes in both the morning and afternoon peak periods. The average daily traffic (ADT) volume in 1981 on Sections 1 and 4 was 68,430 vehicles per day. The ADT volume on Sections 2 and 3, for the same year, was 60,700 vehicles per day. Classification counts supplied by the City Traffic Engineer's Office in Louisville, Kentucky, showed the amount of truck traffic to be 15 percent of this ADT volume.

Test Vehicle and Research Equipment

Jones and Potts (5) investigated the variability of acceleration noise with different drivers. They found that acceleration noise did, in fact, vary with drivers, depending on the individual driver's speed. For example, acceleration noise was essentially the same for two drivers driving below the design speed of the highway. However, if both drivers exceeded the design speed, acceleration noise was greater for the faster driver. On the



Section	Length (miles)
1	2.0
2	1.0
3	1.0
4	1.4

FIGURE 1 Selected study sections of I-65 in Louisville, Kentucky.

basis of the conclusions of the Jones and Potts study, a decision was made to use a single driver for the test vehicle and thus to avoid variations in the data caused by differences in drivers.

In previous studies of acceleration noise, data were recorded by various instruments including an accelerometer, a tachograph, a recording speedometer, and a traffic analyzer (4,6-8). For this study, a Greenshields Traffic Analyzer was installed in a 1971 Plymouth Fury and used to collect the required data for determination of acceleration noise. The traffic analyzer was cable connected to the test vehicle's transmission where the speedometer cable is located. A special connector was used to allow the speedometer to remain operational. The data recorded included (a) the number of positive and negative 2-mph speed changes over the highway section, (b) the running time of the test vehicle, and (c) the length of each highway section. The traffic analyzer can be set on automatic mode that allows for recording all of these data on a constant-distance (every 0.10 mile) basis. There is also a manual switch that allows the operator to record data at any time during a test run. The data are printed in a digital format on recording tape during each data-collection run and can easily be reduced and analyzed at a later time. The format of the recording tape is shown in Figure 2.

Number of +2-MPH Speed Changes	Number of -2-MPH Speed Changes	Incremental Distance (x.xx Miles)	Incremental Time (sec)	Vehicle Stopped Time (sec)	Vehicle Speed (mph)	Counter for Cumulative Runs		
00000100	90	07	00	53	00	P	END RUN
00000001	00	06	00	55	00		
00000001	00	07	00	55	00		
00000001	00	06	00	55	00		
00000001	00	07	00	55	00		
00000001	00	07	00	54	00		
00000001	00	07	00	54	00		
00000001	00	06	00	55	00		
00000001	00	06	00	54	00		
00100001	00	07	00	55	00		
00000101	00	06	00	53	00		
00000001	00	07	00	54	00		
00000001	00	07	00	55	00		
00000001	00	07	00	54	00		
00000000	10	01	00	55	00	P	BEGIN RUN
00000001	00	06	00	55	00		
00000101	00	07	00	55	00		

FIGURE 2 Format of traffic analyzer recording tape.

The "floating car" technique was used to collect data relating the traffic flow parameters of speed, volume, density, and acceleration noise on the selected freeway sections. This technique has long been established as providing an accurate determination of the average speed of the traffic stream on a given highway section. It is assumed that the acceleration data for a floating car would also represent an average of the accelerations of the vehicles in the traffic stream (4).

Data Collection

Total acceleration noise on a given freeway section is the sum of the natural acceleration noise and the acceleration noise due to traffic interaction. The natural acceleration noise is the acceleration noise of the driver and the vehicle in response to the geometric characteristics of the highway. The natural acceleration noise can be measured by making test runs with the floating car on the highway section in the absence of significant traffic. This can be accomplished by making runs between the hours of 1:00 a.m. and 5:00 a.m. when traffic volume on this urban freeway is low. A total of 65 test runs was made on the selected freeway sections between the hours of 1:00 a.m. and 5:00 a.m. The data were then reduced by an interactive computer program to calculate the acceleration noise.

Acceleration noise due to traffic interaction can be determined by making test runs with the floating car at various times of the day when traffic volume is considered to be low, moderate, and high. During the peak hours (7:00 a.m. to 9:00 a.m. and 3:00 p.m. to 6:00 p.m.), traffic volumes are directional. Inbound traffic volume in the morning is generally high, while outbound traffic volume is relatively low. Conversely, outbound traffic volume in the afternoon is generally high, while inbound traffic is relatively low. A single round trip yields data with regard to high and low traffic volumes. Test runs during the morning and afternoon peak periods yielded data over a wide range of speeds and densities compatible with levels of service A through E. Traffic flow during the test runs was never stopped; however, at level of service E, traffic flow was generally slow-and-go.

Approximately 100 test runs that generated more than 488 vehicle-miles of travel were made on the selected freeway sections to collect the data required to calculate acceleration noise due to traffic interaction. During these test runs 346 data points were collected at various levels of service. A single data point reflects the acceleration noise measured over the length of a test section. A single test run over the four freeway sections generated four data points.

Traffic volumes during test runs were determined using the manual count method at two count station locations along the test sections. Two observers were located at each count station with instructions to record the vehicle count during each 15-min period listed on the volume data sheets. One observer was responsible for recording the traffic count in the northbound lanes while the other was responsible for recording the traffic count in the southbound lanes. It was extremely important that the traffic counts be coordinated with the data-collection runs of the test vehicle. Before going into the field, the driver of the test vehicle and the observers conducting the traffic counts synchronized their watches. The format of the traffic count field sheet is shown in Figure 3.

Data Reduction

The values of acceleration noise, speed, volume, and density were reduced from data collected during the test runs on the designated sections of I-65. The test runs were conducted when traffic flow condi-

Date: 6-16-82
Southbound I-65
Fern Valley Road to Outer Loop Road--Afternoon Peak

Time Period	15 Minute Count
3:00 - 3:15	741
3:15 - 3:30	792
3:30 - 3:45	862
3:45 - 4:00	874
4:00 - 4:15	908
4:15 - 4:30	899
4:30 - 4:45	801
4:45 - 5:00	946
5:00 - 5:15	896
5:15 - 5:30	917
5:30 - 5:45	742
5:45 - 6:00	657

FIGURE 3 Format of typical traffic count field sheet.

tions ranged from level of service A to level of service E. The numbers of test runs associated with each level of service were 76 for level of service A, 111 for level of service B, 79 for level of service C, 57 for level of service D, and 23 for level of service E. There are 346 sets of data points in the data base.

Two computer systems were used in the reduction and analysis of the data collected for this study. The DECsystem10 computer, at the University of Louisville, was used to calculate acceleration noise, average speed, and traffic density using an interactive program developed for this study. The IBM 3081-K computer at Clemson University was used to statistically analyze the reduced data and generate a major portion of the graphic information related to this study.

Level of service criteria for freeway sections are given in The Highway Capacity Manual (HCM), Table 9.1 (9), where the maximum allowable volume for each level (or maximum allowable volume-to-capacity ratio) is tabulated against speed. The use of the HCM table implies a correlation between the volumes and speeds shown, when, in fact, no such relationship exists. Meeting one criterion (either speed or volume) does not guarantee that the other has been met.

Roess et al. (10) under a research contract from the FHWA and the National Cooperative Highway Research Program revised the criteria for establishing level of service for freeway sections. They considered it critical that any revised criteria be based on correlated values of volume and speed. Figure 4 shows the speed-volume relationship for freeway capacity.

These curves are based on limited amounts of data available in the literature as well as three pilot studies conducted specifically by Roess et al. (10). The curves show two major characteristics of freeway sections. First, there is a significant range of volumes over which speed is insensitive to volume. Second, as volume approaches capacity, speed decreases rapidly (10).

The use of the HCM table to determine level of service is impractical because speed is insensitive to volume over a large range. A more reliable parameter is density. Therefore, in this study, speed and density were used to define level of service. Table 1 gives the level of service recommendations based on speed and density for basic freeway sections (10). This table was used to determine levels of service associated with this research.

Acceleration noise can be expressed mathematically as

$$\sigma = \left[(1/T) \int_{[0,T]} (a(t))^{1/2} dt\right]^{1/2} \quad (1)$$

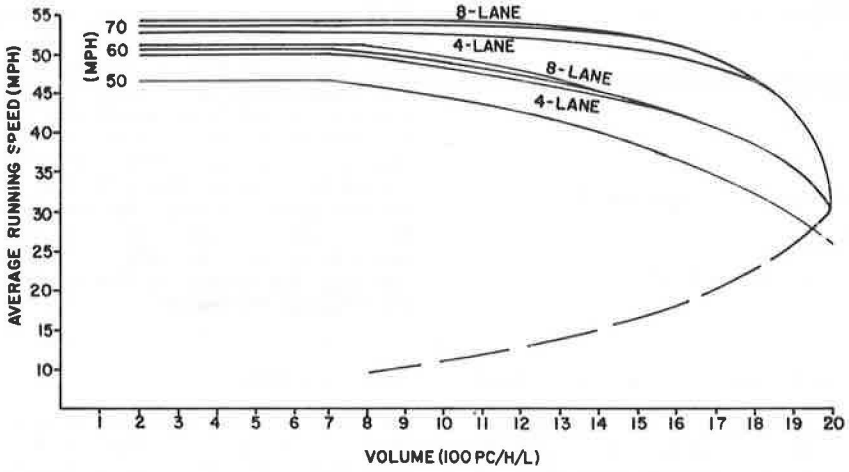


FIGURE 4 Recommended speed-flow curves for freeway capacity *(10)*.

TABLE 1 Levels of Service for Basic Freeway Sections *(10)*

Performance Criteria for Levels of Service			Maximum Service Volumes (one direction) for Levels of Service During Uniform Periods of Flow (passenger cars/hour)			
Level of Service	Speed (mph)	Density (passenger cars/mile/lane)	4 Lanes	6 Lanes	8 Lanes	Each Additional Lane
Average Highway Speed = 70 mph						
A	≥ 50	≤ 15	1,600	2,400	3,280	820
B	≥ 50	≤ 25	2,500	3,900	5,400	1,350
C	≥ 48	≤ 35	3,400	5,100	6,800	1,700
D	≥ 40	≤ 47	3,850	5,775	7,700	1,925
E	≥ 30	≤ 67	4,000	6,000	8,000	2,000
F	< 30	> 67	highly variable			
Average Highway Speed = 60 mph						
A	–	–	–	–	–	–
B	≥ 45	≤ 25	2,300	3,525	4,800	1,200
C	≥ 43	≤ 35	3,050	4,575	6,100	1,525
D	≥ 38	≤ 47	3,600	5,400	7,200	1,800
E	≥ 30	≤ 67	4,000	6,000	8,000	2,000
F	< 30	> 67	highly variable			
Average Highway Speed = 50 mph						
A	–	–	–	–	–	–
B	–	–	–	–	–	–
C	≥ 40(64)	≤ 35	2,800	4,200	5,600	1,400
D	≥ 35(56)	≤ 47	3,300	4,950	4,950	1,650
E	≥ 30(48)	≤ 67	4,000	6,000	6,000	2,000
F	< 30(48)	> 67	highly variable			

Note: Dashes indicate level of service not achievable due to restricted average highway speed.

where

σ = acceleration noise (ft/sec^2),
$a(t)$ = acceleration (positive or negative) at time t(ft/sec^2), and
T = total time in motion (sec).

Because this equation (4,11-13) is difficult to evaluate precisely using field data, the following equation can be used to approximate acceleration noise from parameters that can easily be collected in field studies (4,13,14). The equation is

$$\sigma = \left\{ [(\Delta V)^2/T] \sum_{i=0}^{k} (n_i^2/\Delta T_i) - [(V_t - V_o)/T]^2 \right\}^{1/2} \quad (2)$$

where

σ = acceleration noise (ft/sec^2),
T = total time in motion (sec),
ΔV = constant increment of velocity change (mph),
n_i = integer denoting the number of speed changes at the constant increment (ΔV) that have occurred over time Δt_i,
V_t = velocity at trip start, and
V_o = velocity at trip end.

The natural acceleration noise (σ_n) on a given highway facility is the acceleration noise during a vehicle trip on the facility that can be ascribed to a driver's natural speed changes in the absence of significant traffic (4,12). The natural acceleration noise for each section of I-65 was determined by averaging the values of acceleration noise calculated from data collected during 65 test runs made in the absence of a significant volume of traffic. Table 2 gives the value of the natural acceleration noise for each section of I-65 and the 95 percent confidence limits on each mean value.

TABLE 2 Natural Acceleration Noise of Four Sections of I-65

Section	Natural Acceleration Noise, σ (ft/sec^2)^n	95 Percent Confidence Limit
1	0.2650	±0.0116
2	0.3180	±0.0192
3	0.2990	±0.0160
4	0.2440	±0.0176

The total acceleration noise (σ_T) of a given highway section is simply the value of acceleration noise measured during various traffic flow conditions. Total acceleration noise combines the effects of geometry and traffic interaction. An equation that represents the total acceleration noise is

$$\sigma_T = \sigma_n + \sigma_t \quad (3)$$

where

σ_T = total acceleration noise (ft/sec^2),
σ_n = natural acceleration noise (ft/sec^2), and
σ_t = acceleration noise due to traffic interaction (ft/sec^2).

Equation 3 is an expression for total acceleration noise based on the definition of acceleration noise as the standard deviation of the accelerations of a vehicle in the traffic stream. It is recognized that addition and subtraction of standard deviations is not possible in a statistical sense. However, acceleration noise is not being used in a statistical sense; therefore, addition and subtraction of the acceleration noise values (standard deviations) is acceptable. Solving Equation 3 for σ_t yields an expression for the acceleration noise due to traffic interaction. This expression is simply

$$\sigma_t = \sigma_T - \sigma_n \quad (4)$$

TABLE 3 Mean Values of Traffic Flow Parameters

	Level of Service				
	A	B	C	D	E
No. of data points	76.0	111.0	79.0	57.0	23.0
Mean speed (mph)	57.03	56.13	54.13	49.26	38.80
Mean 15-min volume	323.0	488.0	739.0	868.0	932.0
Mean density (veh/mile)	22.69	34.80	54.70	70.64	97.43
Mean acceleration noise (ft/sec^2)[a]	0.0517	0.0908	0.1825	0.3307	0.6536

[a]This parameter is the value due to traffic interaction.

The value of the acceleration noise due to traffic interaction (σ_t) is independent of the freeway section because the effects of geometry are removed by subtracting the natural acceleration noise (σ_n) from the total acceleration noise (σ_T). Acceleration noise due to traffic interaction is the parameter used to quantify level of service in this study. In the remaining portions of this study, the term acceleration noise (σ) will be used to denote acceleration noise due to traffic interaction unless otherwise noted.

The average speed for each test run was calculated using total elapsed time and total distance data from the traffic analyzer tape. The average speed calculated from this time and distance data is called the space-mean speed. The data associated with the time and distance were input into the computer program developed for this study and the average speed was calculated and printed.

For each level of service, the mean value of speed, 15-min traffic volume, density, and acceleration noise was calculated. These values were determined by averaging each parameter at each level of service. For example, the space-mean speed for each run at level of service A was summed and divided by 76. These values are given in Table 3.

Data Analysis

The specific relationship between acceleration noise and level of service was analyzed. SAS was used to analyze each of the relationships, and SAS/GRAPH was used to develop the graphic displays for each relationship. SAS and SAS/GRAPH are statistical analysis programs developed by the Statistical Analysis System Institute (15-17).

As a first step in the analysis process, a scatter diagram was developed describing the relationship between each of the dependent and independent variables. Each relationship was then analyzed using an F-test to determine the significance of the relationship between the variables. The value of F and the probability that a larger value of F exists by chance alone (PR > F) were calculated for each relationship. If the value of PR > F is 0.05 or less, the relationship between the dependent variable and the independent variable is considered significant. The 0.05 level of significance is generally accepted for most engineering work. The value of the square of the correlation coefficient (R^2) was calculated for each relationship.

The relationship between acceleration noise and each of the three traffic flow parameters was investigated and analyzed. Figures 5-7 are scatter diagrams that show the relationship of acceleration noise to speed, volume, and density, respectively. Acceleration noise is considered the dependent variable, and the three traffic flow parameters are considered the independent variables. Included in each figure is the regression equation that describes the relationship between the variables and the values of F, PR > F, and R^2, which describe the statistical

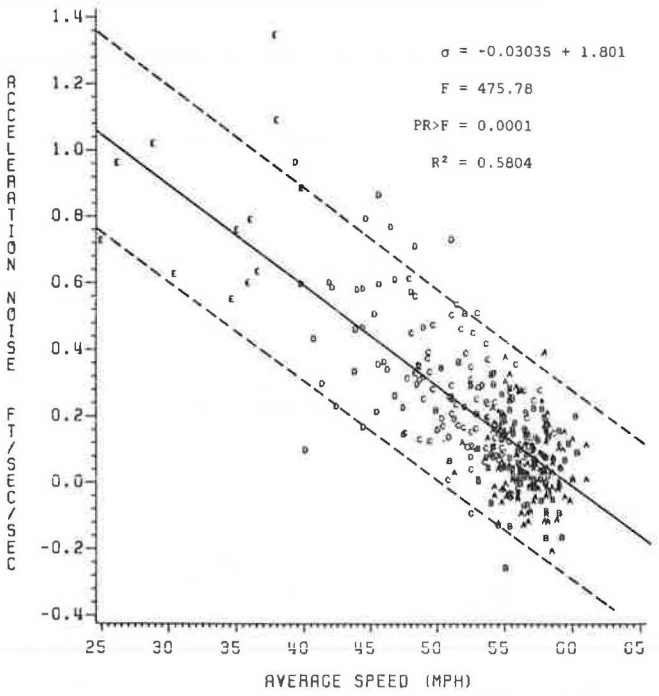


FIGURE 5 Scatter diagram of acceleration noise versus speed.

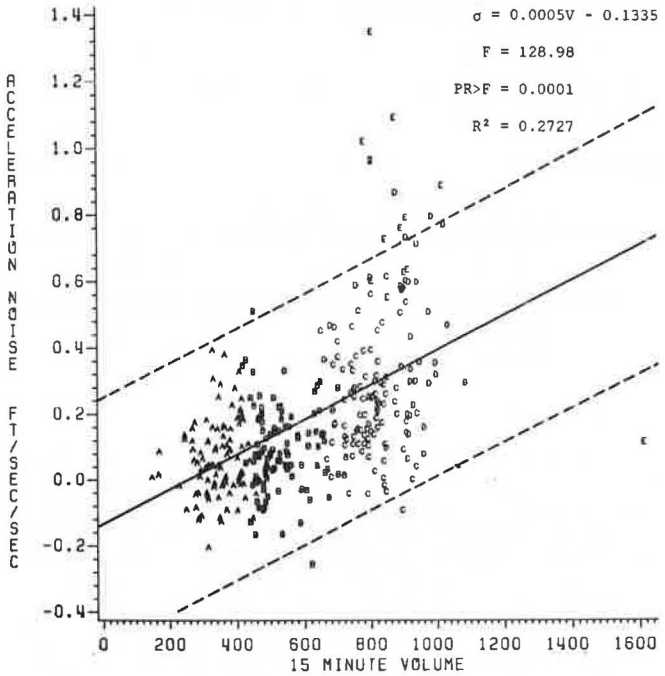


FIGURE 6 Scatter diagram of acceleration noise versus 15-min volume.

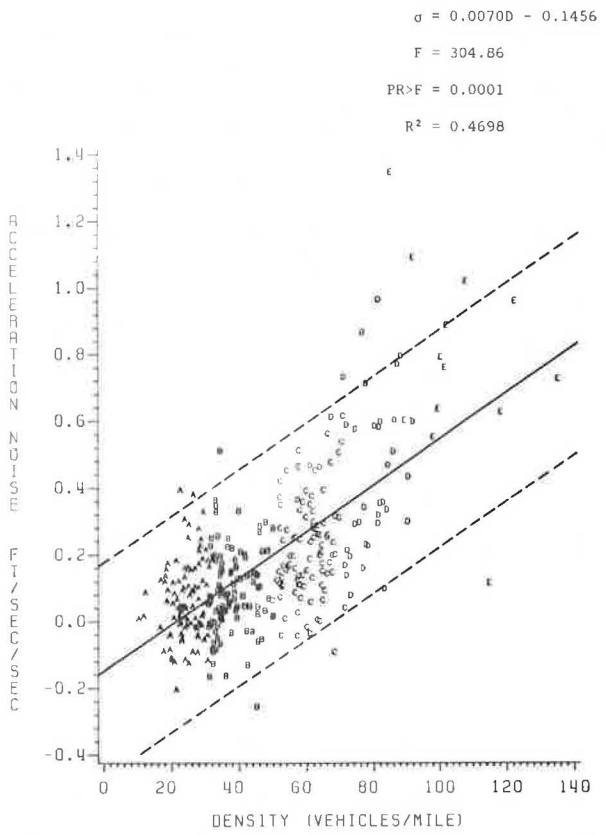


FIGURE 7 Scatter diagram of acceleration noise versus density.

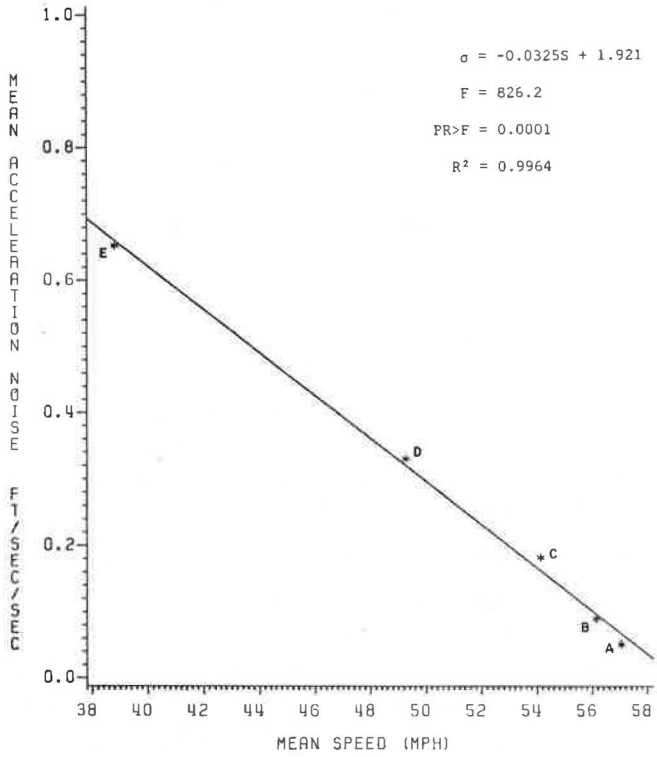


FIGURE 8 Average acceleration noise versus average speed at each level of service.

significance of the relationship between the variables. The solid line through the data represents a linear regression fit and the dashed lines represent the upper and lower 95 percent confidence limits.

There is a significant amount of scatter present in Figures 5-7. This scatter is due to the high degree of variability in the traffic flow data. The amount of scatter present masks the significance of the relationship between the variables.

To gain a better understanding of the relationship between acceleration noise and the three traffic flow parameters, plots showing the average value of acceleration noise at each level of service versus the average value of speed, volume, and density, respectively, at each level of service were developed. These plots are shown in Figures 8-10. Included in each figure, as in the case of the scatter diagrams, is the regression equation that relates the variables and the values of F, PR > F, and R^2.

Results

Table 3 gives the relationship between level of service and the mean values of the parameters of speed, volume, density, and acceleration noise. The table shows that speed decreases by a factor of 1.47 as level of service deteriorates from A to E, whereas volume and density increase by factors of 2.89 and 4.29, respectively, for the same changes in level of service. Acceleration noise increases by a factor of 12.64 over this range of level of service. The data given in Table 3 suggest that acceleration noise is more sensitive to levels of service than are the traditional parameters of speed, volume, and density.

To understand the relationship between acceleration noise and level of service, an analysis of the relationships between acceleration noise and each of the traffic flow parameters of speed, volume, and density was performed. The results of the analysis are discussed in the following paragraphs.

Figure 5 shows the relationship between acceleration noise and speed. The results of the F-test, which determines if a significant relationship

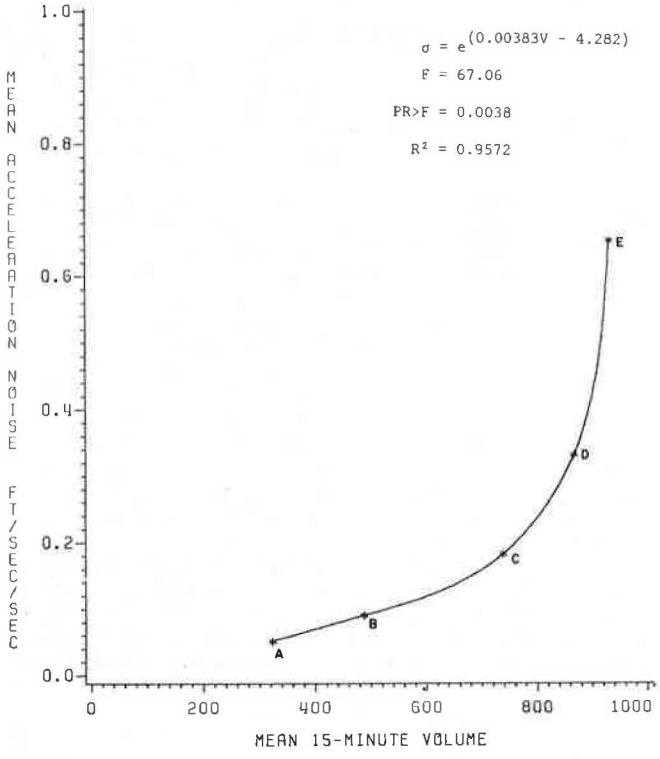


FIGURE 9 Average acceleration noise versus average 15-min volume at each level of service.

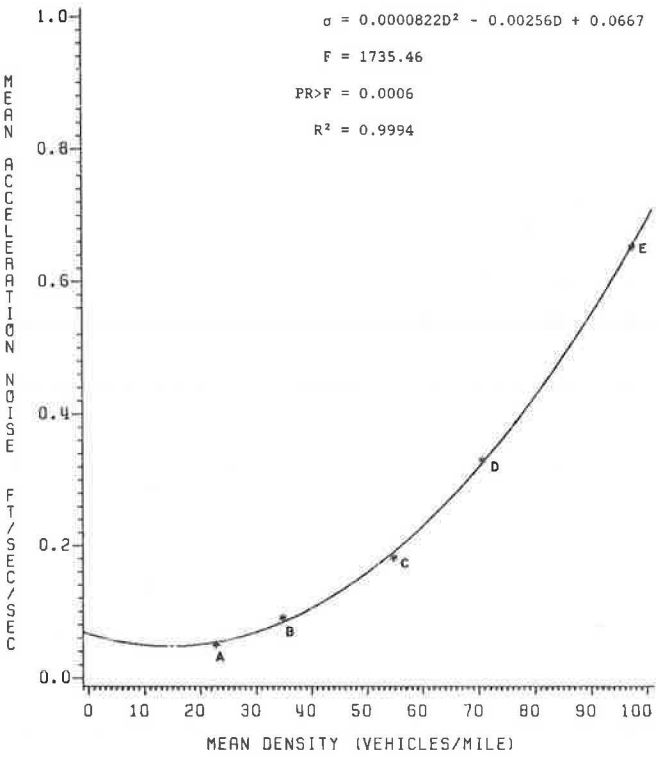


FIGURE 10 Average acceleration noise versus average density at each level of service.

exists between the two variables, show the values of F and PR > F to be 475.78 and 0.0001, respectively. These values indicate that the relationship between acceleration noise and speed is significant.

The value of R^2 is 0.5804. This value indicates that some linear dependence is present between the variables, but not enough to warrant complete linear dependence of acceleration noise on speed. This value of R^2 indicates that the data have a great deal of variance or scatter as can be observed in the scatter diagrams.

Figure 6 shows the relationship between acceleration noise and 15-min volume. The results show that F = 128.98, PR > F = 0.0001, and R^2 = 0.2727. These results are similar to the results of the analysis of acceleration noise versus speed. The relationship between the variables is significant, but the value of R^2 indicates that there is only a small degree of linear dependence between the variables. Visual inspection of the figure shows that the data are widely scattered as in the case of acceleration noise versus speed.

Figure 7 shows the relationship between acceleration noise and density. The results show that F = 304.86, PR > F = 0.0001, and R^2 = 0.4698. Again, as in the previous two relationships, the results indicate a significant relationship between the variables with little or no linear dependence.

Figure 8 shows the relationship between mean acceleration noise and mean speed at each level of service. The results show that F = 826.2, PR > F = 0.0001, and R^2 = 0.9964 for a linear model. These results indicate that a significant relationship exists between the mean values of the variables and that the relationship is linear in nature, with a high degree of correlation.

Figure 9 shows the relationship between mean acceleration noise and mean 15-min volume at each level of service. Visual inspection of the figure indicates that the relationship is nonlinear in nature, so various nonlinear models were analyzed in an attempt to define the relationship between the variables. The results show that F = 67.06, PR > F = 0.0038, and R^2 = 0.9572 for an exponential model. These results indicate that the relationship between the variables is significant and that the relationship is exponential.

Figure 10 shows the relationship between mean acceleration noise and mean density at each level of service. Again, as in the case of mean acceleration noise versus 15-min volume, the relationship is nonlinear; therefore, various nonlinear models were analyzed in an attempt to characterize the relationship between the variables. The results of this analysis show that F = 1,735.46, PR > F = 0.0006, and R^2 = 0.9994 for a quadratic model. The results again indicate statistical significance between the variables and a high degree of correlation with a quadratic model.

The scatter diagrams in Figures 5-7 indicate that acceleration noise is related to level of service. However, wide scatter in the data masks the true relationship between the variables. Figures 8-10 show strong relationships between the variables when the scatter is removed.

CONCLUSIONS

The results of this study support the premise that acceleration noise reflects the internal energy component of the traffic stream and is a quantitative measure of the quality of traffic flow. This conclusion is based on the analysis of the scatter diagrams shown in Figures 5-7 and the graphs shown in Figures 8-10.

The scatter in the data shown in Figures 5-7 increases as level of service deteriorates from A to E. This increasing scatter can be explained by the shape of the speed-flow curves for freeway operations (Figure 4) and the criteria for determining level of service given in Table 1. These curves show that the average running speed remains nearly constant at approximately 50 to 55 mph for a significantly wide range of traffic volumes. At a volume of approximately 1,500 passenger cars per hour per lane, speed drops rapidly as volumes increase only slightly. The maximum capacity for a freeway lane is approximately 2,000 vehicles per hour at a speed of approximately 35 mph.

This rapid decrease in speed for such a small increase in volume has a critical impact on the relationship between acceleration noise and level of service. At low volumes (levels of service A and B), the speed of the traffic stream will be relatively constant and acceleration noise, which is dependent on the number of speed changes, will be at or near the natural acceleration noise value. At higher volumes (level of service C to E), the speed of the traffic stream can be as high as 50 mph (Table 1). The critical parameter for each level of service is density. At higher volumes, a wide range of speed changes, which affect the parameter acceleration noise, can occur. It is entirely possible that under some conditions only a few speed changes may occur at these high volumes resulting in some low values of acceleration noise. On the other hand, there may be conditions under which several speed changes occur resulting in high values of acceleration noise. This implies that acceleration noise may not be as sensitive to volume changes as might be expected.

Scatter diagrams tend to show the relative trend in the relationship between two variables. Such diagrams can be useful in determining the significance of the relationship between two variables through an F-test. However, the true relationship between two

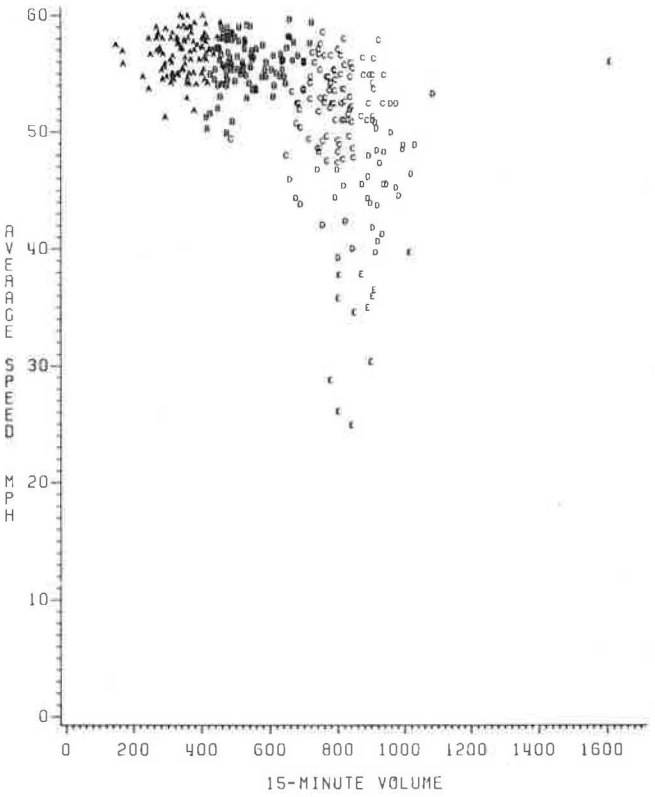


FIGURE 11 Scatter diagram of speed versus 15-min volume.

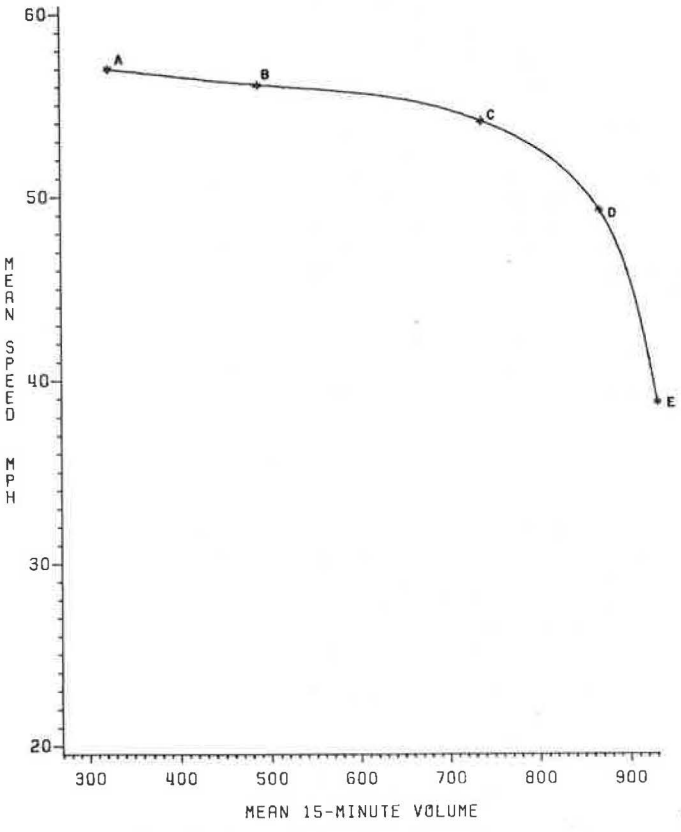


FIGURE 12 Speed-volume relationship with scatter removed.

variables can be masked by the scatter and is often difficult to determine. Figure 4 shows the relationship between speed and volume on freeway sections. The shape of these curves is accepted by traffic engineers as a typical speed-volume relationship with the maximum volume being approximately 2,000 vehicles per hour per lane at a speed of 35 mph. Real speed-volume data from traffic flow studies have the same general shape when graphed as does the ideal speed-volume curve; however, there is a significant amount of scatter associated with the real data. This is evident in Figure 11, which is a scatter diagram showing the speed and volume data collected in this study. Figure 12 shows the same speed-volume relationship after the scatter is removed by calculating the mean values of each variable at each level of service. The curve shown in Figure 12 represents the true relationship between speed and volume for the test sections associated with this study. The shape of this curve is close to the shape of the theoretical curve. Therefore, Figures 8-10, which are based on the mean values of the variables, show the true relationships between the variables. Figures 8-10 show that a strong relationship between acceleration noise and level of service exists, and, furthermore, each relationship can be modeled by the regression equation presented.

REFERENCES

1. C.A. Rothrock. Urban Congestion Index Principles. Bull. 86. HRB, National Research Council, Washington, D.C., 1954, pp. 26-39.
2. C.A. Rothrock and L.E. Keefer. Measurement of Urban Traffic Congestion. Bull. 156. HRB, National Research Council, Washington, D.C., 1957, pp. 1-13.
3. E.M. Hall and S. George. Travel Time--An Effective Measure of Congestion and Level of Service. In Proceedings of the 38th Annual Meeting, HRB, National Research Council, Washington, D.C., 1959, pp. 511-529.
4. D.R. Drew and C.L. Dudek. Investigation of an Internal Energy Model Evaluating Freeway Level of Service. Texas Transportation Institute, Texas A&M University, College Station, June 1965.
5. T. Jones and R.B. Potts. The Measurement of Acceleration Noise--A Traffic Parameter. Operations Research, Vol. 10, 1962, pp. 745-763.
6. J. Lee and J.C. Yu. Energy Change Noise: A Measure of the Quality of Freeway Traffic. Traffic Engineering, Feb. 1974, pp. 28-35.
7. R. Herman, E.W. Montroll, R.B. Potts, and R.W. Rothery. Traffic Dynamics: Analysis of Stability in Car Following. Operations Research, Vol. 7, 1959, pp. 86-106.
8. R.T. Underwood. Acceleration Noise and Traffic Congestion. Traffic Engineering & Control, July 1968, pp. 120-123.
9. Highway Capacity Manual 1965. Special Report 87. HRB, National Research Council, Washington, D.C., 1965, 411 pp.
10. R.P. Roess, W.R. McShane, E.M. Linzer, and L.J. Pignataro. Freeway Capacity Analysis Procedure. ITE Journal, Dec. 1980, pp. 16-21.
11. D.R. Drew, C.L. Dudek, and C.J. Keese. Freeway Level of Service as Described by an Energy-Acceleration Noise Model. In Highway Research Record 162, HRB, National Research Council, Washington, D.C., 1967, pp. 30-85.
12. D.R. Drew. Traffic Flow Theory and Control. McGraw-Hill Book Company, New York, 1968.

13. D.R. Drew and C.J. Keese. Freeway Level of Service as Influenced by Volume and Capacity Characteristics. In Highway Research Record 99, HRB, National Research Council, Washington, D.C., 1965, pp. 1-47.
14. D.L. Gerlough and M.J. Huber. Special Report 165: Traffic Flow Theory--A Monograph, TRB, National Research Council, Washington, D.C., 1975, 222 pp.
15. SAS Users Guide: Basic. Statistical Analysis System Institute, Cary, N.C., 1982.
16. SAS Users Guide: Statistics. Statistical Analysis System Institute, Cary, N.C., 1982.
17. SAS/GRAPH Users Guide. Statistical Analysis System Institute, Cary, N.C., 1981.

Publication of this paper sponsored by Committee on Highway Capacity and Quality of Service.

Highlights of the Canadian Capacity Guide for Signalized Intersections

S. TEPLY

ABSTRACT

A unified approach to the treatment of capacity-related issues in urban networks has been emerging in Canada during the last 10 years. In 1982 the Executive of District 7 (Canada) of the Institute of Transportation Engineers appointed a committee to develop a series of documents that, eventually, will form a Canadian Urban Transportation Capacity Guide. The committee decided to proceed with the section on signalized intersections as the first task. The main reason for this decision was that the capacity of traffic signals is usually the key factor in all urban capacity considerations, and, as a result, a chapter on signalized intersections was most urgently needed. Moreover, a number of analytical and design procedures related to traffic signals have been tested in the Canadian context in the past decade. Although capacity research and development have been only marginally coordinated in Canada, a common philosophy has been forming, as may be seen in documents prepared in Ontario and Alberta. The first edition of the Canadian Capacity Guide for Signalized Intersections was preceded by three draft versions that were discussed both within and outside the committee. One of the guide's principal objectives is to test the approach and procedures and to elicit comments from users and researchers on a country-wide basis. The objective of this paper is to inform the North American transportation research community about the document and to highlight its philosophy and associated techniques. In essence, capacity analysis is based on a lane-by-lane saturation flow procedure that allows for calibration to local community conditions.

The need for a specifically Canadian document on capacity arises mainly from differences in climate; driver behavior; structure of cities; traditional traffic engineering practices; and political, judicial, and legal systems compared with those of other countries. In addition, it has been recognized that, in a country as vast as Canada, there is a great need for a common philosophy that can accommodate a wide variety of regional issues. Such a philosophy has been forming (1-3).

The objectives of the guide (4) can be detailed as follows:

- To consolidate current Canadian practice and research and to emphasize common features of the techniques used in different regions,
- To make it possible to incorporate parameters specific to a community or region,
- To identify the "missing links" and to focus future development of Canadian practice on a common philosophy,
- To set up the background for such a philosophy, and
- To provide a direction for the future education of users without restricting the development of regional and individual expertise

Although the document should provide basic guid-

ance for both experienced and novice practitioners, it also leaves room for additional analytical or design considerations and procedures where appropriate or necessary. This is why it is titled a "guide" instead of a "manual."

The foundations of individual techniques of the guide are described in sufficient detail to allow users to judge the applicability to a specific case or to use shortcuts where appropriate. Consequently, no distinction is made between so-called "planning" and "operations" techniques.

USER-RELATED PRINCIPLES

The editor and the authors of the guide have followed the following principles to enhance user accessibility:

1. Understandability. Not all users will enjoy the benefits of an advanced traffic flow theory education. The guide should be a self-contained document for traffic engineering professionals who may have limited practical or research experience.
2. Explainability. The guide has not been written as a textbook. Nevertheless, it is expected that it will be used both in academic and in professional development courses. Consequently, the procedures of the guide should be easily explainable with the use of traffic flow theory and practice.
3. Scientific soundness and practicality. It has been difficult to combine these two aspects. The guide, however, is mainly intended to serve practicing professionals, and, as a result, some compromises with theory were necessary. For example, although Webster's treatment of the left-turning movements with opposing traffic (5) has been considered theoretically most correct, a simplified, more easily understandable procedure has been incorporated. The results are accurate within less than 10 percent, which is justifiable from a practical point of view.
4. Accountability. The procedural setup of the guide should not only make it possible to incorporate local and regional parameters but, possibly more important, provide an opportunity to verify individual computational steps by direct measurements of existing situations. For that reason, volume manipulations have been restricted to a minimum and values of saturation flows adjusted to local conditions are directly verifiable by brief surveys. The use of local insight is advocated throughout the guide.

Most of the techniques have been tested and evaluated in terms of their theoretical soundness and practicality for Canadian conditions. An experienced user can, however, complement the analytical and design techniques outlined in the guide by using more detailed procedures in those cases in which the guide indicates a critical issue.

The guide covers individual, not necessarily isolated, signalized intersections. The strategic and network context is discussed and systemwide objectives are emphasized. Such analytical or design objectives must be reflected in performance goals. For example, although equalizing delays on all approaches may be a worthwhile goal at an intersection of two major arterial roadways, longer delays may be desirable at the collector approach of an intersection with an arterial street in order to minimize shortcutting through a residential neighborhood.

VOLUME: FLUCTUATIONS AND CONVERSIONS

An import distinction is made between "demand" and "supply" volume. Traditional intersection surveys consider only the volumes of individual movements within the intersection space. This represents the supply volume because it is limited by the capacity supplied by the geometric and timing features of the facility. Consequently, the supply volume-to-capacity ratio cannot exceed 1.0.

Demand volume represents the number of vehicles approaching the intersection. It is counted at the end of an intersection approach queue or derived from demand models. As a result, under growing queue conditions, the demand volume-to-capacity ratio is greater than 1.0. The guide generally employs the demand volume.

Changes of the demand in time represent its fluctuations. Although 1 hr is acceptable in many situations, the guide also recommends the use of rate of flow based on shorter periods for other conditions, such as for smaller communities. All flows are, however, expressed as hourly volumes. As will be demonstrated later, delay (the major evaluation parameter) is quite sensitive to the time base.

Vehicle categories are converted into passenger car units (pcu's) both for volumes and for saturation flows. Typical Canadian conversion factors (pcu equivalents) have been determined using a least-squares optimization procedure that reflects the composite effect of individual vehicle categories (6). They are as follows:

Vehicle Category	Passenger Car Unit Equivalent (pcu/veh)
Passenger cars, vans, pickups	1.0
Single-unit trucks	1.5
Multiunit trucks	2.5
Multiunit trucks heavily loaded	3.5
Buses or streetcars	1.75
Motorcycles	0.5
All trucks and buses combined (approximation)	2.0

As a rule, the guide uses lane-by-lane analytical or design techniques. Consequently, all volumes and saturation flows must be expressed separately for each lane.

SAFETY CONSIDERATIONS

Because certain safety parameters form inviolable constraints to capacity considerations, they are explicitly discussed in the guide for both vehicular and pedestrian movements.

Vehicular Requirements

Two features of vehicular safety are included. They involve the duration of the amber interval and the total intergreen period.

Amber interval can be determined by two methods. Both of them use a common formula but differ in the recommended acceleration rates. Nevertheless, they somewhat simplify the true decision- or dilemma-zone problem. As can be seen in Figure 1 the driver of a vehicle that is in the decision zone at the onset of amber has a choice of continuing at a steady speed (v) and legally entering the intersection during the last portion of the amber interval, or of stopping at the stop line at a reasonable deceleration rate. Typical perception and reaction time (t_{pr}) is 1.0 sec; corresponding distance is denoted d_{pr}. Specific adjustments are recommended for steep downhill intersection approaches.

The intergreen period is defined as the time between the end of the green interval for the traf-

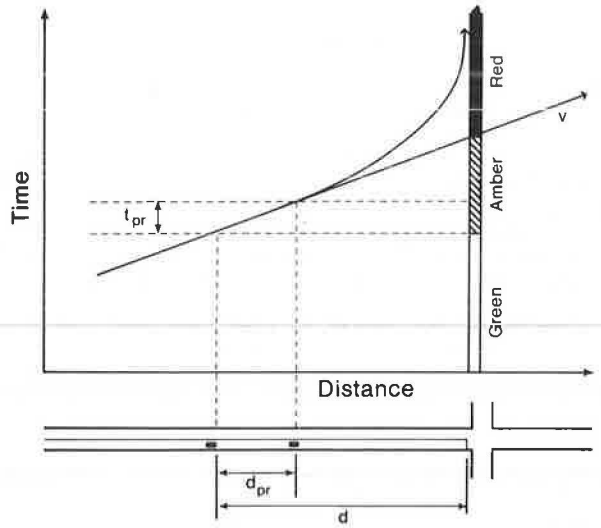


FIGURE 1 Determination of the amber interval.

fic stream losing the right-of-way and the beginning of the green interval for the conflicting traffic stream gaining the right-of-way. It consists of an amber interval and an all-red period.

The proposed technique to minimize potential vehicle-to-vehicle conflicts is based on the need to clear the last vehicles of the ending phase before the approaching vehicles of the starting phase reach the potential conflict area (Figure 2). The last vehicle legally entering the intersection during amber must clear the conflict area (Z_1 or Z_2, respectively) before the first vehicle of the starting phase, legally entering the intersection at the onset of green with a "rolling" start, reaches the conflict area. The formula includes the amber overrun portion, time needed to reach the conflict area, time needed to clear the conflict area (based on the length of a passenger car because longer vehicles act as a barrier), less the time needed by the first vehicle of the starting phase to reach the near end of the conflict area. The suggested range of input values reflects regional differences in driver behavior and in legal practices.

For vehicles clearing and pedestrians approaching, an identical procedure is recommended, but the length of a passenger car is excluded. When a vehicle is in the crosswalk area, it acts as a barrier.

The all-red interval fills up the remaining time between the amber interval and the intergreen period.

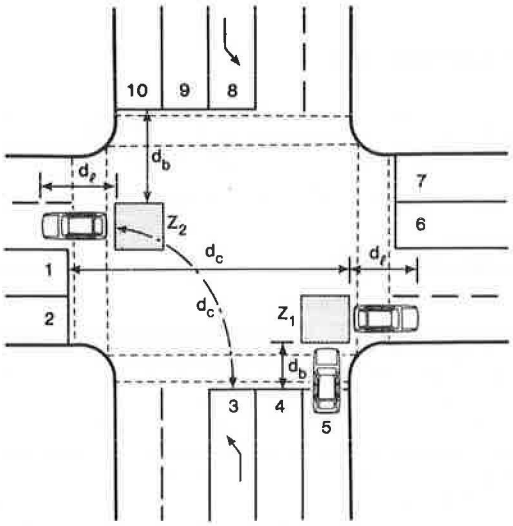
Pedestrian Requirements

For pedestrians, duration of walk intervals and clearance periods must be determined. These depend on the availability and dimensions of refuge areas. Again, basic parameters and considerations are identified. Naturally, the sum of the longest pedestrian walk interval and the clearance period for each phase must not exceed the sum of the longest vehicular green interval and the associated intergreen period for that phase.

CAPACITY ANALYSIS AND SIGNAL DESIGN

In signal operations there are two major tasks related to capacity. The first one is the quantitative evaluation of the performance of existing or planned signalized intersections (titled "Analysis" in the guide) and the second one is the design of new facilities. Both tasks are schematically shown in Figure 3.

The analytical task is based on a definition of the basic or the initial saturation flow for a given community or region and identifies the adjustment procedures for specific intersection conditions. Many individual geometric or traffic factors are included.



Z_1: Conflict Area for Lanes 2 and 5
Lane 2 Clearing (ending phase)
Lane 5 Approaching (starting phase)

Z_2: Conflict Area for Lanes 3 and 10
Lane 3 Clearing
Lane 10 Approaching

FIGURE 2 Determination of the intergreen period.

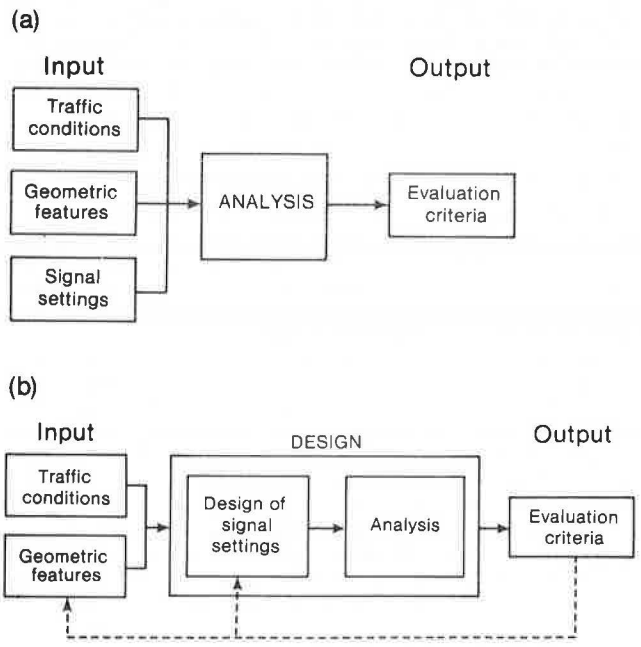


FIGURE 3 Schematic illustration of the signal operations analytical process (a) and schematic illustration of the signalized intersection design process (b).

Quantified performance parameters constitute the output of the analytical task. Capacity is naturally preeminent, but other expressions of capacity, such as volume-to-capacity ratios, reserve capacity, and delays, are also defined.

The design process concentrates on finding the best set of signal settings to suit local conditions, desired objectives, and performance parameters. This process is concerned with phasing schemes, cycle time determination, and green allocations. The suggested procedures use volume-to-capacity ratios or, alternatively, the probability of clearance, delays, and queues as tools to achieve the specified goals.

SATURATION FLOW CONCEPT

The saturation flow concept is shown in Figure 4. Three types of saturation flow are defined in the guide: basic, initial, and adjusted.

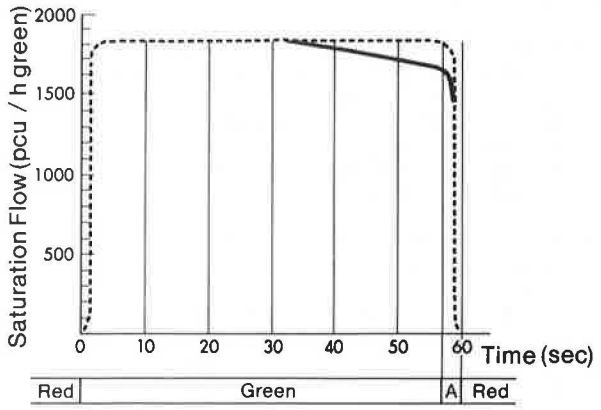


FIGURE 4 Saturation flow concept indicating a decline after about 40 sec of green (solid line).

Basic saturation flow is defined as the number of passenger car units that can discharge across the stop line of an "ideal" intersection lane (width 3.0 to 3.5 m) and move straight through (i.e., no turning movements) without any additional traffic friction (e.g., no parking, no bus stops). Ideal Canadian weather conditions during an optimum length of the green interval must also be included. The value of the basic saturation flow provides a good stable measure of driver behavior in a given community and, as such, can be used as a comparative indicator. Canadian research (6) suggests that not only the population size of a community but also its socioeconomic features influence saturation flows. Another important finding confirmed by international experience (7) indicated that the duration of the green interval should be considered because saturation flow declines after about 30 to 40 sec of green (solid line in Figure 4).

Initial saturation flows reflect a set of typical conditions that modify intersection performance. Long-lasting winter driving conditions and intersection environments are identified in the guide. It is recommended that a set of initial saturation flow values be developed for every community or region (based on local investigations). The differences may be significant as shown in Figure 5, which shows the typical ranges of basic saturation flow for summer and corresponding initial saturation flows for winter conditions as a function of the duration of

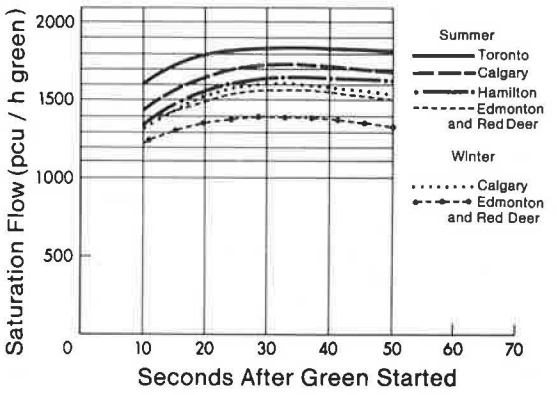


FIGURE 5 Typical ranges of Canadian basic and initial saturation flows in a cumulative average format.

the green interval for five Canadian cities. The values are depicted in a cumulative average format for a given green interval. For example, the values shown at 20 sec after green started show the average saturation flows for green intervals of 20 sec (i.e., between seconds 0 and 20). As a consequence of this format, Figure 5 cannot be directly compared with Figure 4.

It has been found that the impact of weather conditions becomes important only with major changes. Although citizens of other countries may find it somewhat strange, professional traffic engineers will understand that, from an intersection capacity point of view, typical Canadian summer conditions may include temperatures as low as -10°C (as long as the roads are not slippery). Typical winter conditions are characterized in the guide by low temperatures (-10°C to -30°C), pavement dry or well sanded, and exhaust fumes restricting visibility. Extreme winter conditions are defined by lower temperatures or heavy snowfalls.

The recommended intersection environment classification is based on the geometric standards and the general activity level, which is usually associated with adjacent land use. For example, a typical central business district intersection approach lane features a low geometric standard in combination with a high level of adjacent land-use activities, such as many pedestrians, frequent loading and stopping, and dense spacing of access points. This means that the initial saturation flow value for such an intersection is substantially lower than that for an industrial area with a high standard of roadway design and little interference from land-use activities. The user must consider individual approaches to the same intersection independently because they can belong to different categories, depending on their dominant features.

Initial saturation flow is used as a starting point for the incorporation of specific intersection conditions. The resulting value is termed the adjusted saturation flow. For existing intersections, this value can also be determined by direct measurements--or, if it was calculated, it can be verified by short surveys.

Procedures for the determination of the adjusted saturation flow include the following situations for through lanes:

1. Geometric conditions:
   - Lane width,
   - Gradient, and
   - Queueing and discharge space.
2. Traffic conditions:
   - Public transit,

- Parking, and
- Duration of the green interval.

Adjustments for left- or right-turn movements may include applicable procedures for the through lanes. In addition, the following geometric or traffic considerations may be necessary:

1. Geometric condition--turning radius.
2. Traffic conditions:
   - Opposing traffic flows,
   - Pedestrians, and
   - Effect of movement combinations that share one lane.

Figure 6 shows a summary of the applicability of different saturation flow adjustments to various lane function combinations.

	Possible Adjustments							
Lane Function	Lane Width	Radius	Grade	Queueing Space	Discharge Space	Bus Stops	On-street Parking	Length of Green
Straight Through and Left	•		•	•	•	•	•	•
Straight Through and Right	•	•	•	•	•	•	•	•
Straight Through Only	•		•	•	•	•	•	•
Left Only	•		•	•		•		•
Right Only	•	•	•	•		•		•
Straight Right and Left	•	•	•	•	•	•	•	•

FIGURE 6 Overview of initial saturation flow adjustments.

Individual cases are discussed in detail and procedures for the adjustment of initial saturation flows are provided. This section is a major part of the guide because the saturation flow concept constitutes its backbone. Unfortunately, the scope of this paper does not allow a description of the individual procedures. Suffice it to say that the techniques have not only been verified under Canadian conditions, but, perhaps more important, they allow for calibration to specific conditions of a given community or region.

If an approach lane features a combination of several factors for which adjustments of the initial saturation flow are necessary, the resulting adjusted value is not necessarily a multiplicative product of individual adjustments. In many instances, one of the factors will govern. For example, where the saturation flow for right-turning traffic is controlled mostly by a high pedestrian flow rate in the adjacent crosswalk, an additional adjustment for a tight radius is not appropriate. Most right-turning vehicles will have to stop, practically eliminating the effect of the radius.

The techniques suggested for the determination of adjusted saturation flows and the allocation of volumes to lanes with more than one movement (shared lanes) also employ auxiliary turning-movement factors that are determined as ratios of the initial saturation flow and the adjusted saturation flow for individual movements. This factor reflects a specific degree of difficulty in making a right or left turn under given circumstances. However, when the lane assignment has been completed, the volume is converted back to the real, measurable volumes.

INTERSECTION PERFORMANCE CRITERIA

Even though capacity in itself can be used as an absolute comparative measure, it does not reflect the operation of a signalized intersection relative to traffic demand. To this end, the following evaluation criteria are used in the guide:

- Lane (or approach) volume-to-capacity ratios,
- Intersection volume-to-capacity ratio,
- Lane reserve capacity,
- Intersection reserve capacity,
- Average lane delays,
- Average intersection delay, and
- Probability of discharge.

The guide employs the probability of discharge (clearance) as an alternative measure for the design task.

Most of the evaluation criteria that relate volume to capacity are well known and will not be discussed in this paper. It should be noted, however, that none of them in itself can fully represent the complexity of functions and objectives. The guide has not attempted to combine them in a single measure, but instead recommends simultaneous assessment.

Delays for individual lanes and the overall intersection represent the most powerful and practical tool for the evaluation of performance because they relate directly to drivers' perception. Total delay is expressed as a sum of uniform and overflow delay.

The uniform delay equation in the guide uses the queueing theory relationship [see Figure 7(a)].

The TRANSYT-7 computer program provided an overflow delay formula suggested by Whiting (8). The program has been extensively used in Canada for a number of years with satisfactory results. The combination of uniform and overflow delay is shown in Figure 7(b). Naturally, even in well-undersaturated conditions, an occasional cycle may feature an overflow.

The delay equation was tested by independent investigations in Toronto and in Edmonton. The formula matched field conditions with a high degree of accuracy. It should be emphasized, however, that for saturated or oversaturated conditions, the evaluation time (i.e., duration of the congestion) has an overriding effect (Figures 8 and 9). The slopes of the delay functions for 60-, 30-, and 15-min periods differ dramatically. In practical analytical or design problems, the length of the congested period can rarely be determined accurately and, consequently, the delay values for oversaturated conditions should be taken as an indication of the magnitude of the problem, not as absolute, accurate values.

The acceptance of the Whiting delay equation in Canada is also evidenced by the range of users of the SINTRAL computer program system for signalized

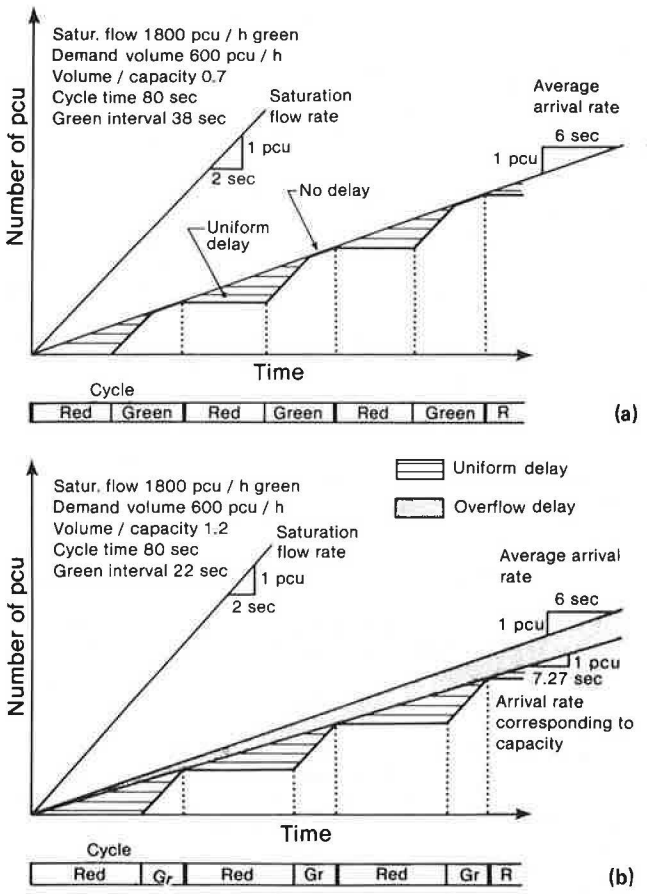


FIGURE 7 Queueing diagrams illustrating the basic delay considerations: idealized conditions well below saturation (volume-to-capacity ratio is 0.7) (a) and idealized conditions when demand exceeds capacity (volume-to-capacity ratio is 1.2) (b).

intersection analysis and design (9) that incorporates the formula.

The guide identifies the detailed delay formula and includes graphic representation of the most common conditions. Figures 8 and 9 are only two of the four graphs used in the guide. These diagrams

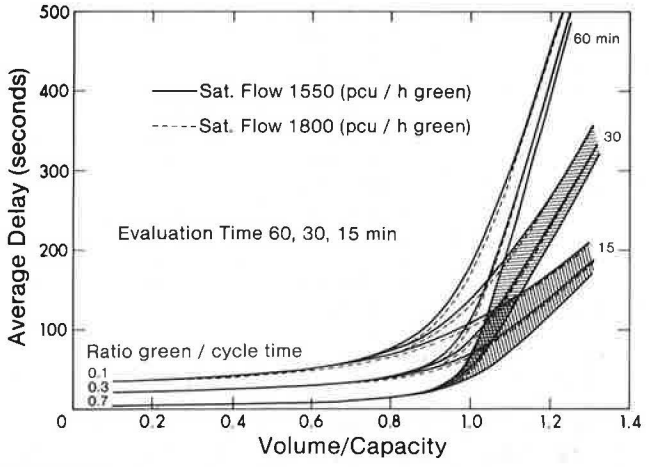


FIGURE 8 Average lane delay as a function of volume-to-capacity ratio (0.1 to 1.3), saturation flow (1,800 and 1,550 pcu/hr green), green interval-to-cycle time ratio (0.1, 0.3, and 0.7), and evaluation period (duration of congestion) (60, 30, and 15 min).

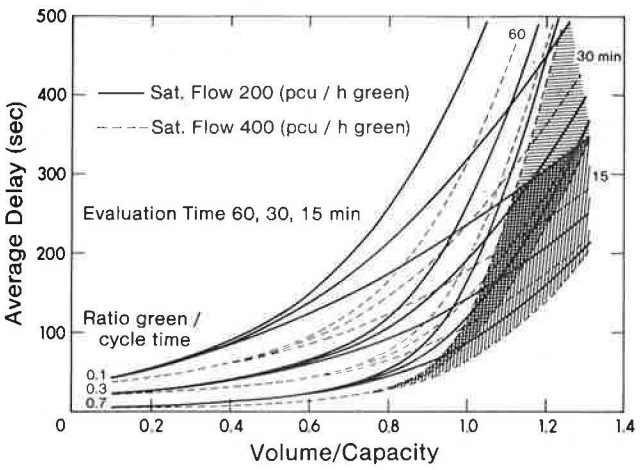


FIGURE 9 Average lane delay as a function of volume-to-capacity ratio (0.1 to 1.3), saturation flow (400 and 200 pcu/hr green), green interval-to-cycle time ratio (0.1, 0.3, and 0.7), and evaluation period (duration of congestion) (60, 30, and 15 min).

can be used for the manual determination of average lane delay.

Because of the shape of the delay function, a volume-to-capacity ratio of 0.9 is generally considered a practical capacity limit. Differences in the delay due to saturation flow and the ratio of the green interval to cycle time become essential for lower values of both parameters. As a consequence of the original Webster cycle time-delay relationship (5) cycle time itself has only a minor impact as long as it remains in the range of 0.75 to 1.5 of the "optimum" cycle time (Figure 10).

DESIGN PROCESS

The principles of signal design are shown in Figure 3(b). The task employs tentative geometric features

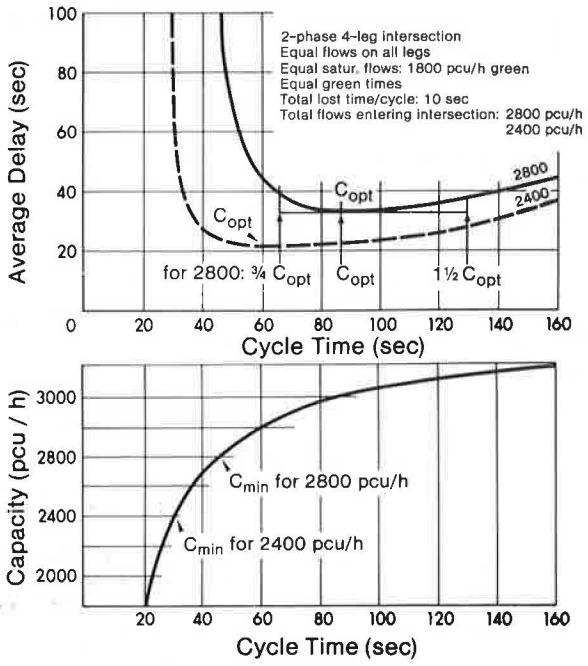


FIGURE 10 Relationship among cycle time, capacity, and delays.

and traffic volumes as the input, and the detailed signal settings constitute the output that may then be compared to the desired performance criteria. Iterative adjustments are usually necessary.

The determination of the minimum and optimum cycle time follows the well-known Webster methodology (5) that needed no adjustment for Canada (Figure 10). In addition, special pedestrian needs in regard to cycle time are considered. The user is also advised to consider systemwide issues, such as coordination.

The following techniques can be used for the allocation of green intervals within cycle time.

- Proportioning volume-to-saturation flow ratios,
- Equalization of the probability of discharge,
- Delay minimization, and
- Congestion management.

The technique of proportioning the volume-to-saturation flow ratio concentrates on the critical lanes. It allocates green intervals for individual phases in the proportion of critical lane volume-to-saturation flow ratios to the total intersection volume-to-saturation flow ratio. This method has been widely used in Canada in the past.

The probability-of-discharge technique attempts to balance the probability of waiting more than one cycle for the critical lanes of individual phases. The Poisson distribution of arrivals in used because it is not only simple (one parameter distribution) but, more important, it accurately represents the traffic conditions in most Canadian situations.

Two applications of this technique are described: starting from green intervals determined by proportioning the volume-to-saturation flow ratio and starting from the highest desired probability of discharge (9).

The delay allocation design technique assigns the green intervals for the critical lanes of individual phases in such a way that the overall intersection delay is minimized or is in agreement with other design objectives.

The goal of the congestion management green allocation technique is to prevent queues from "spilling over" the available storage space. In some instances, a revision of the previously determined cycle time may be needed.

OTHER CONSIDERATIONS

Although the guide is primarily concerned with the operation of individual signalized intersections, system aspects are emphasized. They may be expressed in the strategic objectives of the design, such as the reduction of shortcutting through residential areas. In such a case, the green allocation objectives will be expressed as "minimization of delays on arterial roadways and a maximum allowable delay on the collector or residential approaches." Coordination of signal operation along a route or in a network may significantly modify the arrival pattern and, as a result, the selection of evaluation criteria and the subsequent design method.

CONCLUSIONS

The currently available edition of the Canadian Capacity Guide for Signalized Intersections is based on international as well as Canadian research and practice. Nevertheless, it has been acknowledged during the development stage of this edition that more theoretical as well as empirical research is needed. Many of the outlined procedures and values should be subjected to real-life tests in different regions of Canada. A country-wide uniformity would not be a worthwhile objective in itself if it were not based on common foundations in driver behavior, traffic engineering, and legal practices.

Several new features of signal design and analysis can be identified in the first edition:

1. Calculations are based on demand instead of supply volumes;
2. Typical Canadian passenger car unit equivalents for different vehicle categories have been established on the basis of their interactive flow impact;
3. Amber intervals, intergreen periods, and pedestrian safety requirements are inherent components of the procedure;
4. Basic saturation flow values facilitate the transferability of the process;
5. Saturation flow values vary with socio-economic characteristics of a community not just its size;
6. The effect of ambient conditions is identified;
7. The impact of long green intervals is quantified;
8. Webster's procedure for determining saturation flow for left turns with opposing traffic is simplified;
9. Factors for saturation flow adjustments are not necessarily multiplicative;
10. Turning factors based on saturation flow values are introduced as auxiliary measures in the procedures for volume allocation to lanes and for determination of adjusted saturation flows for lanes with a combination of movements;
11. The procedure allows verification of intermediate results for existing situations;
12. Several simultaneous criteria are used for the assessment of capacity-related intersection performance; and
13. Green interval allocation parameters can be selected on the basis of design objectives.

The editor, the authors, and the committee anticipate significant contributions from users. Several ITE sections have already established committees to this end. Periodic revisions will also be needed in the future.

ACKNOWLEDGMENTS

The author of this paper would like to express his sincere appreciation to the coauthors of the guide. The commitment and effort of Dave Richardson, Brice Stephenson, and John Schnablegger made it possible to incorporate current Canadian research and practice and to devise many of the modifications.

The Canadian District of the Institute of Transportation Engineers has taken a bold step in assuming a leading role in the development of the guide.

The National Sciences and Engineering Research Council of Canada has supported some of the basic research for many years through operations grants.

Toronto, Edmonton, Hamilton, and several other communities assisted in data collection and in the development of the techniques incorporated in the guide.

Last, but certainly not least, the author would like to thank many of the individuals, within or outside the committee, who provided valuable comments or direct assistance on the drafts of the guide. Moral support from many Canadian colleagues was also extremely important, especially during the difficult revision and rewriting stages. The suggestions of the reviewers of this paper are also gratefully acknowledged.

Discussion

Robert H. Wortman*

The Canadian Capacity Guide for Signalized Intersections represents a significant contribution to the literature on highway capacity in that it presents material, concepts, and numerical values that have been found to be pertinent to conditions in Canada. All of those who contributed to the guide are to be commended for their efforts that resulted in this document.

The publication of the guide comes at a time when the signalized intersection chapter of the new Highway Capacity Manual is being put in final form and prepared for release. A review of the two documents reveals similarities as well as marked differences in philosophies and approaches. For example, both documents use saturation flow concepts as the basis for the capacity calculations, and rate of flow is preferred to the use of hourly volumes. Although delay is used as the measure of level of service for signalized intersections in the new Highway Capacity Manual, the Canadian guide suggests delay along with volume-to-capacity ratio and reserve capacity as measures of the adequacy of intersection operation.

Even though delay appears to be a common measure in both documents, there are differences that should be noted by users. The Canadian work uses average delay and the delay equation from TRANSYT-7. In contrast, the new Highway Capacity Manual will employ stop-time delay and a delay equation that has been modified to reflect the findings of field studies. Furthermore, there is some variation in the specific values of delay that define the operational levels for intersections.

In the discussion of capacity values for left-turning vehicles, it is interesting to note the apparent variation in values that was observed in various Canadian cities. Although the equation for permissive left-turn conditions is based on an empirical relationship developed in Edmonton, typical maximum values for other cities are also given. This certainly serves to support the observation that considerable variation in values can be found in different locations. The field data that have been used in the development of the new Highway Capacity Manual show similar results.

The Canadian guide contains material in a number of areas that should be of interest to those who are involved in capacity analyses. There is a quantification of the effect of winter conditions on capacity as well as of the influence of right-turn-on-red situations. Also, there is a discussion of network objectives that may be considered in evaluating operation and design problems. Basically, the network objectives reflect policies that give priority to certain movements or approaches.

Further work in several areas could possibly improve and enhance the current version of the guide. This work should include consideration of the following comments:

1. Average signal delay generally increases as cycle length increases; thus the average delay values as shown in the guide may not be appropriate when cycle lengths are long.
2. It would appear that the quality of progression is a major variable that is not included in computing delay. With volume-to-capacity ratios of less than 0.8, the average delay with excellent progression can be significantly less than for a condition with adverse progression.
3. The maximum delay for acceptable operation is defined as 30 sec; this value may be too large for intersections where motorists expect to benefit from signal progression.
4. The variation in the flows for left turns has been noted; however, it would help the user if the expected variation in all of the values and predictive equations used in the capacity calculations were indicated. In essence, the user should be apprised of the basis and validity of all of the numerical values that are shown in the guide.

In closing, comments should also be directed to all individuals and groups who are involved with the development of capacity procedures as well as to the users of those procedures. For those who are involved in the development of procedures, there is a need to resolve the differences between the various procedures, approaches, and computations. Such differences tend to confuse the user community. With respect to the users, there is a need to become thoroughly familiar with the procedures. The proper application of capacity procedures requires that the user have a certain level of understanding of the material and use good judgment in applying the guidelines that have been developed.

*Department of Civil Engineering, University of Arizona, Tucson, Ariz. 85721

Author's Closure

Robert Wortman's comments on the Canadian Capacity Guide for Signalized Intersections are sincerely appreciated.

During the past 20 years techniques for capacity analysis have converged significantly. Nevertheless, a deep understanding of the applicability of the procedures to local and regional conditions has been lagging somewhat behind the "philosophy." The guide and the new edition of the Highway Capacity Manual as well as other national documents play an important role in the development of a common basis on which we can compare our experience. It is not surprising that some of the specific values are different, but it is encouraging to see that major trends are similar.

The first edition of the guide will soon be critically reviewed. Wortman's suggestions will certainly be included in that review. At this time, however, I would like to point out that the committee deliberately divorced the broader "systems" aspects from individual (not necessarily isolated) intersection issues. Even a relatively simple linear or network coordination of traffic signals introduces an additional dimension to the problem and usually implies a modified set of objectives. Insofar as delay is considered a measure of performance its values at individual intersections may no longer govern. Although an overall delay reduction would be expected in such a system, some intersection approaches may experience a delay increase for the benefit of the larger system. Depending on a designer's objectives, delay may be supplemented by other measures, such as a performance index consisting of delay, running time, and a weighted number of stops. In some systems, the number of stops along certain network links may become the sole criterion.

Having stated that, I should note that I agree with Wortman that a better treatment of these issues is needed. The users must exercise a great deal of judgment and we should provide them with the best information we can give within the scope of the guide.

REFERENCES

1. D.B. Richardson. The Metropolitan Toronto Intersection Capacity Guide, 4th ed. Municipality of Metropolitan Toronto, Ontario, Canada, 1982.
2. Ontario Manual of Uniform Traffic Control Devices. Ontario Ministry of Transportation and Communications, Downsview, Ontario, Canada, 1981.
3. Saturation Flow Manual. The City of Edmonton and the University of Alberta, Edmonton, Canada, 1980.
4. D.B. Richardson, J. Schnablegger, B. Stephenson, and S. Teply. Canadian Capacity Guide for Signalized Intersections. Institute of Transportation Engineers, District 7, University of Alberta, Edmonton, Canada, Feb. 1984.
5. F.V. Webster and B.M. Cobbe. Traffic Signals. Road Research Technical Paper 56. Her Majesty's Stationery Office, London, England, 1966.
6. S. Teply. Saturation Flow at Signalized Intersections Through a Magnifying Glass. Proc., Eighth International Symposium on Transportation and Traffic Flow Theory 1981, Toronto University Press, Ontario, Canada, 1983, pp. 588-622.
7. Traffic Capacity of Major Routes. OECD Road Transport Research, July 1983.
8. D.I. Robertson. Traffic Models and Optimum Strategies of Control--A Review. Proc., International Symposium on Traffic Control Systems, Vol. 1, Berkeley, Calif., 1977, pp. 262-288.
9. J. Schnablegger, B. Stephenson, and S. Teply. Edmonton Version of Critical Lane Analysis: SINTRAL, An Interactive Computer Program System. ITE Journal, Aug. 1981, pp. 20-26.

Publication of this paper sponsored by Committee on Highway Capacity and Quality of Service.

Signal Delay with Platoon Arrivals

JAMES M. STANIEWICZ and HERBERT S. LEVINSON

ABSTRACT

Delays at signalized intersections assuming "platoon" flow are analyzed. Graphic analysis of vehicle platoon arrivals is used to develop equations from which the average travel time delay per vehicle can be estimated. Delay for two different, basic conditions is analyzed: (a) when the first vehicle in the platoon arrives during a green interval and is unimpeded and (b) when the first vehicle in the platoon arrives during a red interval or is impeded by queued vehicles. Delay based on the resulting relationships is compared with delay obtained by three conventional methods: the Webster method, May's continuum model method, and the new 1985 Highway Capacity Manual method. Where the platoon leader is unimpeded, there is no delay when the capacity of the through-band equals or exceeds the approach volume. Thus, a high volume-to-capacity ratio may provide a high level of service. This contrasts with delays based on random or uniform arrivals, which are sensitive to the volume-to-capacity ratio. However, where the first platoon vehicle is impeded by a red interval or by queue interference, a chain reaction may occur in which following vehicles are also impeded. This situation may create considerable delay and effectively reduce progression. Effective traffic signal coordination, therefore, can substantially reduce delay and improve levels of service.

Delay has become an important means of assessing level of service at signalized intersections. Consequently, accurate measurements of this delay are essential. Delay computations and computer simulations often assume uniform or random vehicle flow, singly or in combination. However, where signals are spaced closely together or form part of a progressive system, platoon flows are common and more closely represent reality. Such cases result in a different pattern of delays.

Delays at signalized intersections are analyzed assuming platoon flow instead of a random or a uniform arrival pattern. The following question is addressed: What average delay does a platoon of traffic encounter at a signalized intersection? A simple graphic analysis of vehicle platoon arrivals

is used to develop equations for estimating average travel time delay per vehicle. Two basic conditions are investigated: (a) when the first vehicle in the platoon arrives during a green interval and is unimpeded and (b) when the first vehicle in the platoon arrives during a red interval or is impeded by queued vehicles. The delay based on the resulting relationships is compared with the delay obtained by three conventional methods: the Webster method, May's continuum model method, and the new 1985 Highway Capacity Manual method.

BASIC ASSUMPTIONS AND PARAMETERS

The analyses relate to arterial street traffic approaching traffic signals. It is assumed that the vehicles have been grouped in platoons by signals upstream of the intersection under study. The procedures for estimating delay depend on the arrival condition of the first platoon vehicle. The platoon leader may arrive during a green interval and proceed unimpeded (Case 1) or arrive during a red interval or be impeded by queued vehicles (Case 2).

Analytical relationships were derived from graphic analysis of each time-space pattern. These relationships were keyed to vehicular volume as measured by the number of through passenger car units per lane per cycle. Green and red periods include the usable and unusable portions of the clearance interval, respectively.

The following assumptions underlie the two delay models:

- All vehicles approaching the intersection arrive in a platoon;
- All vehicles in the platoon follow each other at a uniform time spacing (headway);
- All vehicles in the platoon travel at the same speed (speed of progression);
- The upstream approach speed equals the downstream departure speed; and
- The approach volume per cycle equals the departure volume per cycle; there are no oversaturated conditions.

Travel time delay is used in the analysis. It represents the difference between (a) the time it takes a vehicle whose approach speed to an intersection is altered to recover that speed downstream and (b) the travel time required if that vehicle were able to continue at its approach speed unimpeded. Thus, it includes the time decelerating from an approach speed, stopped time, reaction time, and the time to accelerate back to the same speed as on the approach (1). The average travel time delay per vehicle is the sum of the individual travel time delays divided by the number of vehicles involved.

For a vehicle that stops at an intersection, the rate of deceleration per se does not influence the travel time delay because that vehicle cannot enter and clear the intersection until the signal turns green. Therefore, it is the red time incurred assuming instantaneous deceleration at the stop line that effectively contributes to the travel time delay.

When the signal turns green, there is an initial reaction time. This time loss is followed by the time required to accelerate to resume the desired speed, which is assumed to be equal to the approach speed. A summary of the total reaction and acceleration time loss (L) is given in the following table and derived in Appendix A.

Speed (mph)	Lost Time (L) (sec)
20	4.5
25	5.2
30	5.9
35	6.6
40	7.3

When there are no queued vehicles at a signal approach, to avoid deceleration of a vehicle that is approaching an intersection while a red signal is displayed, the signal must turn green at or before the time when the vehicle would begin to decelerate. Values of this time offset (t_d) for various approach speeds are given in the following table and derived in Appendix B.

Speed (mph)	t_d (sec)
20	2.7
25	3.1
30	3.6
35	4.7
40	6.1

When vehicles are queued at an intersection,

$$t_d = L + H_D(S) \quad (1)$$

where

t_d = travel time offset to avoid deceleration of first platoon vehicle (sec),
L = lost time (reaction and acceleration loss) in seconds,
H_D = departure headway (sec/vehicle), and
S = number of queued vehicles (vehicles/cycle/lane).

PLATOON FLOW MODELS

The delay equations vary for each of the two basic flow models. Accordingly, it is important to determine whether the first vehicle in the platoon is (a) unimpeded or (b) impeded. A graphic check of the time-space diagram will readily indicate when the first vehicle arrives in relation to the start of the green interval. A further check is needed to ensure that, if the first vehicle arrives during a green interval, it is not impeded by queued vehicles.

Case 1: First Vehicle in Platoon Arrives on Green and Is Unimpeded

This arrival condition is shown in Figure 1. The first vehicle arrives at the intersection during a portion of the green interval and is not impeded by any queued vehicles. However, the tail portion of the platoon may arrive at the intersection on the red and then leave at the beginning of the next green interval.

If the first vehicle arrives at the intersection during a green interval and is not impeded by any queued vehicles, the number of vehicles that may pass unimpeded (T) can be determined from the following expression:

$$T = (W - t_d + H_A)/H_A \quad (2)$$

where

T = through-band capacity (vehicles/lane/cycle),
W = bandwidth (sec),

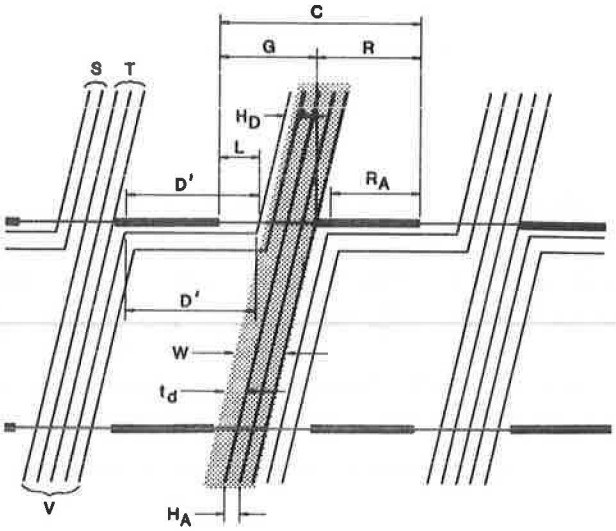


FIGURE 1 Case 1—graphic simulation, platoon leader unimpeded.

t_d = time offset to avoid deceleration of the first vehicle (sec), and
H_A = approach headway (sec).

The number of impeded vehicles (S) is the difference between the approach volume (V) and the number of unimpeded vehicles (T):

$$S = V - T \quad (3)$$

Total Delay

The length of the effective red time (R_A) that the first stopped vehicle must wait is equal to the red interval (R) minus one approach headway:

$$R_A = R - H_A \quad (4)$$

This is because the definition of the red interval (R) includes the unused portion of the clearance interval and commences when the last through vehicle in the platoon enters the intersection. Thus, the next vehicle in this platoon stops on the red but arrives one approach headway later.

The delay to the first stopped vehicle (D') equals the length of the red time (R_A) that it must wait until the start of the green, plus the reaction time and acceleration loss (L):

$$D' = R_A + L = \text{Effective impedance time} \quad (5\text{-}1)$$

The delay to each successive stopped vehicle is as follows:

$$\text{Delay to second stopped vehicle} = R_A + L + H_D - H_A \quad (5\text{-}2)$$

$$\text{Delay to third stopped vehicle} = R_A + L + 2(H_D - H_A) \quad (5\text{-}3)$$

and so forth, and

$$\text{Delay to last stopped vehicle (S)} = R_A + L + (S\text{-}1)(H_D - H_A) \quad (5\text{-}S)$$

where

R_A = effective red time (sec),
L = time loss due to driver reaction and acceleration (sec),
S = number of stopped vehicles (vehicles/lane/cycle),
H_A = arrival headway (sec/vehicle), and
H_D = departure headway (sec/vehicle).

The total delay (D_T) to S stopped vehicles represents the sum of the delays to the first, second, third . . . and Sth vehicle (i.e., the sum of Equations 5-1 through 5-S).

Thus,

$$D_T = S(R_A + L) + [0 + 1 + 2 \ldots + (S - 1)] \times (H_D - H_A) \quad (6)$$

This may be expressed as

$$D_T = S(R_A + L) + \sum_{i=1}^{i=S} (i - 1)(H_D - H_A) \quad (7)$$

where i represents the number of stopped vehicles ranging from 1 to S, or

$$D_T = S(D') + F(H_D - H_A) \quad (8)$$

where

$$F = \sum_{i=1}^{i=S} (i - 1)$$

Calculated values of F are given in the following table.

S	F	S	F
1	0	11	55
2	1	12	66
3	3	13	78
4	6	14	91
5	10	15	105
6	15	16	120
7	21	17	136
8	28	18	153
9	36	19	171
10	45	20	190

Average Delay

The average travel time delay per vehicle (D) is the sum of the individual delays (D_T) divided by the approach volume (V). It is given by the following expression:

$$D = D_T/V = [S(D') + F(H_D - H_A)]/V \quad (9)$$

where V is approach volume per lane per cycle.

This equation may be simplified as follows when the approach headway is equal to the departure headway:

$$D = S(D')/V = S(R_A + L)/V \quad (10)$$

Because S = V - T, Equation 10 becomes

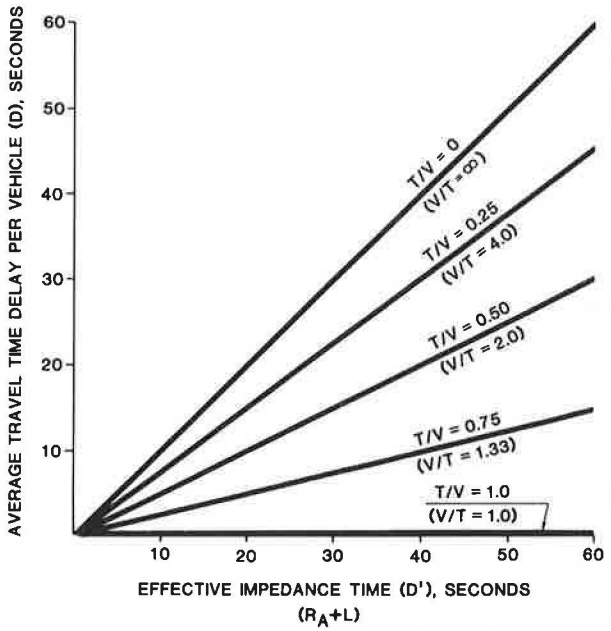
$$D = (R_A + L)(V - T)/V = (R_A + L)[1 - (T/V)] \quad (11)$$

or

$$D = (R_A + L)[1 - 1/(V/T)] \quad (12)$$

Thus, when the approach volume and the through-band volume are equal, there is no delay.

Figure 2 shows how the average travel time delay per vehicle (D) relates to the effective impedance time (D') for various percentages of through-band



*Assumes Base Speed = 30 M.P.H., Headway = 2.1 Seconds

FIGURE 2 Case 1—average travel time delay per vehicle versus effective impedance time.

volume (T) versus approach volume (V), assuming equal approach and departure headways.

As an example, when the effective impedance time is 30 sec and the ratio of the approach volume to through-band capacity (V/T) is 1.0, there is no delay. However, when this ratio is 2.0, the average delay per vehicle is 15 sec. Appendix C contains a sample calculation.

Case 2: First Vehicle in Platoon Arrives During a Red Interval or Is Impeded by Queued Vehicles

This arrival condition, shown in Figure 3, requires a somewhat different delay estimation procedure.

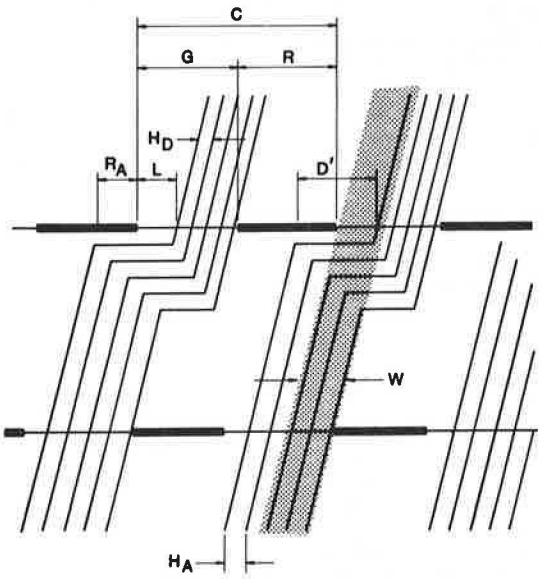


FIGURE 3 Case 2—graphic simulation, platoon leader impeded.

Because the first vehicle in the platoon arrives during a red interval or is impeded by queued vehicles, a chain reaction may result in which following vehicles are also impeded. The number of impeded vehicles and the total delay must be determined.

Total Delay

The delay to each stopping vehicle is calculated by the same method as in Case 1. However, in this second case the number of stopping vehicles first must be determined.

The number of stopping vehicles may be found by solving the delay equation (Equation 5-S) for the condition when the delay equals zero. This condition will occur for vehicle (S + 1). Substituting in Equation 5-S yields the following equation:

$$R_A + L + [(S + 1) - 1](H_D - H_A) = 0 \quad (13)$$

Solving for S, the number of stopping vehicles, gives

$$S = -(R_A + L)/(H_D - H_A)$$

or

$$S = (R_A + L)/(H_A - H_D) \quad (14)$$

The value of S calculated from Equation 14 cannot exceed the total number of arriving vehicles per cycle (i.e., $S \leq V$). Where a calculated value of S is greater than the approach volume (V), the value of S should be assumed to be equal to the approach volume (V). A graphic representation of this limit is shown in Figure 4.

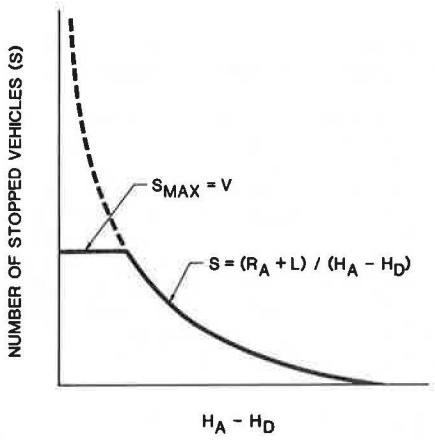


FIGURE 4 Case 2—Equation 14 limit.

The delay to the first stopped vehicle (D') is algebraically the same as for the first case (Equation 5-1):

$$D' = R_A + L$$

where R_A is effective red time (i.e., remaining red time after arrival of first impeded vehicle) in seconds and L is lost time (sec).

However, in this second case, the length of red time (R_A) that the first stopped vehicle must wait is determined graphically from the time-space diagram or relationship.

Average Delay

The average travel time delay per vehicle (D) is also the same as for the first case (Equation 9):

$$D = [S(D') + F(H_D - H_A)]/V$$

where

S = maximum number of vehicles impeded (vehicles/lane/cycle),
D' = delay to first stopped vehicle (sec),
$F = \sum_{i=1}^{i=S} (i - 1)$,
H_D = departure headway (sec),
H_A = arrival headway (sec), and
V = approach volume (= departure volume) in vehicles per lane per cycle.

For most applications, the approach headway will equal or exceed the departure headway of the impeded vehicles. When the approach headway equals the departure headway and the first vehicle is impeded, each vehicle in the platoon will be delayed the same amount of time. When the approach headway is greater than the departure headway, each subsequent stopping vehicle will be delayed less than the first and, depending on the difference between the approach and departure headways, some vehicles may not be impeded as the turbulence clears. Appendix D contains a sample calculation.

COMPARISON WITH OTHER METHODS

The delay estimates for both arrival conditions are compared with delays obtained from equations of three conventional methods: the Webster method, May's continuum model method, and the new 1985 Highway Capacity Manual method.

- The Webster delay formula adjusts uniform delay for random (Poisson) arrivals (2). (See Appendix E.)
- May's continuum model (2) assumes uniform, or regular, arrivals as a continuous function at a signal. (See Appendix F.)
- The new 1985 Highway Capacity Manual (HCM) (3) method employs uniform arrivals and includes an additional term to account for random arrivals. It then adjusts the delay based on platooning characteristics by reducing the delay for effective signal progression and increasing the delay for adverse progression conditions. (Appendix G contains the basic equation for average conditions.)

Delay may be measured in many forms. The Webster method uses approach delay, the continuum model method appears equivalent to approach delay, the HCM method uses stopped delay, and the platoon equations herein use travel time delay. Accordingly, adjustments were necessary so that all methods could be compared in terms of travel time delay.

Approach delay is similar to travel time delay except that it does not include acceleration losses beyond the intersection being evaluated. Assuming a 30-mph base speed and employing Greenshield's departure model (4,p.351), the additional acceleration losses beyond the intersection for each vehicle in the queue are given in the following table:

Vehicle in Queue	Additional Acceleration Loss Beyond Intersection (sec)
1	2.1
2	1.1
3	0.5
4	0.2
5 or more	0.1

In the Webster method or May's continuum model method, the estimated average travel time delay per vehicle (D) is equal to the approach delay (A) obtained directly from the method's equations, plus the sum of the additional acceleration losses beyond the intersection (Z) divided by the approach volume per lane per cycle (V):

$$D = A + Z/V \quad (15)$$

The second term in this equation (Z/V) represents the average acceleration loss beyond the intersection. This term has a relatively small contribution, typically less than 1 sec.

Stopped time delay represents the time spent while the vehicle is motionless. One source (1) estimates that approach delay equals the stopped delay multiplied by a factor of 1.3. Using the stopped delay (d) from the new 1985 HCM method, the average travel time delay per vehicle (D) was estimated by the following equation:

$$D = 1.3(d) + Z/V \quad (16)$$

Case 1: Comparison

Given in Table 1 and shown in Figure 5 is a comparison of the average travel time delay per vehicle for unimpeded platoons arriving on the green with delays estimated for the other methods. These exhibits show data for the following conditions:

- 60-sec cycle, 29 sec green, 31 sec red;
- Base speed = 30 mph;
- Departure headways = 2.1 sec per vehicle;
- Signal capacity per lane per cycle = 12 vehicles;
- Approach volumes for 3, 6, 9, and 12 vehicles per lane per cycle;
- Band capacities (platoon flow) for 3, 6, 9, and 12 vehicles per lane per cycle; and
- Arrival headways (platoon flow) = 2.1 sec per vehicle.

With platoon flow arrivals, there are no delays when the volume-to-band capacity ratio is less than or equal to 1.0. When the volume-to-band capacity ratio is greater than 1.0, delays will result. In some cases, the delays exceed those obtained by other methods.

To illustrate, for a volume-to-capacity ratio of 0.75, the following delays are computed:

Method	Delay (sec)
Webster	19.8
May	15.0
New HCM	
Average conditions	19.0
Ideal progression	11.8
Platoon Method at volume-to-band capacity ratio of	
3.0	23.2
1.5	11.6
1.0	No delay
0.75	No delay

Platoon delay depends on the volume-to-band capacity ratio instead of the traditional volume-to-capacity ratio. This is because the arrivals are "controlled" and concentrated and, therefore, are able to use available green time relatively efficiently. Thus, unused green time, which implies a greater capacity or lower volume-to-capacity ratio, does not reduce delay. This finding contrasts with

TABLE 1 Delay Comparison: Case 1, Average Travel Time Delay per Vehicle[a] (sec/vehicle)

Volume (veh/cycle/lane)	Volume/Signal Capacity	Delay			Platoon Method			
		Webster's Method[a] (sec/veh)	May's Method[b] (sec/veh)	New 1985 HCM Method[c] (sec/veh)	Bandwidth Capacity (veh/cycle/lane)	Volume-to-Bandwidth Capacity Ratio	Bandwidth Capacity-to-Signal Capacity Ratio	Delay (sec/veh)
					3	1.0	0.25	0
					6	0.5	0.50	0
3	0.25	12.8	12.1	12.2	9	0.33	0.75	0
					12	0.25	1.00	0
					3	2.0	0.25	17.4
					6	1.0	0.50	0
6	0.50	15.3	13.4	13.8	9	0.67	0.75	0
					12	0.50	1.00	0
					3	3.0	0.25	23.2
					6	1.5	0.50	11.6
9	0.75	19.8	15.0	19.0	9	1.0	0.75	0
					12	0.75	1.00	0
					3	4.0	0.25	26.1
					6	2.0	0.50	17.4
12	1.00	630	17.6	51.0	9	1.33	0.75	8.7
					12	1.0	1.00	0

[a]Given: C = 60 sec, G = 29 sec, R = 31 sec, signal capacity = 12 veh/cycle/lane, base speed = 30 mph, H_D = 2.1 sec/veh; with platoon flow: first vehicle arrives unimpeded, H_A = 2.1 sec/veh.
[b]Adjusted to obtain average travel time delay per vehicle.
[c]Average conditions; adjusted to obtain average travel time delay per vehicle.

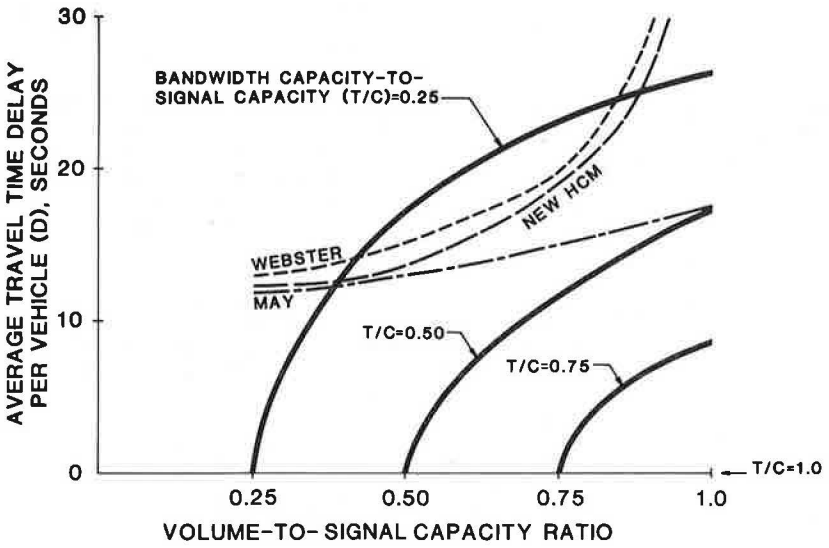


FIGURE 5 Delay comparison: Case 1 (refer to Table 1).

methods that employ random arrivals or uniform arrivals because these arrival patterns are more dispersed throughout the signal cycle. Therefore, random arrivals or uniform arrivals use the additional signal capacity to reduce delay.

Case 2: Comparison

Given in Table 2 and shown in Figure 6 is a comparison of the average travel time delay for impeded platoon arrivals with delays estimated for the other methods. Three different lengths of red time that the first vehicle arriving must wait (R_A = 10, 20, and 30 sec) and two different arrival headway situations (H_A = 3.0 and 2.1 sec) are considered. Other assumptions are the same as for Case 1.

As may be expected, delay is higher the longer the first vehicle of the platoon must wait through the red period. Also, the more the arrival headway exceeds the departure headway, the more the delay tends to reduce. When the arrival headway equals the departure headway, each vehicle is delayed the same amount of time.

SUMMARY AND CONCLUSIONS

Several findings have important traffic capacity and performance implications:

1. Platooning of traffic is desirable to minimize delay along arterial streets. However, the advantages of platooning may be lost if the leading vehicle is forced to stop because the following vehicles may be delayed as well.
2. The through-band of a standard time-space diagram is most meaningful when the first vehicle (or the "leading edge") of the platoon is unimpeded. This is because it describes the unimpeded flow of vehicles through a series of signals and assumes that the platoons travel at the progressive speed.

TABLE 2 Delay Comparison: Case 2, Average Travel Time Delay per Vehicle[a] (sec/vehicle)

Volume (veh/cycle/lane)	Volume/Signal Capacity	Delay			Platoon Method		
		Webster's Method[b] (sec/veh)	May's Method[b] (sec/veh)	New 1985 HCM Method[c] (sec/veh)	Red Time (R_A) 1st Vehicle Waits (sec)	Delay: Arrival Headway = 3.0 Sec	Delay: Arrival Headway = 2.1 Sec
3	0.25	12.8	12.1	12.2	10	15.0	15.9
					20	25.0	25.9
					30	35.0	35.9
6	0.50	15.3	13.4	13.8	10	13.7	15.9
					20	23.7	25.9
					30	33.7	35.9
9	0.75	19.8	15.0	19.0	10	12.3	15.9
					20	22.3	25.9
					30	32.3	35.9
12	1.00	630	17.6	51.0	10	11.0	15.9
					20	21.0	25.9
					30	31.0	35.9

[a]Given: C = 60 sec, G = 29 sec, R = 31 sec, signal capacity = 12 veh/cycle/lane, base speed = 30 mph, H_D = 2.1 sec/veh; with platoon flow; first vehicle arrives on red and is impeded.
[b]Adjusted to obtain average travel time delay per vehicle.
[c]Average conditions; adjusted to obtain average travel time delay per vehicle.

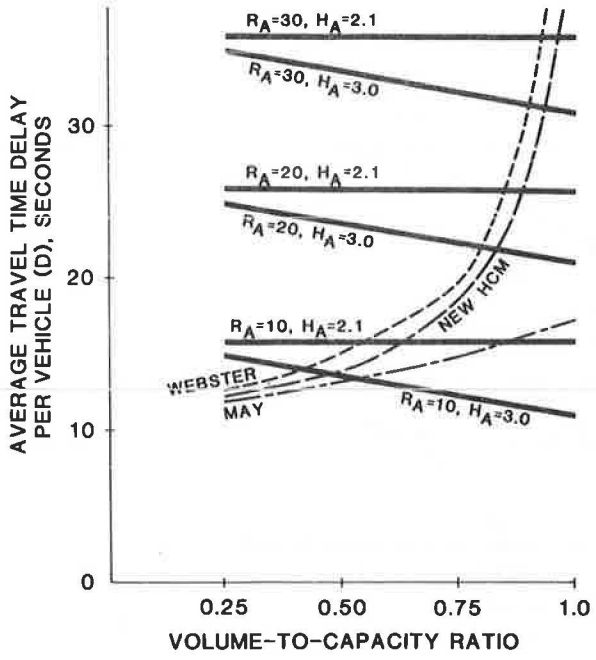


FIGURE 6 Delay comparison: Case 2 (refer to Table 2).

However, when the first vehicle in the platoon arrives during a red interval or encounters queued vehicles, the platoon experiences turbulence and the through-band representation becomes less valid.

3. Where the lead vehicle in the platoon is unimpeded, there is no delay when the capacity of the through-band equals or exceeds the approach volume. Thus, a high volume-to-capacity ratio may provide a high level of service. This contrasts with delays based on random or uniform arrivals, which are sensitive to the volume-to-capacity ratio.

4. However, where the first platoon vehicle is impeded by a red interval or by queue interference, a chain reaction may be introduced in which following vehicles are also impeded. This situation may create considerable delay and effectively reduce progression.

5. The concept of band capacity emerges as an important index of arterial street performance. For minimum delay conditions, volumes should not exceed the band capacity and the platoon leader should arrive at the intersection unimpeded.

6. Effective traffic signal coordination along arterial streets, therefore, can substantially reduce delay and improve levels of service.

The suggested methods provide a realistic means of estimating intersection performance in a progressive signal system. Accordingly, it may be desirable to reassess current delay formulations and computer simulation assumptions, especially where effective signal coordination exists.

Logical next steps include investigating techniques that account for nonplatoon traffic and turning movements, preparing additional delay tables for other signal timing assumptions, and conducting field tests to experimentally verify the research findings and to identify any needed adjustments.

Discussion

Edmond C. Chang*

This study analyzes the coordinated signal delay at signalized intersections mainly from the assumed platoon flow on the arterial street travel directions. A set of graphic analyses was used to investigate the signal delay with respect to the theoretical platoon arrival flow. This method suggested a theoretical approach to estimate the intersection performance between two intersections under its specified assumptions. Two major conditions studied are

- Platoon arrives on green or unimpeded flow (Case 1) or
- Platoon arrives on red or impeded flow (Case 2).

*Traffic Operations Program, Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843-3135

To simplify the study environment, several assumptions were made:

1. There is no random delay; only uniform delay is used in the study;
2. No platoon dispersion exists;
3. No queue spill-back passes over to another intersection or continues through another signal cycle;
4. All vehicles travel in the progression platoon;
5. All vehicles travel at the same free-flowing speed between intersections;
6. All vehicles travel with uniform headway;
7. Vehicular speeds, arriving from an upstream intersection, are the same as they are leaving the downstream intersection; and
8. Traffic operates under the undersaturated condition with arrival flow rate less than departure flow rate.

The most important measurement of effectiveness used in this study is travel time delay defined as the travel time consumed between the time required for a vehicle to recover the downstream approach speed as it approaches the upstream intersection and the time required to continue its unimpeded approach speed. Microscopic traffic characteristics, such as perception and reaction time, acceleration and deceleration rate, lost time and number of vehicle in queue, were considered explicitly by the modified Greenshields departure model. A deterministic delay estimation model was made separately for each successive stopped vehicle in the queue to estimate the total travel time delay under the given arrival rate, approach headway, progression bandwidth, and time offset to avoid deceleration of the first vehicle in the platoon.

The concept of band capacity was introduced in this paper as another important index for measuring arterial street performance. It was suggested that, for minimum delay operations, volumes should not exceed the band capacity and the platoon leader should arrive at the intersection unimpeded. A platoon delay calculation procedure depending on the volume-to-band capacity ratio instead of the traditional volume-to-capacity (V/C) ratio was developed. Most of the traffic is assumed to arrive in the "controlled" and "concentrated" progression band; therefore, vehicles are able to use available green time more efficiently in this study. Under this particular study assumption, the adjustment of progression bandwidth within the unused green time does not reduce delay. This unused green time may be a result of greater capacity or lower V/C ratio. On the other hand, methods that employ random arrivals or uniform arrivals indicate a totally different result because those arrival patterns are more dispersed throughout the signal cycle. Therefore, random arrivals or uniform arrivals can be guided to use the additional signal capacity provided by the slack green time to further reduce delay.

Effective traffic signal coordination can substantially reduce delay and improve levels of service. However, unsynchronized traffic signal operations will impede the progression band for carrying the through traffic movements. This inefficient progression operation will not only create undue signal delay but will also propagate these delays throughout the signalized network. Therefore, it is believed that the bandwidth and time offset as employed in Equation 2 to estimate the arterial travel time delay can heavily influence the calculations of through-band capacity. The method used to derive the bandwidth and offsets can significantly affect how efficiently the progression platoon can use the through green time for better signal coordinations.

In this study, the approach headway was assumed to be greater than or equal to the departure headway of the impeded vehicles. When the approach headway equals the departure headway and the first vehicle is impeded, each vehicle in the platoon is assumed to be delayed the same amount of time. When the approach headway is greater than the departure headway, each subsequent stopping vehicle will be delayed less than the first and, depending on the difference between the approach and departure headways, some vehicles may not be impeded as the vehicular queue clears the intersection.

In reality, because the first platoon vehicle is impeded by a red signal phase or by queue interference, a chain reaction phenomenon may be developed in which following vehicles are also impeded. This situation may create considerable "shock-wave" delay and effectively reduce progression especially onto the downstream intersections. [See Messer et al. (5) and papers by Chang et al. and Chang and Messer in this Record.] Therefore, it is suggested that the further revision of the assumption of "all platooned traffic" be enhanced to consider the random arrival flow rate onto the downstream intersection. It can be more helpful in estimating the through-bandwidth capacity and the resultant signal delay calculation. An approach, similar to the platoon interconnection factor, as used by PASSER II-84 to adjust for the difference in arrival rates between green and red phase of the cycle, is suggested for possible consideration in the further development and application of this study. Essentially, a version of the tentative NCHRP delay estimation equation was modified to adjust the arrival flow rate, especially the downstream through movements as affected by the effect of travel time on platoon dispersion. Techniques that account for nonplatoon traffic, platoon dispersion effects of the progression band beyond the downstream intersections, and additional delay tables for revised study assumptions are also recommended in order to provide more realistic applications of this research effort.

Authors' Closure

Edmond C. Chang sets forth an interesting overview of our paper and suggests some possible directions for further study. We agree that it is important to accurately identify the width of the real or effective through-band, account for random perturbations in platoons, and consider the consequences of turn-in or turn-off traffic. Certainly, such additional analyses can produce more realistic results. Even more important, however, are actual field studies that analyze delays under conditions of optimal progression. Analyses of delays along one-way streets with near-perfect progression or arterial streets with preferential offsets would prove useful for comparing our "boundary" formulations with actual observations. Such real-world analyses will permit our delay tables to be revised for practical application in determining levels of service at intersections.

REFERENCES

1. W.R. Reilly, C.C. Gardner, and J.H. Kell. A Technique for Measurement of Delay at Intersec-

tions. JHK and Associates; FHWA, U.S. Department of Transportation, San Francisco, Calif., 1976.
2. M.J. Huber. Traffic Flow Theory. In Transportation and Traffic Engineering Handbook, W.S. Homburger, Ed., 2nd ed., Institute of Transportation Engineers; Prentice-Hall, Englewood Cliffs, N.J., 1982, pp. 465-467.
3. Transportation Training and Research Center, Polytechnic Institute of New York, and Texas A&M Research Foundation, Texas A&M University System. Signalized Intersections. Draft 2, Chapter 9, NCHRP Project 3-28(B) to update the 1965 Highway Capacity Manual. TRB, National Research Council, Washington, D.C., Sept. 1984.
4. L.J. Pignataro. Signalization of Isolated Intersections. In Traffic Engineering: Theory and Practice. Prentice-Hall, Englewood Cliffs, N.J., 1973.
5. C.J. Messer, D.B. Fambro, and D.A. Anderson. Effects of Design on Operational Performance of Signal Systems. Research Report 203-2F. Texas Transportation Institute, Texas A&M University, College Station, Aug. 1975.

APPENDIX A--REACTION AND ACCELERATION LOSS (L)

Refer to Figure A-1. Assumptions made are:

1. Time after start of green that first stopped vehicle crosses intersection curb line = t_g = 3.8 seconds (4),
2. Reaction time = t_r = 1.0 sec,
3. Distance from stop bar to intersection curb line = d_s = 15 ft, and
4. Acceleration rate = a = 3.3 mph/sec (A. French. Vehicle Operating Characteristics. In Transportation and Traffic Engineering Handbook, W.S. Homburger, Ed., 2nd ed. Institute of Transportation Engineers; Prentice-Hall, Englewood Cliffs, N.J., 1982, p. 168).

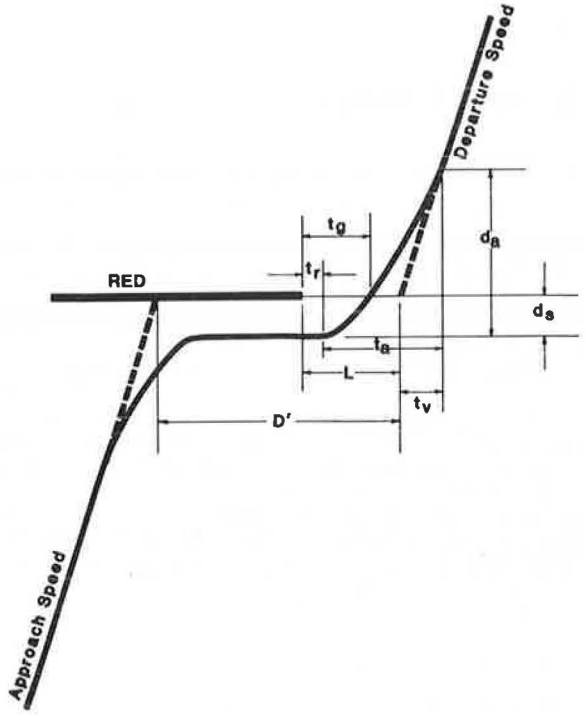


FIGURE A-1 Illustration of terms used to determine L.

General equations are as follows:

Time during acceleration (t_a) is

$$t_a = (v - v_o)/a \quad \text{(A-1)}$$

where

v = velocity after speed change;
v_o = initial velocity (e.g., zero velocity at stop); and
a = acceleration rate.

Distance traveled during acceleration (d_a) is

$$d_a = v_o\ (t) + 0.5\ (a)\ t_a^2 \quad \text{(A-2)}$$

Time traveled at base speed (t_v) is

$$t_v = (d_a - d_s)/v \quad \text{(A-3)}$$

Reaction time and acceleration time loss (L) is

$$L = t_r + t_a - t_v \quad \text{(A-4)}$$

Values of L are as follows:

Speed (mph)	L (sec)
20	4.5
25	5.2
30	5.9
35	6.6
40	7.3

APPENDIX B--TIME OFFSET (t_d) TO AVOID DECELERATION OF FIRST PLATOON VEHICLE

Refer to Figure B-1. Assumptions made are:

1. There are no queued vehicles at the signal approach. (Note: If queued vehicles are present, then $t_d = L + H_D(S)$, where L, H_D, and S are as defined in the text.)
2. Deceleration rate = a and
a = -4.6 mph/sec for velocity changes ranging from 0 to 30 mph and
a = -3.3 mph/sec for velocity changes ranging from 30 to 40 mph (A. French. Vehicle Operating Characteristics. In Transportation and Traffic Engineering Handbook, W.S. Homburger, Ed., 2nd ed. Institute of Transportation Engineers; Prentice-Hall, Englewood Cliffs, N.J., 1982, p. 168).
3. Distance from front of stopped vehicle to curb of intersection = d_s = 15 feet.

General equation are as follows:

Time during deceleration (t) is

$$t = (v - v_o)/a \quad \text{(B-1)}$$

where

v = velocity after speed change (e.g., zero velocity at stop),
v_o = approach velocity, and
a = deceleration rate.

Distance traveled during deceleration (d_d) is

$$d_d = v_o\ (t) + 0.5\ (a)\ t^2 \quad \text{(B-2)}$$

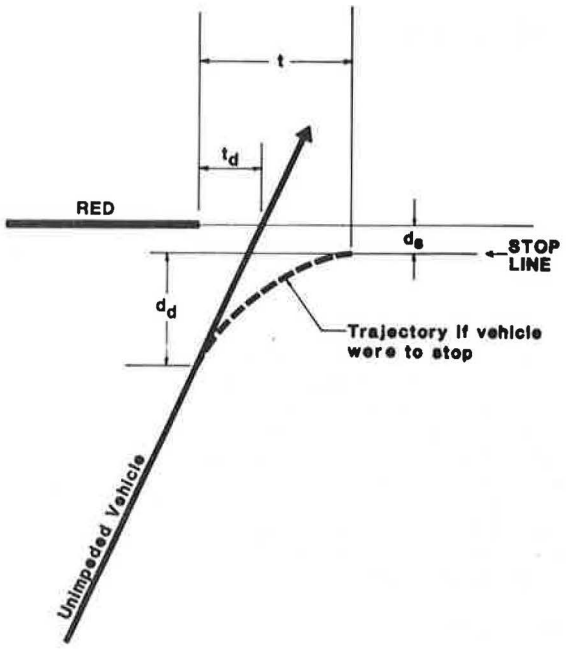


FIGURE B-1 Illustration of terms used to determine t_d.

Time offset to avoid deceleration of first platoon vehicle (t_d) is

$$t_d = (d_d + d_s)/v_o \quad \text{(B-3)}$$

Values of t_d are as follows:

Speed (mph)	t_d (sec)
20	2.7
25	3.1
30	3.6
35	4.1
40	6.1

APPENDIX C--EXAMPLE OF PLATOON ARRIVALS WHEN FIRST VEHICLE ARRIVES DURING GREEN INTERVAL

Given: C = 60 sec, R = 31 sec, G = 29 sec, H_A = 3 sec, H_D = 2.1 sec, V = 9 vehicles/lane/cycle, base speed = 30 mph, W = 19 sec, first vehicle is not impeded, and no queued vehicles at upstream signal.

Find average travel time delay per vehicle (D):

$$T = (W - t_d + H_A)/H_A$$
$$T = (19 - 3.6 + 3)/3 = 6 \text{ vehicles/lane/cycle} \quad \text{(C-1)}$$

$$S = V - T$$
$$S = 9 - 6 = 3 \text{ vehicles/lane/cycle} \quad \text{(C-2)}$$

$$R_A = R - H_A$$
$$R_A = 31 - 3 = 28 \text{ sec} \quad \text{(C-3)}$$

$$D' = R_A + L$$
$$D' = 28 + 5.9 = 33.9 \text{ sec} \quad \text{(C-4)}$$

$$D = [S(D') + F(H_D - H_A)]/V$$
$$S = 3, F = 3$$
$$D = [3(33.9) + 3(2.1 - 3)]/9 = 11 \text{ sec/vehicle} \quad \text{(C-5)}$$

APPENDIX D--EXAMPLE OF PLATOON ARRIVALS WHEN FIRST VEHICLE ARRIVES DURING RED INTERVAL

Given: C = 60 sec, R = 31 sec, G = 29 sec, H_A = 3.0 sec, H_D = 2.1 sec, V = 9 vehicles/lane/cycle, base speed = 30 mph, and first vehicle arrives 10 sec before start of green (therefore, R_A = 10 sec).

Find average travel time delay per vehicle (D):

$$S = (R_A + L)/(H_A - H_D)$$
$$S = (10 + 5.9)/(3 - 2.1) = 18 \quad \text{(D-1)}$$

S cannot be greater than V; therefore, use S = 9

$$D' = R_A + L$$
$$D' = 10 + 5.9 = 15.9 \text{ sec} \quad \text{(D-2)}$$

$$D = [S(D') + F(H_D - H_A)]/V$$
$$D = [9(15.9) + 36\ (2.1 - 3.0)]/9$$
$$= 12.3 \text{ sec/vehicle} \quad \text{(D-3)}$$

APPENDIX E--WEBSTER'S DELAY EQUATION

$$A = c(1 - \lambda)^2/[2(1 - \lambda x)] + x^2/[2q(1 - x)] - [0.65(c/q^2)^{1/3}]\ [x^{(2+5\lambda)}]$$

where

A = average delay to passenger car unit (pcu) on the approach (sec);
c = cycle time (sec);
g = effective green time (sec);
r = effective red time (sec);
s = saturation flow on the approach (pcu/sec);
λ = g/c, proportion of the cycle that is effectively green;
y = q/s, ratio of average arrival rate to saturation flow; and
x = qc/gs, ratio of average number of arrivals per cycle to the maximum number of departures per cycle.

APPENDIX F--MAY'S UNIFORM DELAY EQUATION

$$A = r^2/[2c(1 - q/s)]$$

where

A = average delay to pcu on approach (sec),
r = effective red time (sec),
c = cycle length (sec),
q = average arrival rate of traffic on the approach (pcu/sec), and
s = saturation flow on the approach (pcu/sec).

APPENDIX G--NEW 1985 HIGHWAY CAPACITY MANUAL EQUATION (AVERAGE CONDITIONS)

$$d = 0.38(C)\ [1 - (g/C)]^2/[1 - g/C)(x)] + 173\ x^2\ \{(x - 1) + [(x - 1)^2 + (16\ x^2/v_a)]^{1/2}\}$$

where

d = average stopped delay per vehicle (sec/vehicle),
C = cycle length (sec),
g/C = ratio of effective green time to cycle length,
x = volume-to-capacity ratio, and
v_a = adjusted volume (vehicles/hour/lane).

Publication of this paper sponsored by Committee on Highway Capacity and Quality of Service.

Analysis of Light Rail Vehicle Clearance Time at Intersections

WULF GROTE and JASON C. YU

ABSTRACT

The intent of this study is to determine the amount of intersection disruption created by the implementation of light rail transit (LRT) through intersections with automobile traffic. The measure for determining intersection disruption is the LRT clearance time, which is the sum of time consumed by the actual presence of an LRT vehicle in the intersection and the time required before and after LRT arrival to prepare the intersection. Clearance time is affected by several factors that have been categorized as LRT vehicle operating characteristics, geometric layout of the intersection, and the traffic control method implemented. Several equations have been developed, with these factors as variables, to determine the LRT clearance time created by various operating conditions. In addition, several graphs show the ranges of clearance time that might be experienced under specific operating characteristics.

Congestion of major roadways is becoming a problem in many cities across the North American continent. In an effort to alleviate the congestion problem, several medium-sized cities (roughly 1 to 2 million inhabitants) have turned to light rail transit (LRT) to provide more efficient usage of right-of-way within major transportation corridors. LRT is a fixed-guideway transit system that has the capability of operating safely at grade, but LRT can also be grade separated at major conflict points. This means that, in many cities, LRT will operate at street level through roadway intersections in an effort to reduce capital expenditures for the transit project.

Intersections where LRT conflicts with automobile traffic must be carefully analyzed to determine the potential delay impacts that may be created. The severity of impact will depend on the specific characteristics of the LRT system being implemented. Several previous studies have identified factors that influence intersection performance. One such study by Larwin and Rosenberg (1) identified delay at an LRT crossing as a function of

- LRT approach speed,
- Train length,
- Location of stop,
- Emergency stopping capabilities of LRT,
- Service frequency,
- Cross-street width, and
- Train detection and signal control requirements.

Although many influencing factors have been identified through previous studies, little work has been done to determine the amount of impact created by each one. The objective of this study is to demonstrate the amount of intersection disruption created by each factor. In addition, the results of this study should serve as a valuable tool for engineers and planners in selecting appropriate LRT system characteristics to minimize intersection delay. Factors affecting the delay experienced at an intersection can be divided into three categories: light rail vehicle operating characteristics, prevailing intersection geometrics, and the traffic control method implemented. Factors within each of these categories are

Light Rail Operating Characteristics	Geometric Layout	Traffic Control
Speed	Width of crossing	Degree of light rail priority
Acceleration	Turning-lane provisions	Traffic control devices
Proximity of LRT station		Effects of emergency stop considerations
Headway		Signal phasing
Train length		Automobile progression

Some of these factors have overlapping impacts and others are difficult to quantify. In this study, each of these factors was analyzed and, if possible, quantified in detail. These factors together determine the amount of time that normal traffic flow through an intersection is interrupted by the presence of a light rail vehicle. This time is defined here as the LRT clearance time. Some of this time is consumed by the physical presence of a light rail vehicle in the intersection; the remaining time is required to prepare the intersection for a rail vehicle arrival and allow necessary steps to resume automobile traffic flow after the rail vehicle has cleared the intersection. This study will use LRT clearance time as the measure for determining intersection disruption.

It is important to note that not all LRT clearance time necessarily results in lost capacity to an intersection. In some cases, traffic will continue to flow on intersection legs not conflicting with the LRT crossing. Even at locations where all traffic will be stopped during the physical presence of LRT in the intersection, vehicles may still be able to flow during a portion of the time the intersection is being prepared for LRT arrival. The intent of this analysis is to show the amount of time that normal intersection operations are disrupted by the implementation of LRT. In many cases, the study results will not necessarily show the lost intersec-

tion capacity created by LRT. Intersection capacity losses may result from a portion or all of the LRT clearance time, depending on the specific operating characteristics of the intersection.

ALTERNATIVE OPERATIONAL CONDITIONS

The first step of the analysis of LRT clearance time is to determine what types of crossing conditions a light rail vehicle might encounter on arrival at an intersection. In general, there are six possible operational conditions:

1. The light rail vehicle approaches the intersection at a constant speed and, because of either preemption of the traffic signal or light rail progression, the light rail vehicle is able to proceed without interruption.
2. Again, the light rail vehicle is able to proceed at a constant speed. However, as a safety precaution, the time for a light rail vehicle to make a full emergency stop is accommodated after cross-street traffic has been halted and before LRT arrival at the intersection. The emergency stop time provision assures that the LRT vehicle operator can stop the train if an intersection blockage occurs. Without this provision, an LRT vehicle will arrive at the intersection the instant that cross-street traffic receives a stop condition.
3. The light rail vehicle is required to stop at the near side of an intersection and then must accelerate back to its operational speed. This condition will occur if a near-side station platform is present or if the light rail vehicle is stopped by the traffic signal (i.e., no LRT priority).
4. The light rail vehicle is able to cross the intersection uninterrupted, but deceleration occurs due to a station platform at the far side of the intersection. No emergency stop considerations are provided. For this analysis, the deceleration condition has been combined with the acceleration case (Condition 3) because acceleration and deceleration rates are assumed to be similar.
5. Again, the light rail vehicle is able to cross the intersection without interruption and then decelerates into a far-side station. However, it is assumed that, as a safety precaution, a full emergency stop from the operational speed is accounted for before a light rail vehicle arrives at the crossing.
6. No priority is granted to the light rail vehicle and, as a result, the vehicle is forced to stop at the near side of an intersection where far-side station platforms are present. This means that the light rail vehicle, after stopping, will have to accelerate as close as possible to the operational speed and then decelerate into the station.

For each of these six conditions, equations were derived for estimating LRT clearance time. Critical factors entering into the equations were light rail speed; train length; width of crossings; and, in some cases, acceleration, deceleration, and emergency braking. The equations derived in this study are expressed by general variables. However, examples and graphs presented herein are calculated by inserting typical values for LRT vehicle length, service acceleration rate, service deceleration rate, and emergency deceleration rate. Although the value of these parameters may vary from one vehicle model to the next, the analysis presented here makes use of values considered typical of new light rail systems. The values used are as follows:

- Vehicle length = 90 ft,
- Service acceleration = 4 ft/sec^2,
- Service deceleration = 4 ft/sec^2, and
- Emergency deceleration rate = 7.3 ft/sec^2.

DERIVATION OF LRT INTERSECTION CLEARANCE TIME

Condition 1

The condition where vehicle speed is constant is best described using the uniform rectilinear motion equation:

$$t = (x - x_o)/s$$

where

x = final position coordinate,
x_o = initial position coordinate,
s = operating speed (a constant), and
t = time to cover the distance from x_o to x.

In the case where LRT crosses an intersection, $(x - x_o)$ can be simplified to x to denote the total distance covered in crossing the intersection. This distance (x), referred to as the effective crossing distance, can further be defined as the intersection width (w) plus the length of the train (L). The train length is determined by multiplying the vehicle length (c) by the number of vehicles (V) forming the train. Therefore, the LRT clearance time can be determined by

$$t = (w + cV)/s \quad (1)$$

where w and c are expressed in feet, s is expressed in feet per second, and t is in seconds.

Equation 1 has been solved for speeds ranging between 10 and 50 mph, intersection widths between 40 and 160 ft, and train lengths of one-, three-, and five-car trains, as shown in Figure 1. Not all possible values for the various factors have been computed, but values not shown can be interpolated from the lines on the graph or by using Equation 1. This graph should give a good indication of how modifications would affect light rail clearance time.

The obvious conclusion that can be drawn from Figure 1 is that as speed increases the total light rail clearance time decreases. At low speeds, particularly when train lengths are long and intersections are wide, a small change in speed results in a significant reduction in clearance time. However, with high speeds, short trains, and narrow intersections, the crossing time is hardly affected at all. For example, a five-car train traveling across a 160-ft intersection at 12 mph would save about 7 sec over a 10-mph speed under the same conditions. Conversely, a one-car train crossing a 40-ft intersection at a speed increased from 40 to 50 mph would save less than half a second. Other conclusions that become apparent by studying Figure 1 are that at slow speeds a change in train length or a change in intersection width can result in a fairly significant change in light rail clearance time. At high speeds, however, changes of this nature have a relatively minor impact.

Condition 2

The clearance time equation developed for this condition is similar to that for Condition 1, except that emergency stop considerations must be added. The suggested emergency stop time (E) is

$$E = (s/e) + 4 \text{ sec}$$

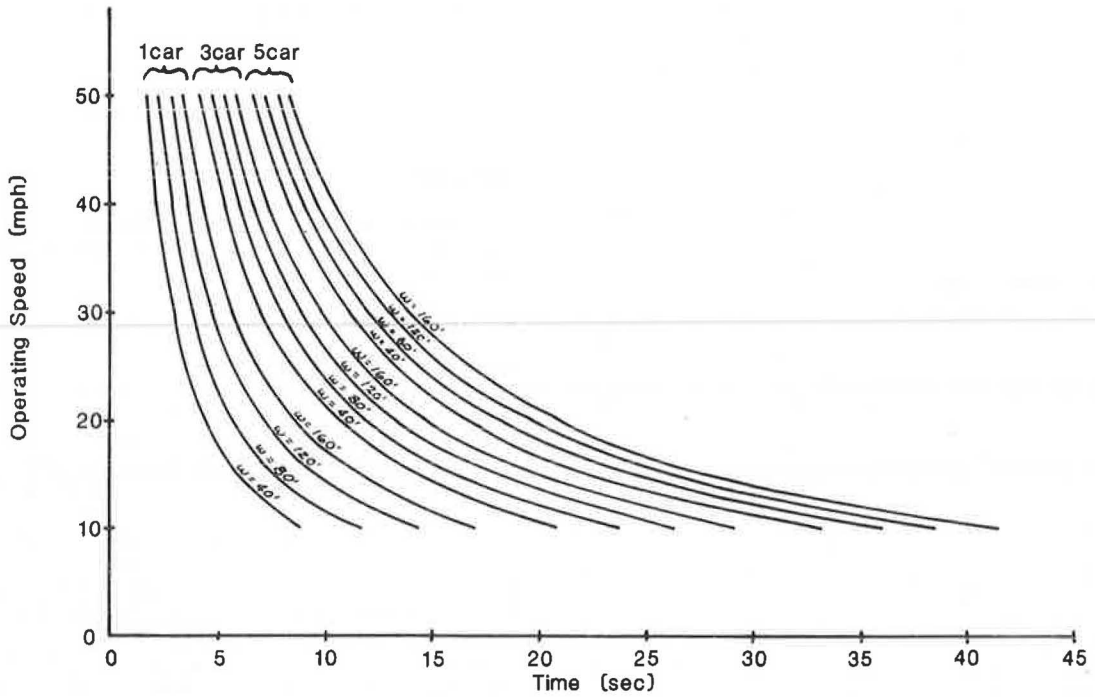


FIGURE 1 Light rail transit clearance time—constant speed.

where E is in seconds, e is the emergency braking rate in feet per second^2, and 4 sec is a factor added to allow driver reaction time and jerk considerations. This equation is added to the clearance time provided by Equation 1 and results in an overall light rail clearance time for this condition as follows:

$$t = [(w + cV)/s] + (s/e) + 4 \quad (2)$$

The emergency stop time is smallest at low speeds and becomes significantly larger as speeds increase.

Again, as was done for Condition 1, Equation 2 was solved for a range of values for each factor as shown in Figure 2. Unlike the graph for Condition 1, there is an optimum speed (s_{opt}) for each of the curves where light rail crossing time can be minimized; s_{opt} is derived by taking the derivative of Equation 2 and setting it equal to zero:

$$s_{opt} = [e(w + cV)]^{1/2} \quad (3)$$

For a given train length and intersection width, it is then possible to determine the corresponding

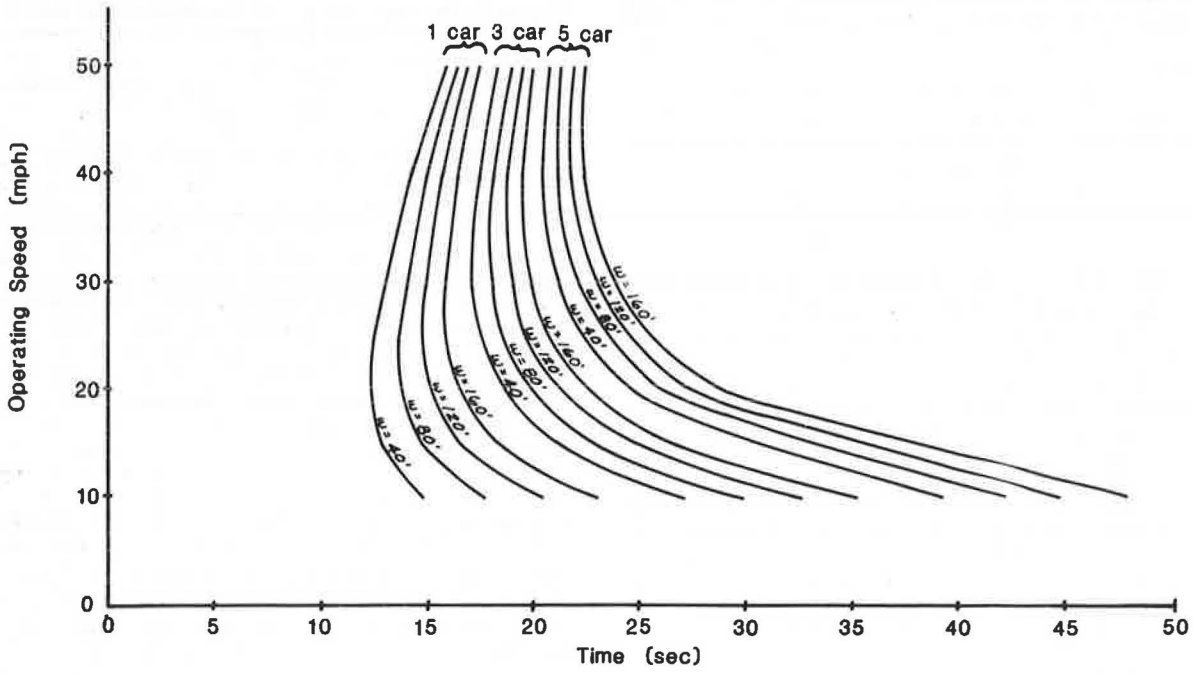


FIGURE 2 Light rail transit clearance time—constant speed including emergency stop.

s_{opt} by using Equation 3. This speed is then inserted into Equation 2 to determine the minimum clearance time.

It can be concluded from Figure 2 that optimum speed increases with increasing train length and intersection width. In addition, as optimum speed is increased, clearance time also becomes longer. Therefore, the least amount of intersection disruption would be at moderate speeds with short train lengths and narrow intersections. At low speeds the intersection clearance time is increased due to the amount of time the light rail vehicle is blocking the intersection, and at high speeds the clearance time is again increased due to the increased emergency stop time required before vehicle arrival at the intersection. Figure 2 also shows that changes in train length or intersection width create more of an impact at slower speeds. Also, speed is more of a delay factor at lower speeds when intersections are wide and trains are long.

Conditions 3 and 4

As mentioned earlier, the cases of acceleration from a near-side stop and deceleration into a far-side stop have been combined in this analysis because a common rate of acceleration and deceleration is used. Therefore, the equations will be identical, but signs will be reversed (negative rate for deceleration, positive rate for acceleration). The discussion that follows will refer to acceleration, but deceleration can be calculated by changing signs.

There are two circumstances that can occur for the conditions depending on the value of the factors involved. The first possibility is for a light rail vehicle to reach operating speed before clearing the intersection. Under this circumstance, a portion of the intersection would be traversed with the vehicle accelerating and the remainder of the intersection would be crossed at a constant speed. The second possibility is for the light rail vehicle to accelerate through the entire intersection without reaching full operating speed. Separate equations have been developed for each of these conditions.

The first step is to determine if the light rail vehicle would be able to cross a portion of the intersection at full operating speed. This is done by subtracting the distance to accelerate to operating speed from the total width of the intersection plus the train length. The distance to accelerate is found by using the uniformly accelerated rectilinear motion equation:

$$x_a = \left(s_f^2 - s_o^2\right)/2a$$

where

x_a = acceleration distance,
s_f = final speed,
s_o = initial speed, and
a = acceleration.

Because either the initial or the final speed will be zero, depending on whether the vehicle is accelerating or decelerating, this equation can be modified with the value for operating speed (s^2) replacing $s_f^2 - s_o^2$. Thus, the equation is further simplified to

$$x_a = s^2/2a$$

where x_a is in feet, s is in feet per second, and a is in feet per second^2. The total effective crossing width is w + cV as defined for Condition 1. Therefore, the distance across an intersection that a light rail vehicle is able to travel at full speed (R) is

$$R = (w + cV) - (s^2/2a) \quad (4)$$

If the value of R is greater than or equal to zero, the vehicle crosses at full speed for this distance and if R is less than zero, the vehicle is unable to attain operating speed in the distance covered by the effective crossing width.

For $R \geq 0$, the total clearance time is calculated by adding the time to accelerate plus the time the vehicle operates at a constant speed through the intersection. The acceleration time (t_a) is calculated by using the uniformly accelerated rectilinear motion equation:

$$s_f = s_o + at_a$$

This simplifies to

$$t_a = s/a$$

because either the final or the initial speed equals zero. The time at constant speed (t_c) is found by using the uniform rectilinear motion equation defined under Condition 1:

$$t_c = (x - x_o)/s$$

In this case $x - x_o$ is replaced by R and

$$t_c = (R/s) = [(w + cV) - (s^2/2a)]/s$$

The clearance time (t) becomes

$$t = t_a + t_c = (s/2a) + [(w + cV)/s] \quad (5)$$

For $R < 0$, the maximum velocity (s_m) attained by the light rail vehicle before clearing the intersection is again calculated by making use of a uniformly accelerated rectilinear motion equation simplified to

$$s_m = [2a(w + cV)]^{1/2}$$

The clearance time (t) is also found by using a uniformly accelerated rectilinear motion equation where

$$t = s_m/a = [2a(w + cV)]^{1/2}/a \quad (6)$$

Using Equations 5 and 6, the clearance times for various factor values have been calculated as shown in Figure 3.

Figure 3 shows that, for higher operating speeds, the full speed is usually not attainable, particularly for shorter trains and narrow intersections. This means that the operating speed under these conditions is irrelevant to the clearance time. However, when slower operating speeds are used, these speeds have a significant impact on clearance time, especially as train length and intersection width increase.

It should be noted that, for the case of deceleration, the clearance times calculated correspond to a far-side stop that occurs as soon as the vehicle is clear of the intersection. If, for example, the platform is longer than the train, the clearance time would be slightly decreased if the train did not stop until it reached the end of the platform farthest away from the intersection. The reason the clearance time would be decreased for this configuration is that the light rail train would be able to continue at operating speed for longer than the times calculated in Figure 3.

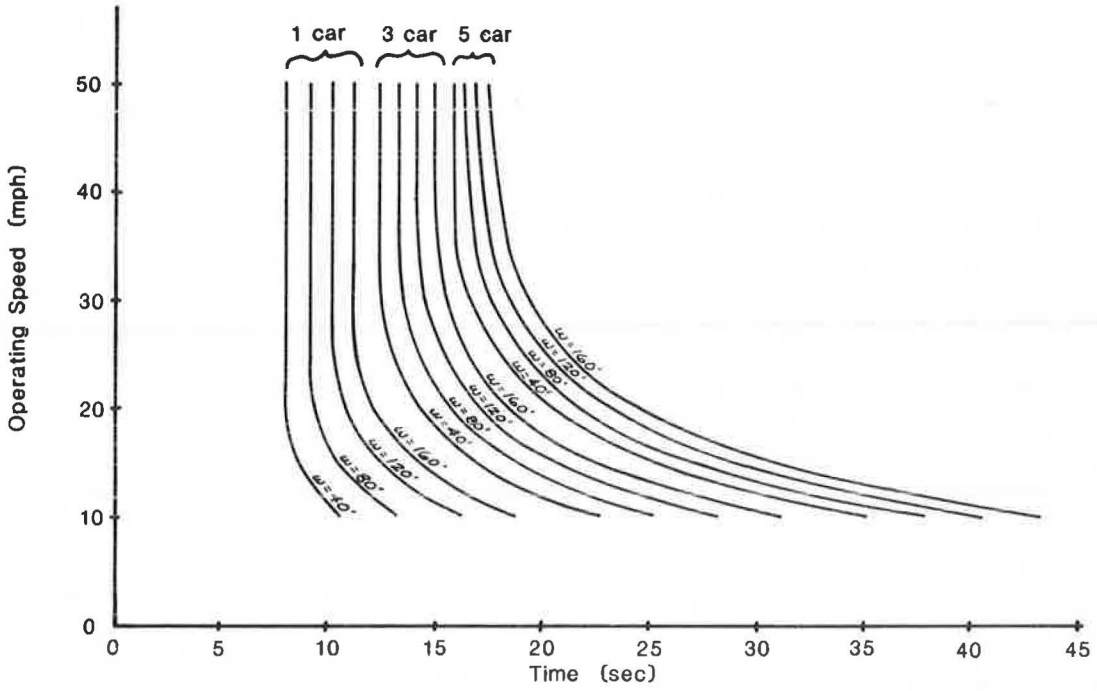


FIGURE 3 Light rail transit clearance time—deceleration to far-side stop or acceleration from near-side stop.

Condition 5

The condition of a far-side stop with emergency stopping considerations is similar to Condition 4 except that time for an emergency stop must be added to the total LRT clearance. The emergency stop factor is identical to that explained for Condition 2. Therefore, the equations for Condition 4 are modified as follows:

$$t = (s/2a) + [(w + cV)/s] + (s/e) + 4 \quad \text{for } R \geq 0 \quad (7)$$

$$t = \{[2a(w + cV)]^{1/2}/a\} + (s/e) + 4 \quad \text{for } R < 0 \quad (8)$$

Figure 4 shows the clearance times for various train lengths, intersection widths, and speeds. For any given train length and street width, there is an optimal operating speed at which crossing time is minimized. Speeds below this value create larger clearance times due to vehicle blockage of the intersection, and speeds above the optimum have greater clearance times due to consideration of emergency stop time. The optimum speed for given

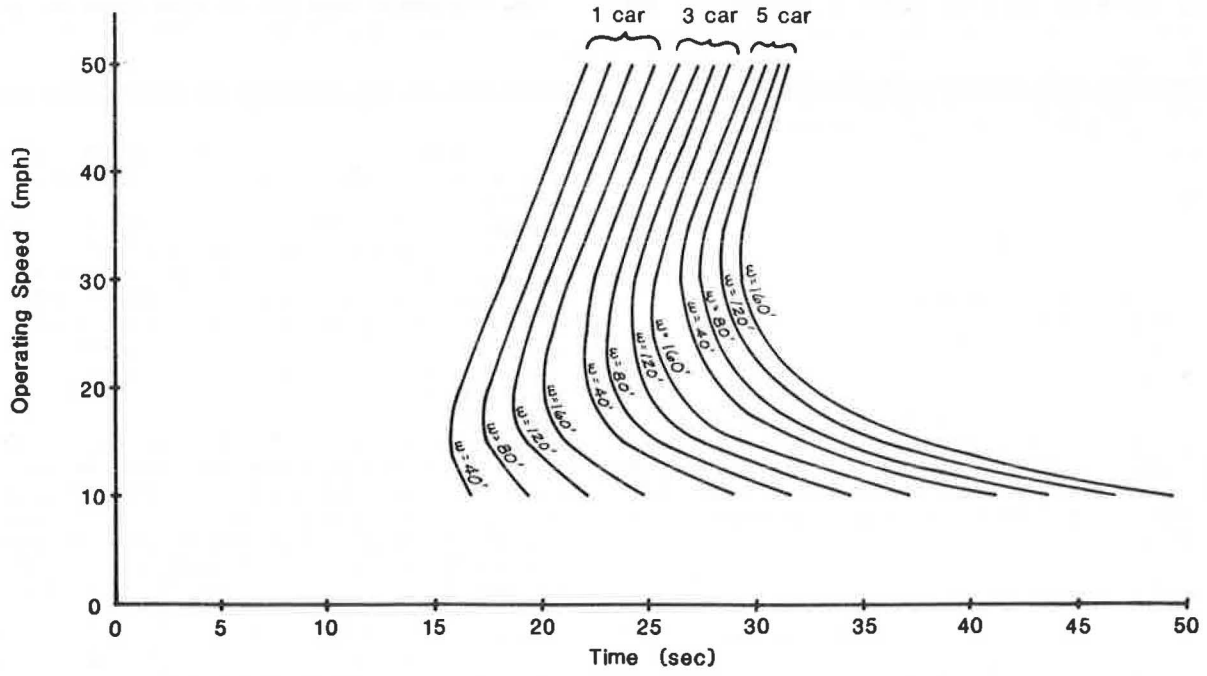


FIGURE 4 Light rail transit clearance time—deceleration to far-side stop including emergency stop.

conditions is calculated by taking the derivative of Equation 7 and setting this equal to zero to obtain:

$$s_{opt} = \{(w + cV)/[(1/2a) + (1/e)]\}^{1/2} \quad (9)$$

Equation 8 is not used for optimum speed calculations because the optimum speed is always attained before R becomes less than zero. As R becomes more negative, clearance time increases in a linear fashion.

Optimum speed increases with increasing train length and intersection width. At slow speeds, modifications in intersection width or train length create more noticeable time changes than they do at high speeds. Also, for slow speeds when train length and intersection width are large, a small change in speed can create a significant amount of change in clearance time.

Condition 6

The case where a light rail vehicle must first stop at the near side of an intersection and then stop again at the far side assumes that the vehicle will attempt to attain full operating speed, or the highest speed possible, before decelerating to a far-side stop. If the operating speed is attainable, the light rail vehicle will operate at this constant speed until it is necessary to begin deceleration from the service rate.

As for previously described conditions, the uniformly accelerated rectilinear motion equation is best used to determine light rail clearance time:

$$s_f^2 = s_o^2 + 2a(x - x_o)$$

This is similar to Condition 3 in that the value of R must first be calculated. For Condition 6 this is done by subtracting the distance required to accelerate to operating speed and then to decelerate back to zero (x_{ad}) from the effective crossing distance (w + cV). Because x_{ad} would be twice the acceleration distance (x_a) defined for Condition 3, x_{ad} can be expressed as

$$x_{ad} = 2x_a = 2(s^2/2a) = s^2/a$$

Therefore, the distance across the intersection that the light rail vehicle is able to travel at operating speed is

$$R = (w + cV) - (s^2/a) \quad (10)$$

If $R < 0$, the operating speed is unattainable for the given factor values.

For $R \geq 0$, the light rail clearance time is calculated by adding the time required for accelerating and then decelerating the vehicle (t_{ad}). The time t_{ad} is equal to twice the time to accelerate (t_a), assuming the service acceleration and deceleration rates are the same, and can be expressed as

$$t_{ad} = 2t_a = (2s/a)$$

The time at operating speed (t_c) is calculated by

$$t_c = R/s = \{[(w + cV) - (s^2/a)]/s\}$$

The clearance time (t) becomes

$$t = t_{ad} + t_c = (s/a) + [(w + cV)/s] \quad (11)$$

For $R < 0$, the maximum attainable speed (s_m) must be calculated using the uniformly accelerated rectilinear motion equation:

$$s_f^2 = s_o^2 + 2a(x - x_o)$$

In this case the vehicle can accelerate for only one-half the distance across the intersection before decelerating as shown in Figure 5. Therefore, $(x - x_o)$, by making substitutions, becomes

$$(x - x_o) = [(w + cV)/2]$$

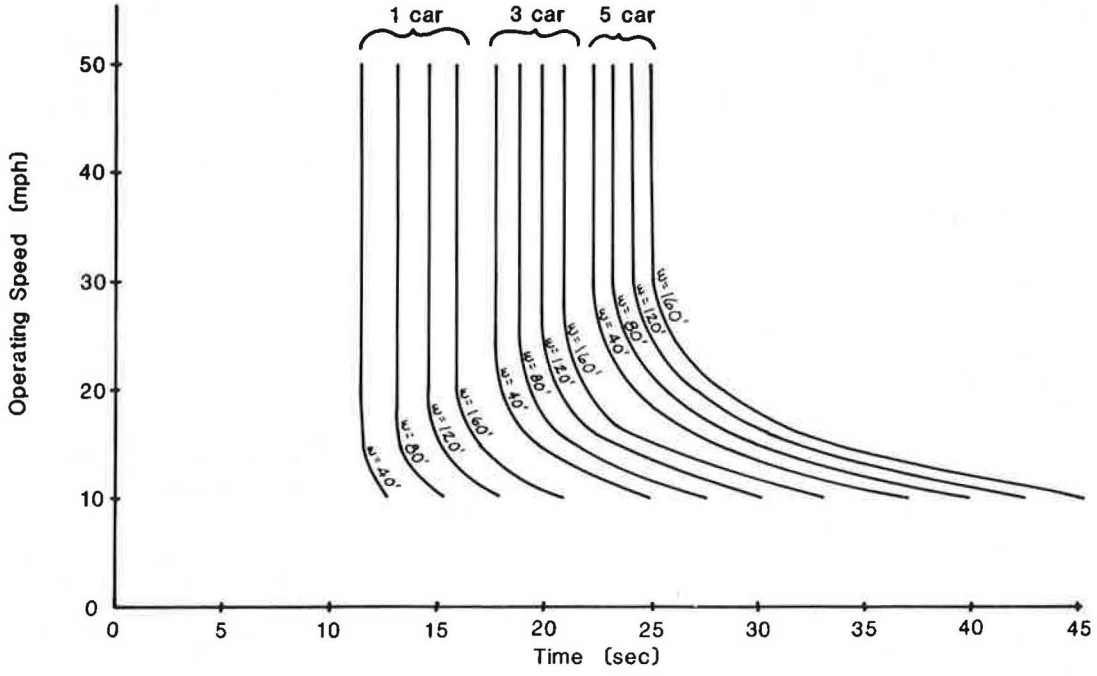


FIGURE 5 Light rail transit clearance time—acceleration from stop and then deceleration to a stop at far side of intersection.

TABLE 1 Summary of Equations

Description	Equation	Equation No.
Constant speed (Condition 1)	$t = (w + cV)/s$	(1)
Constant speed with emergency stop (Condition 2)	$t = [(w + cV)/s] + (s/e) + 4$	(2)
Acceleration or deceleration (Conditions 3, 4)	$R = (w + cV) - (s^2/2a)$	(4)
	for $R \geq 0$: $t = (s/2a) + [(w + cV)/s]$	(5)
	for $R < 0$: $t = [2a(w + cV)]^{1/2}/a$	(6)
Deceleration with emergency stop (Condition 5)	$R = (w + cV) - (s^2/2a)$	(4)
	for $R \geq 0$: $t = (s/2a) + [(w + cV)/s] + (s/e) + 4$	(7)
	for $R < 0$: $t = \{ [2a(w + cV)]^{1/2}/a \} + (s/e) + 4$	(8)
Acceleration then deceleration (Condition 6)	$R = (w + cV) - (s^2/a)$	(10)
	for $R \geq 0$: $t = (s/a) + [(w + cV)/s]$	(11)
	for $R < 0$: $t = 2[a(w + cV)]^{1/2}/a$	(12)

Substituting this and the speed components, the equation becomes

$$s_m^2 = 2a[(w + cV)/2] = a(w + cV)$$

$$s_m = [a(w + cV)]^{1/2}$$

The clearance time is then found by adding the time to accelerate (t_a) to the time to decelerate. Because acceleration and deceleration rates are equal for this analysis, $2t_a$ can be used, which becomes

$$t = 2t_a$$

$$t = 2(s_m/a) = 2[a(w + cV)]^{1/2}/a \quad (12)$$

Figure 5 shows the clearance times for varying street widths, operating speeds, and train lengths. The values calculated here, again, assume that the far-side stop occurs the instant the light rail vehicle is clear of the intersection (as explained for Condition 4).

Figure 5 shows that most higher operating speeds are unattainable regardless of the effective crossing distance and that the operating speed is only a factor in clearance time at slower speeds. Also, at slower speeds, the impact on clearance time becomes greater as speed decreases and train length and intersection width increase.

COMPARISON OF OPERATIONAL CONDITIONS

A summary of the equations that are critical in calculating clearance times for each of the six described conditions is given in Table 1.

For Conditions 2 and 5, in which emergency stop considerations are included, an optimum speed exists that minimizes overall clearance time. The optimum speed equations are given in Table 2.

The variables for all the equations given in Tables 1 and 2 are summarized as

t = total clearance time (sec),
w = intersection width (ft),
c = LRT vehicle length (ft),
V = number of vehicles in train,
s = operating speed (ft/sec),
e = emergency braking rate (ft/sec^2),
R = distance light rail vehicle is able to travel across an intersection at operating speed (ft),
a = acceleration (or deceleration) rate (ft/sec^2), and

TABLE 2 Optimum Speed Equations

Condition	Equation	Equation No.
2	$s_{opt} = [e(w + cV)]^{1/2}$	(3)
5	$s_{opt} = \{(w + cV)/[(1/2a) + (1/e)]\}^{1/2}$	(9)

s_{opt} = optimum operating speed across intersection (ft/sec).

Figure 6 shows selected curves for each of the previously discussed operating conditions. Only the curves for selected train lengths (one and five cars) and intersection widths (40 and 160 ft) have been included. This should give a feel for the range of impacts on clearance time without cluttering and confusing the graph.

The curves in Figure 6 show that constant light rail speed without emergency stop considerations obviously results in the lowest clearance times. However, in most cases, emergency stop time should be considered to allow a margin of safety. Given this, the acceleration condition (in which no emergency stop consideration is necessary) would create the least intersection impact. At slower speeds, or where intersection widths and train length are minimized, the condition under which the vehicle accelerates from a near-side stop and then decelerates to a far-side stop (Condition 6) would result in the next lowest intersection clearance times. In most cases, decelerating to a far-side stop with provision for emergency stopping results in the worst intersection clearance time.

LIGHT RAIL HEADWAYS

All light rail clearance times presented so far account only for an individual LRT vehicle interruption of an intersection. Most analyses involve the comparison of impacts on an hourly basis. Therefore, the light rail headways during a 1-hr period should be considered.

The hourly clearance time (T_h) required by LRT is the individual interruption clearance time (t) times the number of intersection interruptions per hour (i). The total number of intersection interruptions accounts for headways in both directions of travel and is normally figured as twice the frequency for one direction. The equation is

$$T_h = ti \quad (13)$$

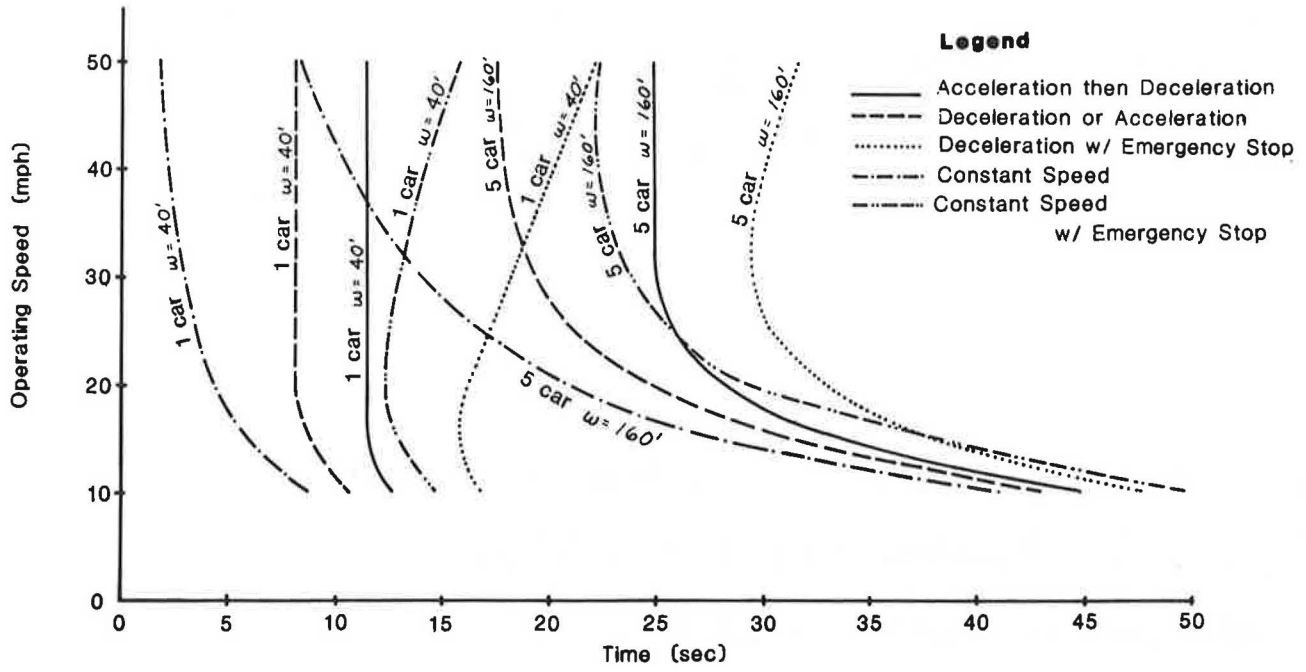


FIGURE 6 Comparison of clearance times under various operating conditions.

OTHER FACTORS AFFECTING CLEARANCE TIME

In addition to the interrelated factors discussed in the previous section, there are other factors that can cause clearance time to be added to the values calculated earlier. Two factors that will be discussed are railroad gates and clearance windows.

Railroad Gates

In some instances railroad gates are used to control traffic at light rail crossing locations. Gates are used either by themselves or in conjunction with traffic signals. The time to lower and raise gates varies somewhat from one railroad to the next, but a typical time would be about 22 sec (2). This includes about 10 sec to drop the gate and up to 12 sec to return the gate to an upright position.

When emergency stopping is considered as part of the control strategy, the railroad gates would be completely down at the time the light rail vehicle would need to begin its emergency braking rate. This allows light rail vehicle operators to see if automobiles are stopped on the tracks after the gate is completely down. If the intersection is blocked, the rail vehicle would be able to stop short of the intersection by applying the emergency stopping rate.

Clearance Windows

When using certain types of control strategies, it may be desirable to add a few more seconds to the LRT clearance interval than what is required to clear the intersection. This time, referred to as a clearance window, allows for irregularities in light rail vehicle operations. This concept is particularly applicable to the strategy in which light rail vehicles are progressed through the corridor. If, for example, the light rail vehicle is unable to maintain the progression speed, a clearance window provides a greater probability that the vehicle will arrive at the intersection without having to stop. The amount of clearance window added to the clearance time depends on several factors including potential disruptions to cross-street traffic and the reliability of the light rail progression speed. The clearance window might typically be in the range of 5 to 10 sec.

CONCLUSIONS

Several factors that influence the performance of intersections when LRT is implemented have been identified in this study. These factors can be categorized as follows:

- Light rail vehicle operating characteristics,
- Prevailing intersection geometrics, and
- Traffic control methods.

The analysis performed here has demonstrated that light rail intersection clearance time varies considerably under different conditions. For the interrelated conditions analyzed, the clearance time ranged anywhere from 3 sec up to about 50 sec. Times would be even greater if a clearance window or railroad gates were provided. In general, it was found that clearance time increases with longer train lengths and greater intersection width. Clearance time is also increased as speeds decrease, as long as emergency stop considerations are not included. When provision is made for emergency stopping, an optimum speed can be computed. This optimum speed becomes higher as train length and intersection width are increased.

It is important to remember that LRT clearance time does not necessarily show the impacts on automobile capacity at an intersection. LRT clearance time is only one component of intersection capacity analysis. Further study is needed to determine an LRT analysis procedure that is compatible with the Highway Capacity Manual.

REFERENCES

1. T.F. Larwin and H. Rosenberg. Traffic Planning for Light Rail Transit. Institute of Transportation Engineers, Washington, D.C., 1978.
2. De Leuw, Cather and Company. Southeast Corridor Preliminary Engineering: Traffic Control. Draft. Denver Regional Transportation District, Denver, Colo., Aug. 1982.

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

Evaluation of Queue Dissipation Simulation Models for Analysis of Presence-Mode Full-Actuated Signal Control

FENG-BOR LIN

ABSTRACT

Full-actuated signal control may rely on long inductive loop detectors for detecting the presence of vehicles. The operation of this mode of control is governed primarily by the interactions between the detectors and queueing vehicles. To facilitate reliable simulation analyses of such a signal control, the queue dissipation characteristics in relation to the detectors should be properly modeled. The queue dissipation models used in the NETSIM program and the Value Iteration Process--Actuated Signals program are evaluated. These models are found to be capable of producing realistic departures of queueing vehicles from a detection area. The models are rather weak, however, in representing other aspects of vehicle-detector interactions. Possible modifications of the models are discussed.

Queue dissipation is a troublesome phenomenon that has to be dealt with in the simulation of traffic flows at a signalized intersection. Proper modeling of this phenomenon is imperative if a model is to be used for simulation analysis of presence-mode full-actuated signal control. The reason for this can be found in the logic of this mode of control.

The basic logic of presence-mode full-actuated control is rather simple. A vehicle can demand or hold the green light by occupying a detection area. The detection area is usually defined by a long inductive loop detector. After the vehicle leaves the detection area, the green is extended by a duration equal to a preset vehicle interval. To continue holding the green, another vehicle must enter one of the detection areas of the same signal phase before the vehicle interval expires. The phase duration is limited by a preset maximum green interval. The timing of this interval begins with the actuation of a detector by a vehicle in an opposing phase.

Because the vehicle interval is usually set at a value close to 0 sec, the queueing vehicles in a lane have a much better chance of holding the green than do those not in the queue. Consequently, the phase durations of this mode of control are governed by the queue dissipation characteristics in relation to the detectors. In a dissipating queue, vehicles enter and depart from a detection area in a dynamic and probabilistic manner. This results in a sequence of detector actuations and departures that determines whether the queueing vehicles can extend the green continuously. If such relationships are not properly modeled, the simulated operation of the signal control will deviate from reality.

The dynamic and probabilistic nature of the interactions between the queueing vehicles and the detectors is rather difficult to simulate adequately. To compound the problem, past efforts at modeling queue dissipation were focused on queue discharge headways (1). Not until recently have efforts been made to investigate the nature of the queue dissipation in relation to presence detectors (2). The lack of a comprehensive treatment of this subject is unsettling. It raises the issue of whether the queue dissipation models used in existing signal simulation programs are realistic.

The purpose of this paper is to explore this issue by evaluating two existing queue dissipation simulation models. These models are part of the microscopic simulation models used, respectively, in the Value Iteration Process--Actuated Signals program (3) and the NETSIM program (4). The evaluation

entails a comparison of simulated queue dissipation characteristics with those observed in the field.

The field data used for the comparison are extracted from data collected as part of a project sponsored by the U.S. Department of Transportation (5). The original data include observations made on queueing flows of various directional movements. These flows were not interfered with by pedestrians and parking maneuvers, nor were they impeded by downstream flows. The approach speeds of the observed vehicles were less than 35 mph. Without sacrificing the generality of the comparison, only straight-through flows are dealt with in this paper. Furthermore, because the field observations include a negligible number of trucks and buses, all vehicles are considered to be passenger cars.

QUEUE DISSIPATION SIMULATION MODELS

Pitt Car-Following Model

The Value Iteration Process--Actuated Signal (VIPAS) computer program uses the Pitt car-following model (6) for simulating queue dissipation. The model was originally developed for freeway simulation. It contains a basic equation and several constraints. The basic equation is

$$a = 2[x^* - y - L - 10 - v(k + T) - bk(u^* - v)^2]/(T^2 + 2kT) \quad (1)$$

where

- a = acceleration of follower in the interval $(t, t + T)$,
- k = car-following parameter (driver sensitivity),
- L = length of the leading vehicle,
- T = time scanning interval,
- y = position of follower at time t,
- v = speed of follower at time t,
- x^* = position of leader at time $t + T$,
- u^* = speed of leader at time $t + T$, and
- b = constant.

The units in this equation are feet and seconds.

The Pitt car-following model is based on the premise that a following vehicle will attempt to maintain a space headway of $L + kv + 10$ ft. To prevent collision between two vehicles in a car-following maneuver, several constraints are incorporated into the model to override the acceleration rate as determined from the basic equation. These constraints ensure that a following vehicle can stop safely behind its leader under two conditions: the leader decelerates to a stop at a specified maximum emergency deceleration rate, and the follower starting at a driver reaction or lag time (c) later decelerates to a stop behind the leader at a deceleration rate within the maximum emergency deceleration limit. Mathematical relationships are used to impose these constraints on car-following behavior.

The model described has been modified in the VIPAS program for simulating car-following maneuvers. One modification involves the replacement of the constant 10 in Equation 1 with

$$A = \begin{cases} 7 + (3v/20) & \text{if } v \leq 20 \quad (2a) \\ 10 & \text{if } v > 20 \quad (2b) \end{cases}$$

Another modification deals with the driver sensitivity parameter (k). The value of k as used in the original model approximates (3)

$$k = (2{,}718 - Q_s)/687.5 \quad (3)$$

where Q_s is the saturation flow rate applicable to the flow in a given lane. The value of k is adjusted in the VIPAS program to

$$k = 3.539 - 0.0012985\ Q_s \quad (4)$$

To account for differences in car-following behavior among individual drivers, a multiplication factor is also used in the VIPAS program. This factor is generated randomly according to a specified probability distribution. In this distribution the multiplication factor is divided into the following seven levels: 1.7, 1.2, 1.1, 1.0, 0.9, 0.7, and 0.5. The probabilities associated with these levels are, respectively, 0.1, 0.2, 0.1, 0.2, 0.1, 0.2, and 0.1. A generated value of this factor is applied to the saturation flow rate to obtain a k value for a driver.

For queue dissipation, the VIPAS program uses b = 0 in Equation 1. The scanning interval is 1 sec. Furthermore, the acceleration rate is limited to 7 ft/sec^2 for passenger vehicles moving at speeds of not more than 16 ft/sec and to 5 ft/sec^2 when the speeds exceed 20 ft/sec. The maximum emergency deceleration rate is assumed to be 10 ft/sec^2.

After an acceptable acceleration rate is calculated from the model, the speed and position of a following vehicle at time t + T (i.e., the end of the scanning interval) are computed as

$$v^* = v + a\ (T - c) \quad (5)$$

and

$$y^* = y + vT + [a(T - c)^2]/2 \quad (6)$$

where

- v^* = speed of follower at time $t + T$,
- y^* = position of follower at time $t + T$, and
- c = driver reaction time ($c < T$).

The driver reaction time (c) has a calibrated value of 0.2 for deceleration and 0.3 for acceleration.

NETSIM Model

The queue dissipation model used in the NETSIM program is essentially a heuristic algorithm consisting of a number of decision rules. These rules are contained in the MOVE, ADJQ, GOQ, and other related subroutines of the program. They determine the speed, acceleration, and position of a vehicle in each scanning interval of 1 sec. The model is not as straightforward as the Pitt car-following model. The flow chart shown in Figure 1 gives an insight into the heuristic nature of the model. The flow chart omits those vehicle-processing functions of the model that are beyond the scope of this paper.

In processing the dissipating queueing vehicles in a given scanning interval, the leading vehicle of a queue upstream of an intersection is identified first. If this vehicle is also the leading vehicle when the signal is red, then it is assigned a loss time. This loss time determines when the vehicle will discharge into a downstream lane. The loss time can assume a deterministic value as specified by the user of the NETSIM program. It may also be generated from one of two imbedded distributions (4). As an example, one such distribution consists of the following 10 levels of loss time: 5.6, 3.6, 3.2, 3.0, 2.6, 2.2, 2.0, 1.6, 1.2, and 0.6 sec. Each of these loss times has the same probability of being generated in the simulation process.

If a leader can discharge in the current scanning

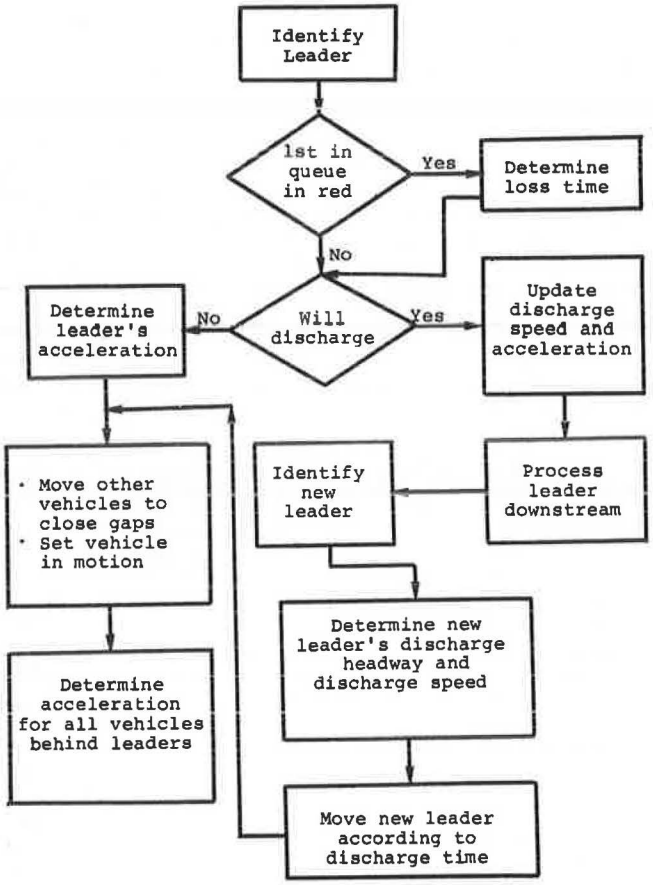


FIGURE 1 Schematic of NETSIM queue dissipation simulation process in a one-second scanning interval.

interval, its discharge speed and acceleration are determined. The vehicle is then processed into a downstream lane. The discharge speed of this vehicle becomes the basis for determining that of the following vehicle. The change in the discharge speed from one vehicle to another follows a predetermined pattern. The discharge speed of the first queueing vehicle at the onset of the green is assumed to be 9 ft/sec. This speed is increased by 4 ft/sec for each additional discharging vehicle until it reaches 40 ft/sec.

Following the processing of a leader that will discharge, a new leader is identified. The discharge headway of the new leader is then generated to determine the corresponding discharge time. This headway may also be generated randomly from an imbedded probability distribution. In addition to the discharge headway, the intended discharge speed of the new leader is determined. The current speed of the new leader is subsequently compared with this discharge speed to determine whether the vehicle should accelerate or decelerate. If an acceleration is warranted, an acceleration rate of 4 ft/sec^2 is applied to passenger cars that are moving at speeds of less than 17 ft/sec. When vehicle speeds exceed 17 ft/sec, a rate of 3 ft/sec^2 is used instead.

These threshold acceleration rates may be revised in the scanning interval when a vehicle first becomes the new leader. In such an interval the time remaining before discharge is computed for the new leader. This time duration is then used to calculate the average speed required for the new leader to discharge on time. The average speed, in turn, forms the basis for determining the needed acceleration rate. This newly calculated rate is limited to a maximum of 5 ft/sec^2. It is further reduced to 3 ft/sec^2 if the current speed exceeds 16 ft/sec.

In every scanning interval the model also checks to see if a queueing vehicle can start moving. It is assumed that queueing vehicles will start moving in succession at 1-sec intervals after a green phase begins. The model sets the queueing vehicles in motion according to this backward wave speed of one vehicle length per second. The spacing of the queueing vehicles that remain stationary is checked. The model adjusts the spacings so that they equal a specified effective length of the vehicles.

For vehicles that are set in motion behind the leader, normal acceleration rates are set at a value of 6 ft/sec^2 for speeds of less than 17 ft/sec and 3 ft/sec^2 for speeds exceeding 17 ft/sec. The initial acceleration rates allowed for stationary vehicles are 6 to 8 ft/sec^2, depending on the queueing positions of such vehicles. A car-following relation (7) is incorporated into the model to prevent a collision in the event that a leading vehicle implements a panic stop at a deceleration of 10 ft/sec^2. This car-following relation defines an upper bound of the acceleration rates. Furthermore, the acceleration of a vehicle is limited to a value that will not cause its speed to exceed the designated discharge speed.

EVALUATION

The movements of dissipating queueing vehicles bring about a series of interactions between the vehicles and the detectors. The basic elements of such interactions in a traffic lane include the departure of a vehicle from a detection area, the arrival of a vehicle at the upstream end of a detector, and the dwell of a vehicle in a detection area. These three elements can be modeled in terms of departure time, arrival time, and dwell time.

Departure time of a vehicle is used herein to denote the time elapsed from the onset of the green to the moment the rear bumper of the vehicle crosses the downstream end of a detector. Departure time can be represented by the sum of a sequence of queue discharge headways. Arrival time represents the time elapsed from the onset of the green to the moment the front bumper of a vehicle reaches a detector. This definition applies only to those vehicles upstream of a detector at the onset of green. Dwell time is the difference between departure time and arrival time.

These time-related variables are under the influence of drivers' behavior. Because drivers differ in their behavior, substantial variations in the values of these variables have been observed. For example, departure times and arrival times of those vehicles in a given queueing position could differ by as much as 6 sec. The corresponding dwell times also vary from 40 percent to more than 200 percent of their average value. The queue discharge headways can be expected to vary from less than 1 sec to more than 5 sec for vehicles in the same queueing position.

These variations are difficult to simulate, but their inclusion in the simulated characteristics of a queueing flow is indispensable. Consider the representative average observed departure times and arrival times given in Table 1. The average characteristics indicate that, with a 30-ft detector, those vehicles in the fourth queueing position or further upstream cannot move into the detection area before the vehicle ahead departs from the same area. If a model applies such average characteristics to individual vehicles, a distorted picture will

TABLE 1 Representative Average Departure Times and Arrival Times of Straight-Through Queueing Vehicles

Queueing Position	Departure Time (sec)	Arrival Time (sec) 30-ft Detector	Arrival Time (sec) 50-ft Detector
1	3.3		
2	5.9	2.9	
3	8.3	5.9	4.5
4	10.6	8.5	7.4
5	12.8	10.9	9.9
6	15.0	13.2	12.2
7	17.1	15.4	14.4
8	19.2	17.6	16.7
9	21.3	19.8	18.8
10	23.4	22.0	20.9

emerge. For example, consider a signal phase that involves only a single-lane flow and has a vehicle interval of 0 sec. In such a case the average characteristics would not allow the fourth queueing vehicle to extend the green. Thus, a premature termination of the green would take place whenever more than three vehicles were in the queue. On the other hand, the average departure times and arrival times associated with a 50-ft detector would allow every queueing vehicle to hold the green.

In reality, not every queueing vehicle in the fourth queueing position or further upstream is unable to move into a 30-ft detector early enough to hold the green. Similarly, not all queueing vehicles can extend the green continuously when a 50-ft detector is used. This phenomenon is shown in Figure 2 in terms of the probabilities of premature termination of the green. These probabilities are derived from observed queue dissipation characteristics of single-lane flows. They reveal that the premature termination of the green could occur even when a 50-ft detector is used along with a vehicle interval of 2 sec. When two lanes are associated with a signal phase, 50-ft detectors in combination with 0-sec vehicle intervals can still lead to a high probability of premature phase termination (5).

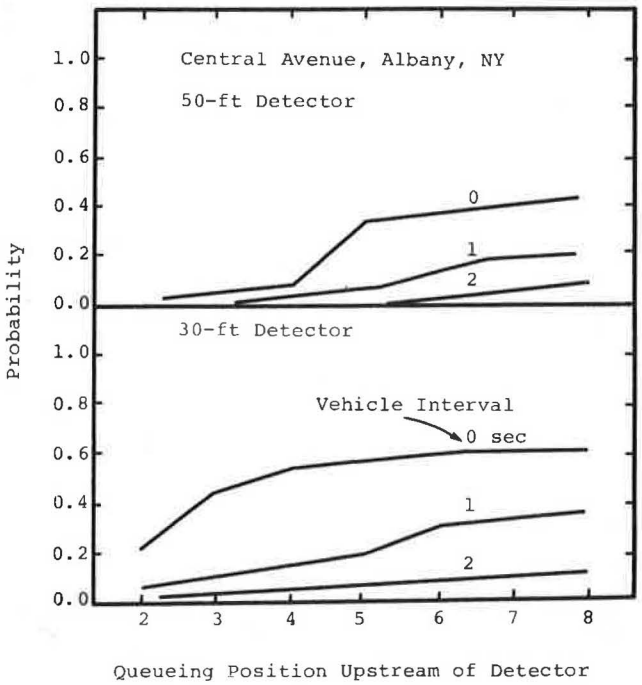


FIGURE 2 Probabilities of queueing vehicles facing prematurely terminated green duration.

Therefore, the variations in the queue dissipation characteristics of individual vehicles have to be realistically represented in a simulation model. Otherwise, significant errors may be introduced into the simulated operation of a signal control. This would reduce the usefulness of a model for comparing alternative control strategies.

In light of these implications, there is a minimum requirement a simulation model should satisfy. This requirement demands that simulated departure, arrival, and dwell characteristics conform to probability distributions that can be identified in the field. These three aspects of the queue dissipation characteristics are not unrelated. For a given vehicle, any two of these aspects define the remaining one. Therefore, it is not necessary to test a model in terms of all the aspects. In this paper the Pitt car-following model and the NETSIM model are tested in terms of departure and dwell characteristics.

Pitt Car-Following Model

The application of this model requires the specification of a saturation flow rate as an input. To test the model, a saturation flow rate of 1,700 veh/hr is used to represent the movements of straight-through queueing flow. This flow rate corresponds to a stabilized queue discharge headway of about 2.1 sec. The driver reaction time (c) used for the test is 0.3 sec for acceleration and 0.2 sec for deceleration as specified in the VIPAS program. A higher value of 0.5 sec for acceleration is also used to examine the sensitivity of the model outputs to the choice of c. Following the VIPAS program, the first queueing vehicle is assumed to accelerate at 7 ft/sec^2 when its speed is less than 17 ft/sec and at 5 ft/sec^2 when its speed exceeds 20 ft/sec. The space headways between stationary queueing vehicles are allowed to vary randomly between 20 and 30 ft.

The departure and dwell characteristics of the queueing vehicles in the Pitt car-following model are internally generated data. Therefore, both aspects have to be compared with the field data. To facilitate the comparison, the characteristics of 60 dissipating queues are simulated.

The data given in Table 2 indicate that the average discharge headways generated from the model agree rather well with the observed values. To test the variations in the simulated individual discharge headways, the headways of those vehicles in a given queueing position are normalized by expressing them

TABLE 2 Observed and Simulated Average Queue Discharge Headways

Queueing Position	Discharge Headway (sec) Observed Case 1	Observed Case 2	Pitt Model
1	3.2	3.6	3.3
2	2.5	2.6	2.8
3	2.3	2.4	2.5
4	2.3	2.4	2.3
5	2.2	2.4	2.3
6	2.2	2.1	2.2
7	2.2	2.2	2.3
8	2.1	2.2	2.2
9	2.2	2.1	2.2
10	2.1	2.1	2.1

as percentages of their average value. The field data reveal that the cumulative probability distributions of such normalized headways fall in a well-defined range. The two solid lines in Figure 3 define the representative range of the observed distributions. The simulated distributions are also shown in the same figure.

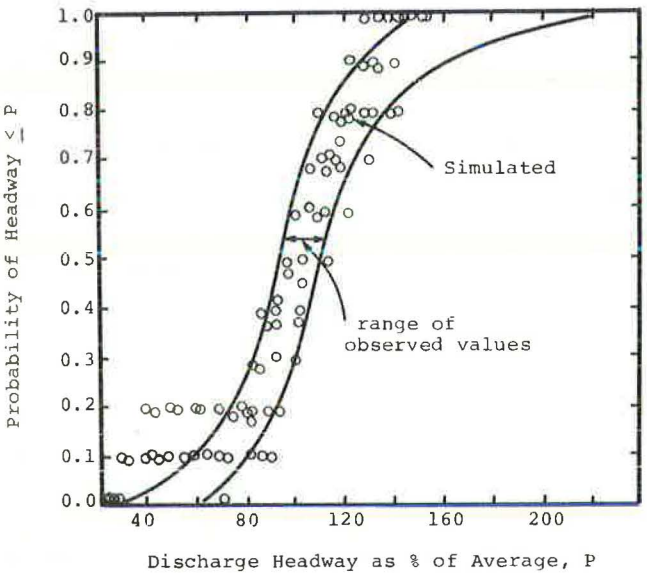


FIGURE 3 Simulated discharge headway distributions by Pitt model in comparison with observed distributions.

It can be seen from Figure 3 that the simulated distributions lack headways that exceed 140 percent of the averages. In contrast, the field data indicate that there is about a 10 percent chance that a discharge headway will exceed 140 percent of the average. The model also tends to produce a larger percentage of small headways. Between 60 and 140 percent of the averages, however, the simulated distributions conform reasonably well to the observed distributions.

The simulated average dwell times follow the general trends of the observed values. Two types of discrepancies, however, are evident in Figure 4. For a vehicle immediately upstream of a detector at the onset of the green, the simulated dwell time is much longer than the observed value. Those vehicles much further back in the queue have simulated values that are smaller than the observed values. The discrepancies involving 50-ft detectors and 80-ft detectors are relatively large. These discrepancies may be manipulated by changing the space headways of the stationary queueing vehicles. However, reducing the discrepancies for those in the front of a queue may increase the discrepancies for those in the back.

Because the simulated departure characteristics are in close agreement with the observed data, such discrepancies must be due to biases in the simulated arrival times. On the basis of Figure 4, it appears that the simulated queueing vehicles just upstream of a detector at the onset of green move into the detection area too early whereas those further upstream move into the detection area too late. One possible cause of this problem is the use of the same car-following rule in the model for vehicles in all queueing positions. This feature of the model makes it difficult to calibrate the model to satisfy the observed departure and dwell characteristics simultaneously.

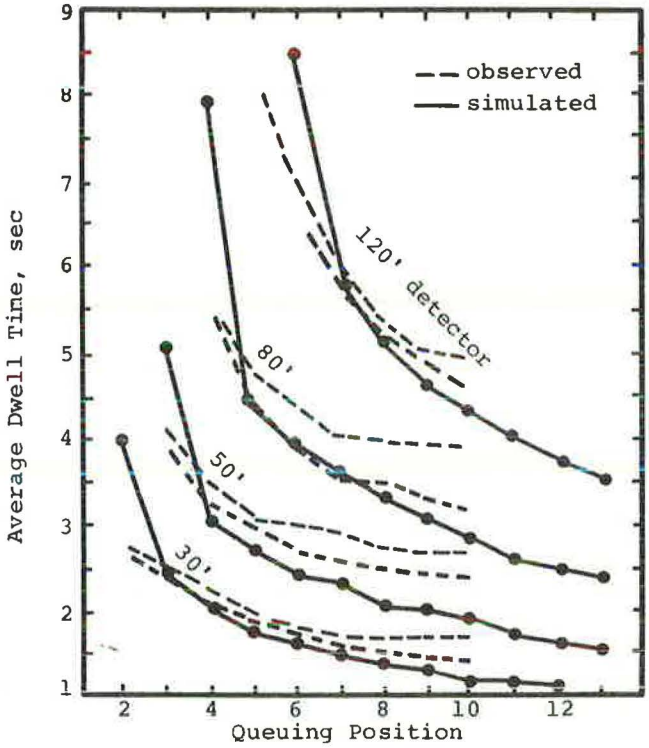


FIGURE 4 Simulated average dwell times by Pitt model in comparison with observed values.

For example, a longer driver reaction time (c) may be used to raise the simulated dwell times. The data in Table 3 indicate how an increase of c from 0.3 to 0.5 sec for acceleration affects the simulated queue discharge headways and dwell times. The use of c = 0.5 sec has appreciable effects on the dwell times of only the first two vehicles upstream of a detector. At the same time, the queue discharge headways become unreasonably high although the input saturation flow rate remains the same. This simple example indicates that the limited number of variables used in the model cannot adequately explain the wide range of queue dissipation characteristics. Allowing vehicles in different queueing positions to follow different car-following rules may be one way of alleviating this problem.

TABLE 3 Sensitivity of Simulated Dwell Times and Queue Discharge Headways to the Choice of c

Queueing Position	Discharge Headway (sec)		Average Dwell Time (sec)			
			30-ft Detector		80-ft Detector	
	c = 0.3	c = 0.5	c = 0.3	c = 0.5	c = 0.3	c = 0.5
1	3.3	3.3				
2	2.8	3.7	4.0	4.5		
3	2.5	2.9	2.4	2.8		
4	2.3	2.5	2.0	2.0	7.9	8.6
5	2.3	2.6	1.8	1.7	4.7	5.0
6	2.2	2.6	1.6	1.7	3.9	4.0
7	2.3	2.6	1.5	1.6	3.5	3.6
8	2.2	2.5	1.4	1.5	3.3	3.4
9	2.2	2.5	1.3	1.4	3.0	3.2
10	2.1	2.5	1.2	1.4	2.8	3.0

In addition to the problem associated with average dwell times, the simulated distributions of in-

dividual dwell times also differ significantly from the observed distributions. Figure 5 shows that the simulated dwell times of individual vehicles in a given queueing position are concentrated in a range from 90 to 110 percent of their average. In contrast, the field data show a variation of from about 40 percent to more than 200 percent of the average. If deterministic space headways are used for stationary queueing vehicles, the simulated distributions have even lower variations. Apparently, the treatment of the driver sensitivity factor as the only probabilistic variable is not sufficient to account for real-life variations. This weakness will probably become insignificant if more variables are included in the model and the driver reaction time (c) is also treated as a random variable.

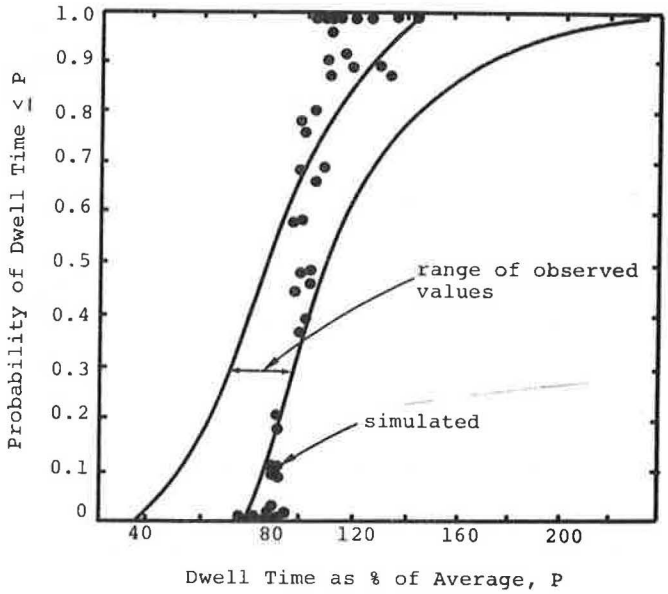


FIGURE 5 Simulated dwell time distribution by Pitt model in comparison with observed distribution.

NETSIM Model

To test the queue dissipation model of the NETSIM program, an effective length of 25 ft for passenger cars is used as an input. The space headway between two adjacent queueing vehicles at the onset of the green is made equal to this effective length. At the same moment, the front end of the first vehicle in a queue is assumed to coincide with the downstream end of the detector.

Another input specified for the model is an average discharge headway of 2.1 sec for those vehicles in the fourth queueing position or further upstream. This input is used with the imbedded discharge time of each vehicle. The headway distribution allows the generated headways to vary from 50 to 170 percent of the specified average of 2.1 sec. For vehicles in the second and the third queueing positions, 0.5 and 0.2 sec are added, respectively, to the generated headways. This implies that vehicles in the second queueing position have an average discharge headway of 2.6 sec and those in the third queueing position have a 2.3-sec average. The discharge times of the vehicles in the first queueing position are generated separately from the loss time distribution described previously.

Discharge time is considered to be the time at which a queueing vehicle moves and reaches the downstream end of a detector after a green phase begins. Therefore, this vehicle will travel a car length before it departs from the detection area. The length of a car is assumed to be 20 ft. The departure times are determined from the discharge times generated in the simulation process. The arrival times are estimated from the speed profiles of individual vehicles. These estimates are then used to compute dwell times. The movements of 120 dissipating queues are simulated to provide needed statistics for examining the model.

The average simulated dwell times as a function of the queueing position are shown in Figure 6. The simulated values differ from the observed values in two respects. First, the simulated values for vehicles in the fifth through the tenth queueing positions tend to be smaller than the observed values. Second, instead of continuing to decrease as expected, the simulated values of those vehicles in the eighth queueing position or further upstream increase with queueing position. These two types of deviations are systematic instead of random among detectors of 30 to 120 ft in length. This suggests that the model has built-in biases in representing observed queue dissipation characteristics. Such biases are also rather evident in Figure 7 when the distributions of the simulated dwell times are compared with those of the observed dwell times.

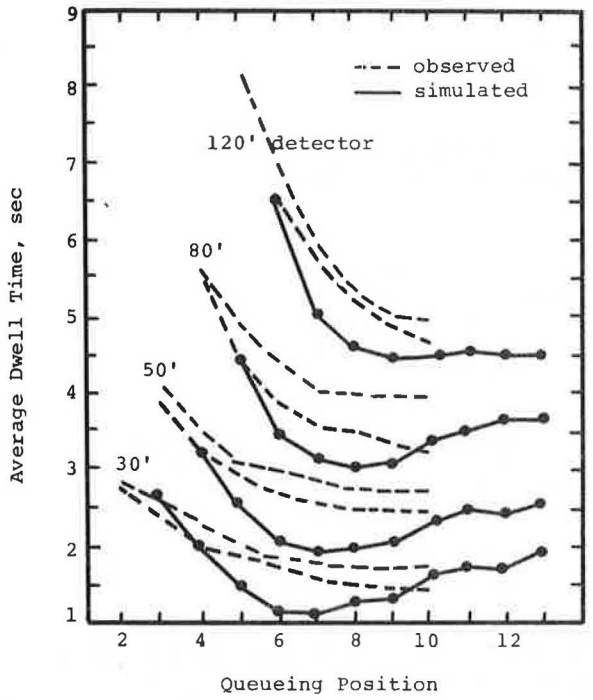


FIGURE 6 Simulated average dwell times by NETSIM in comparison with observed values.

The simulated departure times do not play a significant role in creating the biases in the dwell times. They are found to be a good representation of the observed characteristics. It should be cautioned, however, that the degree of realism in the simulated departure times depends in part on the loss time generated for the vehicle in the first queueing position. Such loss time is a parameter not clearly defined in the documentation of the NETSIM program. An examination of the MOVE subroutine indicates that loss time represents the time elapsed

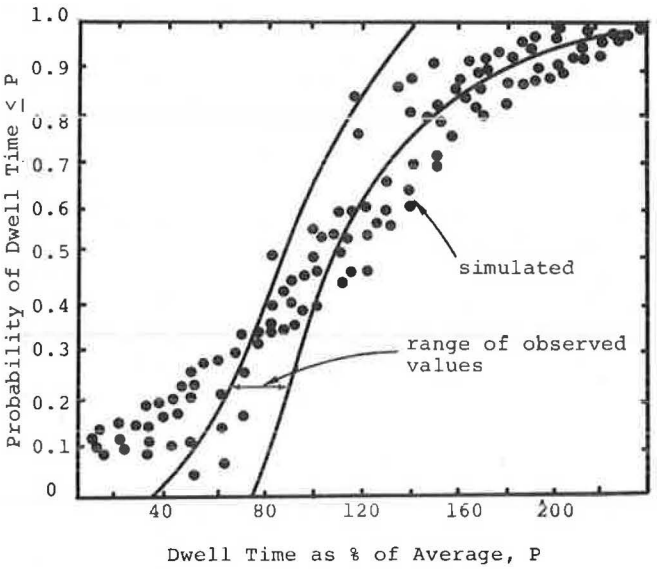


FIGURE 7 Simulated dwell time distribution by NETSIM in comparison with observed distribution.

from the onset of green to the moment the first queueing vehicle is allowed to move. The imbedded loss time distribution used for the test has an average value of about 2.6 sec. If this distribution were replaced by a user-specified input of 1.6 sec, the average departure time of vehicles in any queueing position would be reduced by about 1 sec. This would aggravate the biases in the simulated dwell times. Therefore, the specification of loss time should take into consideration the departure times that are being simulated.

A major weakness of the model is its lack of responsiveness to variations in the discharge times of those vehicles in a given queueing position. This feature could lead to incompatibilities between the generated discharge times and the speed profiles of individual vehicles. As described previously, the discharge time of a vehicle is not determined until the vehicle becomes the new leader in a certain scanning interval. This discharge time is used only in the same scanning interval to adjust vehicle speed. After this interval the speed of the new leader is adjusted according to predetermined acceleration rates. Furthermore, before a vehicle becomes a new leader, its acceleration is governed by a different set of threshold values, by a need to prevent collisions, and by the requirement that the resulting speed be no more than the designated discharge speed. Overall, this simulation process has little to do with the time at which a vehicle is expected to discharge. Consequently, by the time a vehicle becomes the new leader, it may be at a location and moving at a speed that does not allow it to discharge at the designated moment in a normal fashion. This is one reason why the simulated dwell times are biased. This problem can be avoided if vehicles are allowed to discharge only according to deterministic headways. In that case, the model could be calibrated readily to fit average characteristics of queue dissipation. However, the simulated distributions of departure times and dwell times would be unrealistic.

There are other factors that may contribute to the discrepancies between the simulated and the observed dwell times. For example, discharge speeds are increased in the model at a rate of 4 ft/sec per discharging vehicle. This allows the discharge speed of the ninth vehicle in a queue to reach 40 ft/sec. On the basis of the observed average dwell times in 30-ft detection areas and a vehicle length of 20 ft, the average speeds in such detection areas can be estimated for vehicles in different queueing positions. These speeds approximate the discharge speeds. Figure 8 shows that the rates of increase in discharge speeds are only about 1 ft/sec per vehicle for those behind the sixth queueing position; the average discharge speed of those vehicles in the ninth queueing position is on the order of 30 ft/sec.

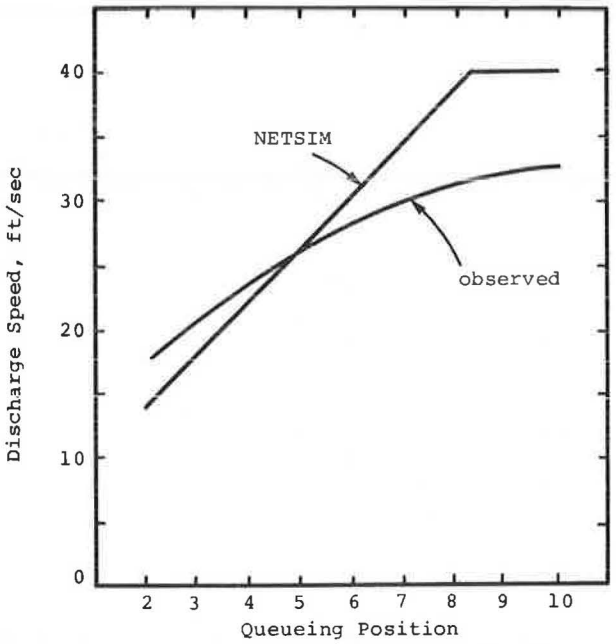


FIGURE 8 Allowable discharge speeds in NETSIM and estimates based on field data.

A few modifications can be made to alleviate the shortcomings of the model. One modification is to adjust the speed of a leader continuously with respect to the discharge time. Another modification is to generate the discharge time of a vehicle either at the onset of the green or in the scanning interval when this vehicle joins a queue. The generated discharge time is then used as a determinant of vehicle acceleration. These modifications provide more time for the model to adjust the speed profile of a vehicle to match the discharge time.

The discharge speeds used in the model can also be revised to conform to the derived relationship shown in Figure 8. This requires only minor changes in the ADJQ subroutine. A critical modification that should be made involves the allowable acceleration rates for vehicles behind a leader. For stationary vehicles, the model allows initial acceleration rates of from 6 to 8 ft/sec^2, depending on the queueing position. When a vehicle has started moving, the allowable rate is either 6 ft/sec^2 or 3 ft/sec^2. These rates should have been calibrated in terms of observed dwell characteristics.

A summary of the allowable acceleration rates that produce a reasonably close fit to the observed dwell times is given in Table 4. Generally, these rates indicate that, in the first few seconds after a vehicle starts moving, the vehicle should be subject to substantial restrictions in its movement. The restrictions need to be more severe for those in the back of a queue than for those in the front.

TABLE 4 Allowable Acceleration Rates of Vehicles Behind a Leader

Distance from Stop Line (ft)	Speed (ft/sec)	Allowable Acceleration (ft/sec^2)
$D \leq 100$	0	2
$D > 100$	0	1
$D < 100$	< 3	3
$150 \leq D < 180$	< 2	2
$180 \leq D < 200$	< 3	2
$200 \leq D < 230$	< 4	2
$D > 230$	< 7	2
All other conditions		3

The acceleration rates given in Table 4 are allowed to be increased or decreased with respect to the discharge times when vehicle speeds exceed 25 mph. They are also subject to the same constraint that was applied in the original model to prevent collisions. The average dwell times simulated according to the modifications described are shown in Figure 9. The agreement between the simulated and the observed average dwell times is good. The simulated dwell time distributions are also found to conform to the observed distribution.

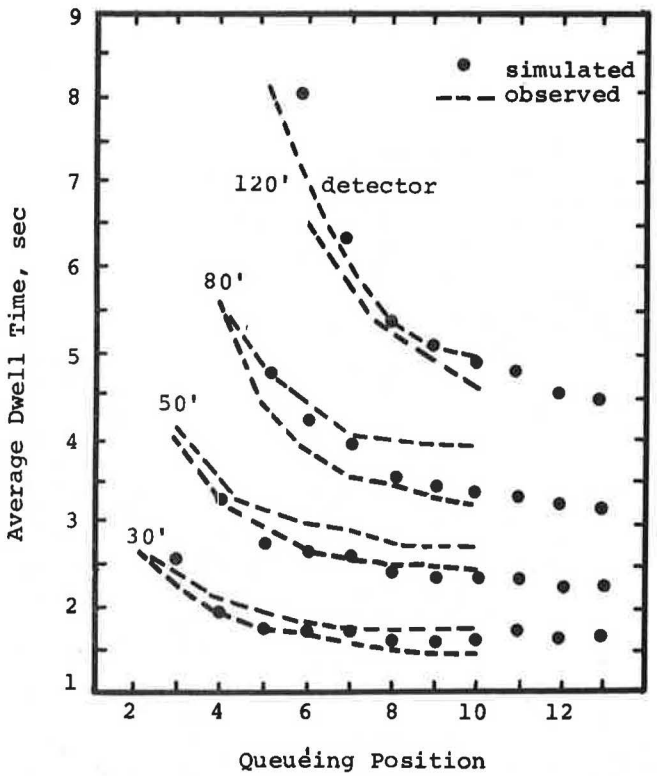


FIGURE 9 Simulated average dwell times by modified NETSIM in comparison with observed values.

CONCLUSIONS

For simulation analysis of presence-mode full-actuated signal control, the queue dissipation models used in the VIPAS program and the NETSIM program, respectively, have significant limitations. Both models are capable of producing realistic departure characteristics of queueing vehicles in relation to detectors, but they are inadequate in representing other vehicle-detector interactions such as the dwell times of vehicles. This reduces the usefulness of both simulation programs for analysis of a wide range of traffic flow and signal control conditions.

The Pitt car-following model used in the VIPAS program relies on mathematical relationships to simulate car-following maneuvers. It can be easily adopted for various applications. However, its limited number of variables and parameters does not provide a flexible model structure for calibration. The outputs of the model cannot be made to conform easily and simultaneously to observed departure, arrival, and dwell characteristics of queueing vehicles. The dwell times produced by the model are also more or less deterministic. This contrasts with the large variations observed in the field. The model needs additional variables to better represent queue dissipation characteristics in relation to detectors. More random variations should also be introduced into the simulated dwell characteristics of queueing vehicles.

The queue dissipation model used in the NETSIM program is essentially a heuristic algorithm for processing vehicles. This program relies on deterministic threshold acceleration rates and discharge speeds to guide the movements of queueing vehicles. A major weakness of the model is that the simulated movements of queueing vehicles have little to do with the discharge times generated separately from a probability distribution. Consequently, the simulated speed profiles may not be compatible with the discharge times. This, in turn, leads to systematic biases in the simulated dwell characteristics.

This model can be modified in several respects. For example, the discharge times of individual vehicles can be generated long before the vehicles are about to discharge. These discharge times can then be used continuously to guide the vehicle movements. And, more important, the threshold acceleration rates and discharge speeds should be modified to reflect actual variations in queue dissipation characteristics from one queueing position to another. With these modifications, a significant improvement of the simulated queue dissipation characteristics can be achieved.

Past efforts in modeling queue dissipation for simulation analysis of traffic-actuated signal controls have paid little attention to vehicle-detector interactions. This negligence could direct the search for improved signal controls in wrong directions.

ACKNOWLEDGMENT

This study is sponsored in part by the U.S. Department of Transportation Program of University Research.

REFERENCES

1. G.F. King and K.M. Wilkinson. Relationship of Signal Design to Discharge Headway, Approach Capacity, and Delay. In Transportation Research Record 615, TRB, National Research Council, Washington, D.C., 1976, pp. 37-44.
2. F.B. Lin. and M.C. Percy. Vehicle-Detector Interactions and Analysis of Traffic-Actuated Signal Controls. In Transportation Research Record 971, TRB, National Research Council, Washington, D.C., 1984, pp. 105-111.
3. J.A. Breon. Value Iteration Process-Actuated Signals, Vol. 1. Fiscal and Systems Management Center, Pennsylvania Department of Transportation, Harrisburg, 1983.

4. Traffic Network Analysis with NETSIM--A User Guide. Implementation Package FHWA-IP-80-3. FHWA, U.S. Department of Transportation, 1980.
5. F.B. Lin. Optimal Timing Settings and Detector Lengths of Presence-Mode Full-Actuated Control. Draft Final Report. Office of University Research, U.S. Department of Transportation, Dec. 1984.
6. Development and Testing of INTRAS, A Microscopic Freeway Simulation Model, Vol. 1. Report FHWA/RD-80/106. FHWA, U.S. Department of Transportation, 1980.
7. Network Flow Simulation for Urban Traffic Control System--Phase II, Vol. 5. FHWA-RD-77-45. FHWA, U.S. Department of Transportation, 1977.

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

Another Look at Identifying Speed-Flow Relationships on Freeways

BRIAN L. ALLEN, FRED L. HALL, and MARGOT A. GUNTER

ABSTRACT

Despite approximately 50 years of research on highway operating characteristics, the way in which the speed-flow relationship moves between free flow and congested flow conditions is still not clearly understood. The speed-flow relationship as it pertains to those transitions is investigated using an extensive data set collected on the Queen Elizabeth Way freeway in Ontario. Two different analytical approaches are used: time-connected plots of mean speed and mean flow and an event-based trace, averaged with respect to the transition to and from congested flow. The results confirm several aspects of conventional understanding but also raise questions that are hard to answer with the conventional interpretations of speed-flow relationship.

Highway capacity and the operational characteristics of uninterrupted traffic flow have been the focus of research for at least the past 50 years. Since Greenshields' work in the 1930s, traffic researchers have devoted considerable attention to investigating and interpreting the fundamental characteristics of traffic flow on freeways. The dozens of research papers produced during those 50 years have typically documented either the results of interpretive empirical studies or the degree of success achieved in relating those results to known or proposed theoretical concepts (models). Certainly it is a well-known and well-researched area. Why then another paper on this subject?

The answer consists of three complementary parts. First, the subject area itself remains relevant. As freeway systems become more complex and experience ever-increasing traffic loads, the effective management of those systems becomes increasingly important. To manage them effectively, operating agencies must have reliable information about traffic flow on which to base appropriate actions. This is true whether overall system management or more specific and data-demanding activities, such as entrance ramp control or incident detection and response, are considered.

Second, it is well accepted that the representation of speed-flow relationships first presented in the Highway Capacity Manual (HCM) (1), and more recently updated in the Interim Materials on Highway Capacity (2), cannot completely reflect all actual conditions witnessed on the many different existing freeway systems. There is little doubt that at the very least a further updating of those curves is required to reflect current operating characteristics, even if the only result is simply to "calibrate," on a site-specific basis, the traditional approach to highway capacity and level of service.

This leads to the third part of the answer. In trying to calibrate the traditional approach with their data, many traffic engineers have had difficulty with the traditional approach to describing the speed-flow relationship as a smooth, continuous curve as depicted in the HCM material and the vast majority of standard traffic engineering references. Some researchers, suspecting that the curves are not in fact continuous and perhaps not always smooth, have proposed other solutions. Perhaps it is appropriate even after 50 years of research to take yet another look at precisely what does happen with traffic flow on freeways.

In this paper only the relationship between speed

and flow on freeways, particularly as the flow moves between the congested and uncongested regimes, is discussed. The next section provides some background for the analysis by including some of the expressions of dissatisfaction with current understanding of traffic flow characteristics. That discussion also helps to focus on possible analytical techniques by suggesting why other analysts are not satisfied with their results. In the subsequent sections the data set used is described, the results of the analysis are presented, and the conclusions that appear to follow from the analysis are provided.

BACKGROUND

Most researchers would surely agree that there are three primary issues related to understanding freeway traffic flow relationships:

1. Identification of the basic variables.
2. Formulation of the fundamental interactions among these variables. This includes consideration of functional relationships that describe the nature of uninterrupted flow throughout the range of low volume-to-capacity ratios (V/C), high V/C, congested operation, and transitions between them.
3. Quantification of the flow relationships, that is, identification of the numerical magnitudes of ideal lane capacity, free flow speed, and so on.

There is certainly agreement regarding the first issue: speed, flow, and density are the variables of interest. Some concern has been expressed in the past over the use of particular speed measures, but that debate appears to have been practically resolved with the more recent adoption of average (running) speed. This will be the speed referred to in the remainder of this paper. Similarly, the vast majority of researchers use short duration volume counts expanded to hourly flow rates. Again, there has been debate over how short the time durations should be, but generally it is accepted that the shorter durations (1, 2, or 5 min) are best suited for investigations of flow interactions and characteristics. Because the present work is concerned exclusively with identifying such interactions, 5-min volumes and resultant hourly flow rates will be used throughout. Finally, the usefulness of density or occupancy rates cannot be denied. The measure is appropriate to level of service concerns and helpful in identifying basic relationships. Because this work is concerned primarily with speed-flow relationships, however, few comments will be made regarding density.

Although the second issue of formulating fundamental relationships has not been as readily resolved, there are important areas of agreement. Most traffic engineers would agree that the basic shape is somewhat similar in nature to the curves shown in Figure 1. In particular, it is accepted that traffic operates in either an uncongested state on the upper branch of the curve or in a congested state on the lower branch of the curve. This basic form was well represented by the HCM curve (Curve 1) in the figure (1). There is also recent agreement that speeds on the upper branch remain relatively constant over much larger ranges of flow rate than previously depicted, as typified by the other plots (Curves 2-4) in Figure 1 (2-4). In addition, there is agreement that congested operations occur approximately as shown by the lower branch in the figure. There is less agreement on the way in which the upper and lower branches are joined and on the operational nature of transitions between the branches. It is possible that these transitions are dependent

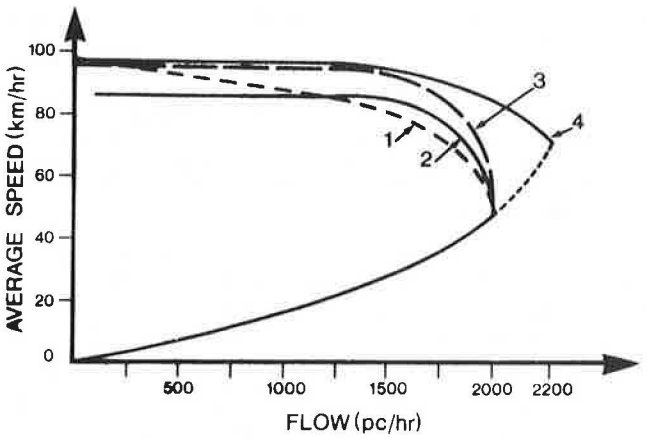


FIGURE 1 Speed-flow relationships from various studies: 1, 1965 Highway Capacity Manual *(1)*; 2, Interim Materials on Highway Capacity *(2)*; 3, Ontario 1978 *(3)*; and 4, Ontario 1983 *(4)*.

on demand and capacity characteristics and location with respect to bottleneck sections. The third issue of quantification is a relatively simple statistical one when the form of the relationship has been resolved. Unfortunately, the shape of the relationship at high flow rates has not been satisfactorily resolved, so magnitude estimation is still a problem.

Most practitioners believed that concerns expressed following publication of the 1950 HCM (5) had been satisfactorily addressed with the revised notions presented in the 1965 version. The speed-flow relationships presented in the 1965 HCM remained at the forefront of acceptance and use for many years. However, as freeway design, operations, and control expertise grew during the late 1960s and the 1970s, researchers and practitioners recognized that, for whatever reasons, the HCM representations were outdated. Despite this recognition, the results of a major 1977 literature review on flow relationships indicated that the published studies "tend to be quite detailed, but for limited sections of highway, and none attempts to generalize the basic relationships" (6,p.3). In other words, although there was dissatisfaction and a need to revise the old understanding, no one had yet provided a new interpretation. It is particularly interesting to note in that review that Roess could find only 15 references that would "probably be useful in the development of narrative and background portions" (6,p.2), and none of these were considered to be of great utility in revising existing relationships. Of the 15, one was a Greenshields' paper published in the 1930s, and five others had been published before the 1965 HCM.

Other researchers have continued to recognize the problem in work published in the 1980s, as exemplified by such comments as the following from Koshi et al. (7,p.403).

> This paper deals with vehicular flow characteristics especially in a congested region, and attempts to describe what really happens and why.
>
> It has been pointed out that vehicular flows oscillate in congested conditions and that there is a discrepancy in the speed-density-volume relationships between free flow and congested flow regions. The phenomena, however, seem not to have been explored thoroughly enough to understand the total picture of vehicular flow characteristics.

Even more recently, Hurdle again voiced the recurrent concern (8,p.127):

> A good understanding of the way in which speed varies with flow is an essential prerequisite to the creation and use of any level-of-service concept for freeways. Unfortunately, misinformation about this relationship abounds.

Presumably the dissatisfaction over the current level of understanding refers to the speed-flow interactions when flow rates are extremely high, or when transitions are made from upper-branch to lower-branch operation and back again, or both. It would appear that the basic nature of the problem facing those researchers lies either with the characteristics of the data that are available to them or with the analysis and interpretation procedures used, or some combination of the two. Although there have been some problems in the past, current technology permits collection of adequate data sets with little difficulty. As a result, the problem appears to lie with the choice of analytical procedures. This is normally dependent on the paradigm selected or hypotheses to be tested. For example, if it were assumed that speed and flow vary according to an HCM-type curve, data could be plotted and standard curve-fitting techniques adopted to generate the entire two-branch relationship. Many researchers have attempted to refine this approach, at least using speed and density data, by fitting curves to the uncongested and congested regimes separately (9-11), achieving some degree of explanatory success. Because there is no doubt about the existence of two regimes of traffic flow, that surely should also have a bearing on the manner in which speed-flow data are analyzed.

To illustrate what such an approach means for analysis, consider the attempts of Mahabir (12) in 1980 to fit a relationship to the rather extensive data set that is used in this study as well. Given a typical data set as shown in Figure 2, he assumed a two-regime model and used a number of regression techniques to try to determine the curve.

The distinction between the congested and uncongested states was made by arbitrarily selecting a speed at which he felt capacity operation occurred. It appears to be just as reasonable to fit the curve by eye, as in fact Mahabir ended up doing, because either way prejudgment of parameters, even on the already-assumed curve, must be made. Hurdle and Datta (8) also discuss the problem of curve fitting at or approaching capacity, and present five curves, all compatible with their data, that could represent the speed-flow relationship in this area. Although three of the curves would be more generally accepted, two highly unusual curves fit the data just as well.

In light of all the uncertainty, how much reliance can be placed on the traditional understanding? Traffic engineers during the past 50 years have repeatedly stated that the speed-flow relationship is not fully understood. Therefore, it may be worthwhile to take a different analytic approach to the problem. The procedures adopted for this investigation represent some different approaches and are discussed following a brief description of the data set.

SPEED-FLOW DATA

As with any analytic exercise directed toward determining the "true" nature of a real-world phenomena from only a sample of data, the researcher must have confidence in that data set, or at least know its limitations. For many years, researchers were hampered by relatively small data samples with consequent low levels of confidence as to their representativeness. During the past 20 years, however, the increasing implementation of freeway surveillance and control systems and the availability of high technology portable equipment have allowed collection of vast amounts of data with relatively high reliability. The problems related to confidence have subsequently been largely removed, leaving only the question of how best to organize and analyze these large data samples.

The data used in this study, originally obtained and analyzed by Mahabir in 1980, come from the Ontario Ministry of Transportation and Communications Freeway Surveillance and Control System (13). The data were collected in 1979-1980, and at that time the system operated on a 5-km section of the Queen Elizabeth Way (QEW) between Oakville and Toronto, where morning commuter traffic created congested flow conditions on the three eastbound lanes (Figure 3). The system comprised nine mainline detector stations with induction loop pairs in each of the three directional lanes, ramp metering on five entrance ramps, and closed-circuit television (CCTV) surveillance cameras operated from a control center. The limiting capacity restriction (bottleneck) was downstream of Station 9 where heavy Highway 10 entrance ramp traffic merged with the three through lanes.

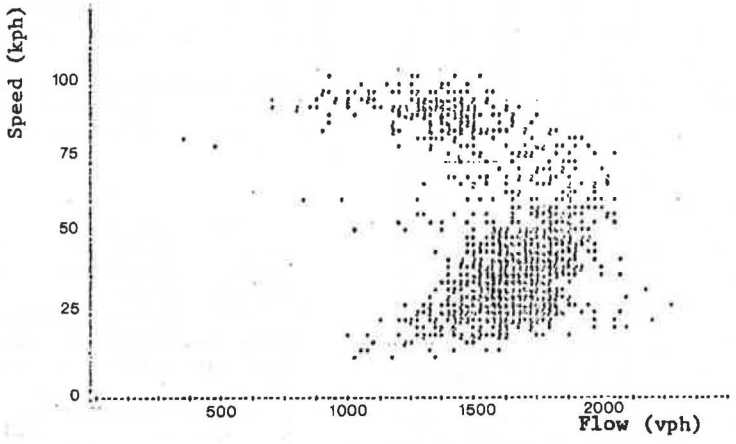


FIGURE 2 Scatterplot of data used by Mahabir *(12,* Figure A2.23, Station 9, middle lane*)*.

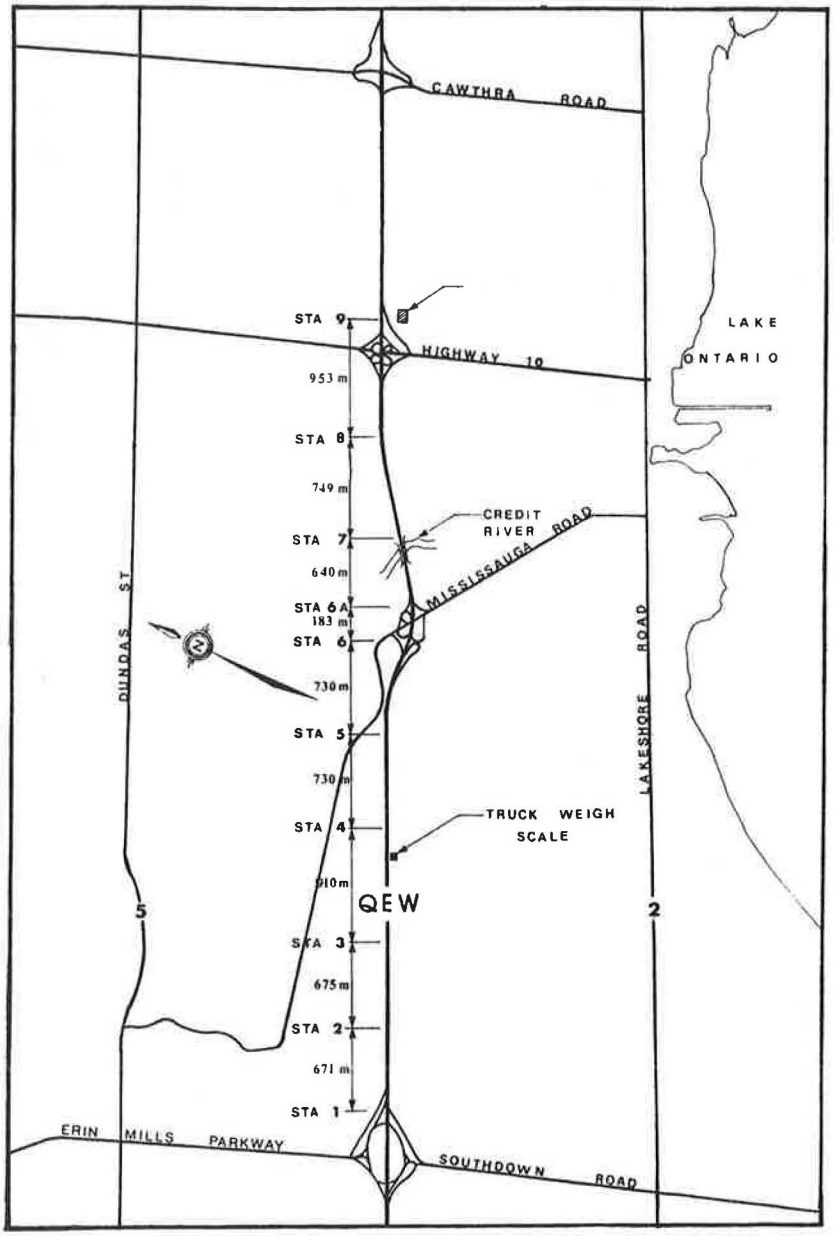


FIGURE 3 Station locations for the Freeway Surveillance and Control System, Queen Elizabeth Way, Ontario [reproduced from Case and Williams *(13)* Figure 3.4].

The data obtained from that system span 8 months from July 1979 to February 1980 and consist of approximately 2.5 hr of collection during weekday morning peak periods. From each pair of induction loops the following information was compiled for each lane for 5-min intervals: occupancy at the downstream loop, occupancy at the upstream loop, volume of vehicles longer than 7.6 m, total volume of vehicles, and average speed. The data were stored on magnetic tapes by the Data General computer that operated the system. In addition, a complete log of daily weather conditions and incidents (accidents, breakdowns, and so forth) was available.

Because the purpose of the current analysis is to identify the basic speed-flow characteristics, it was decided that only data representing "ideal" conditions would be used. Mahabir had already identified all days for which no incidents, accidents, or adverse weather had been logged; 68 such days were extracted from the 8-month period. This compilation conservatively represented only the "best" days of operation and has been used for all analyses reported in this paper.

The data were maintained and analyzed on a lane-by-lane basis because Mahabir's work suggested considerable differences in results according to lane. This distinction required that the analysis be performed in passenger car equivalents in order that the shoulder lane with 12 to 25 percent trucks, middle lane with 2 to 10 percent, and median lane with 0 to 2 percent could be easily compared. A passenger car equivalency of 2 was used, as recommended in TRB Circular 212 (2).

ANALYSIS

The appropriate analytical procedures to use with data such as these are not obvious. Mahabir's scat-

terplots (Figure 2) provide a useful picture of the range of likely and possible values of speed-flow points, but they do not provide an understanding of what actually occurs on the roadway. The present analysis, then, began with his results within the context of current speed-flow ideas and attempted to sort out the data to best describe freeway speed-flow behavior. For the most part, graphic techniques instead of statistical tests have been used. The reason for this is that statistical curve fitting is appropriate only when there is some theory to suggest the type of curve to use. The present question is simpler: what is the general pattern underlying the data?

Two types of analysis are reported here. The two procedures are presented in the order they were tried because the first analysis provided some insights and further questions that led to the second method of analysis. Other approaches were also attempted but do not merit discussion because the results added relatively little to understanding the pattern. (Those approaches were plotting speed versus time of day, and flow rates versus time of day for several days; and plotting the 68-day average speed and flow versus time.) In this discussion the focus will be mainly on Station 9 because it is situated closest to the capacity restriction and, therefore, experiences the longest period of congestion. However, there are few differences in the general trends between stations, and any of significance will be noted.

Time-Connected Plots of Mean Speeds Versus Mean Flows

The first analysis considered how speeds and flows are related over time. Such an approach has been used before to assist in resolving the problems of

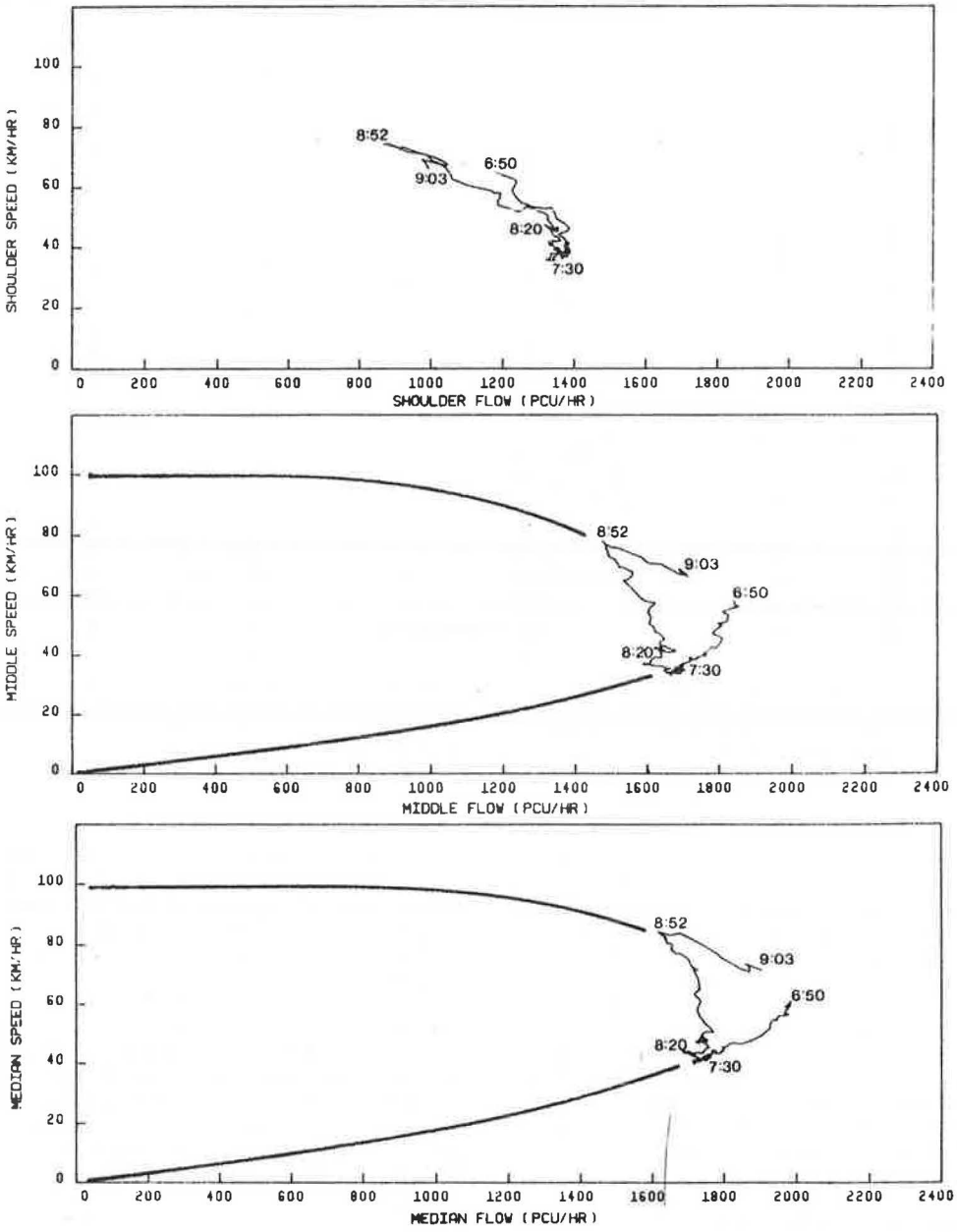


FIGURE 4 Station 9 mean speeds versus mean flow rates (over 68 days) connected by clock time; heavy lines suggest an HCM-type curve consistent with the data.

distinguishing between congested and uncongested speed-flow pairs, but only with daily data not averages over many days (14). Mean speeds and flows are plotted in the time-connected diagram of Figure 4. There are several interesting points to examine in the median and middle lane plots, which are quite similar in nature, particularly in relation to conventional interpretation. First, these data can be interpreted to be consistent with the standard HCM-type speed-flow representation as suggested by the heavy solid lines that have been positioned to fit these data. Second, the speed decrease takes a long time to occur (6:50 a.m. to 7:30 a.m.). If that decrease takes place along the lower branch of the suggested relationship, the duration is not too surprising. However, if operation at 6:50 a.m. was in the uncongested state, a more sudden drop in speeds at approximately the same flow rate would be expected. Because the speeds start at about 60 km/hr, acceptance of the lower-branch operation argument would seem reasonable. Third, regardless of the manner in which speeds and flows decreased, the relatively steady operation between 7:30 a.m. and 8:20 a.m. is to be expected. Queues and the resultant storage requirements imposed by the downstream bottleneck remain relatively constant during the congested period and appear to fluctuate less than 10 percent in speed and flow. Fourth, the speed increase between 8:20 a.m. and 8:52 a.m. appears reasonable in relation to the sketched curves, but detailed examination raises some questions.

The first question relates to duration. It is clear that the increase can be related to a transition from congested operation on the lower branch to uncongested operation on the upper branch. If so, why does the recovery take at least 0.5 hr? Traffic flow theory suggests that in steady-state conditions such recoveries take place over extremely short time periods. There is a strong possibility that the unexpected duration arises because of the averaging technique used in the analysis. That possibility will be examined in the next subsection.

The second question about the speed increase relates to the flow rates at the start (8:20 a.m.) and end (8:52 a.m.) of it. Figure 4 suggests that flow rates after the transition are lower than those at the start of transition. One line of reasoning from steady-state conditions (i.e., no demand changes upstream or downstream of the data acquisition location) suggests that recovery takes place at almost identical flow rates in the congested and uncongested regions and that, therefore, the result in Figure 4 is a consequence of the averaging. A second line of reasoning suggests that these are not steady-state conditions and that the queue clears because of a demand decrease on the main line (QEW). Which of these two is correct is examined further in the next subsection.

In the shoulder lane, the plot is quite different. Where the other two lanes show speed and flow decreasing from 6:50 a.m. to 7:30 a.m., the shoulder lane has flow increasing as speeds decrease. The heavy line in Figure 5(A) suggests that operations were close to capacity and were not forced downward along the lower branch [i.e., capacity is about 1,400 passenger car units (pcu) per hour]. However, it appears more reasonable that there is some operation on both the upper and lower branches, with the necessary transitions between them. Consequently, there could be a higher capacity, approximately 1,700 pcu/hr, as shown in Figure 5(B). If this is the case, why again do the speed changes take so

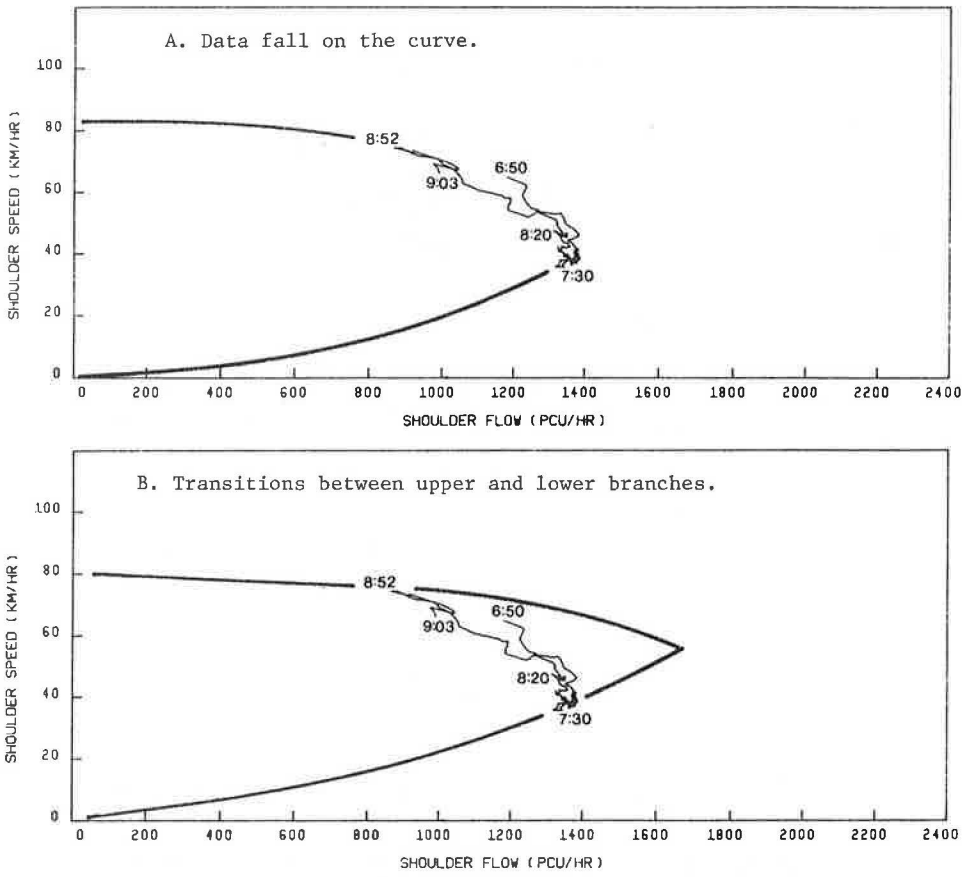


FIGURE 5 Representation of a speed-flow curve to fit shoulder lane data for Station 9.

long, with the end points at quite different flow rates? Particularly, during the supposed transition, why does the speed decrease with an increase in volume?

One possible answer is the same as suggested previously: the averaging technique may bias the view. Another possibility is that the station location strongly affects the result because it is immediately upstream of an entrance ramp feeding the bottleneck. To test this possibility, consider the plots in Figure 6 for Station 4, which is considerably upstream from any entrance ramps and at least 4 km from the bottleneck. In this instance, the operations on the shoulder lane, although at lower volumes, are similar in nature to those of the other two lanes with transitions to and from what appears to be definite lower-branch operation. It would, therefore, appear that the usual steady-state demand and shock wave notions are inadequate to describe the unusual speed-flow results for the shoulder lane at Station 9, which may be strongly related to site-specific driver behavior characteristics.

An important result of viewing the plots in Figures 4 and 6 is further confirmation of the approach of examining operations on a lane-by-lane basis instead of averaging across all lanes. When attempting to identify fundamental operating characteristics it is obviously important to avoid potential confusion and consider the lanes separately. In addition, it would appear equally important to examine more than one location, as indicated by the rather dramatic differences between shoulder lane operations shown in the two figures.

Event-Based Analysis

One possible explanation for the unexpected results in Figures 4 and 6 is that they arise because of the

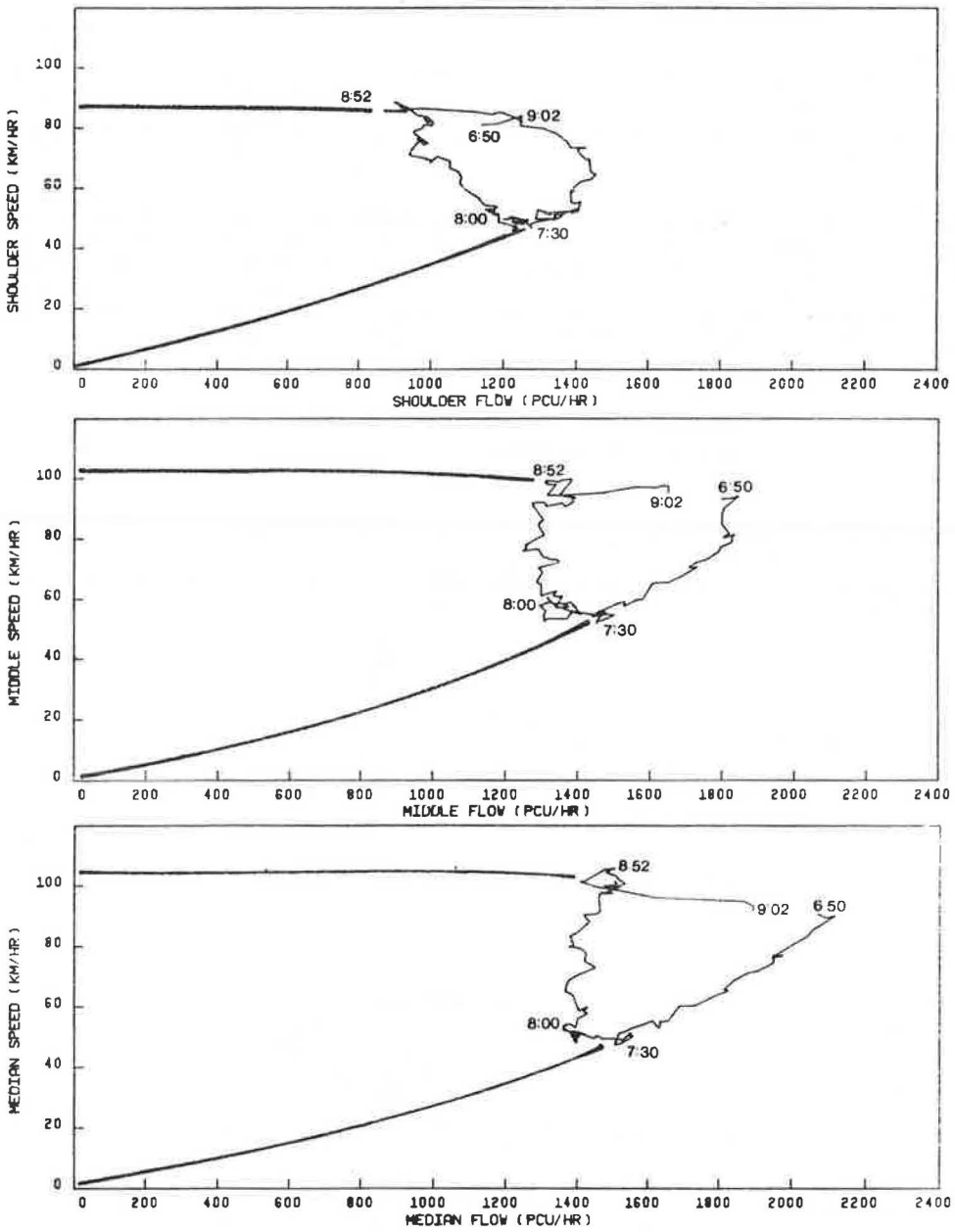


FIGURE 6 Station 4 mean speeds versus mean flow rates (over 68 days) connected by clock time; heavy lines suggest HCM-type curve consistent with the data.

use of moving average speeds and flows. In addition, averaging over clock time is not likely to mean averaging over identical traffic events. For example, congested operation will probably commence at slightly different times on different days. However, if the occurrence of a transition and average speeds and flows before and after that change in operation could be identified, without reference to clock time, a different and more representative picture might emerge. The procedures used to do that and the results for the transitions to and from congested (lower-branch) operation are described in this section.

The basis for identifying the change in conditions was the change in speed between successive intervals, because theoretically there will be a sudden decrease in average speed representing a move from the upper to the lower branch of the speed-flow curve, or the reverse. A change of 15 km/hr was selected on the basis that it was larger than most of the random fluctuations. In the case of the start of congestion, the first occurrence of a drop of 15 km/hr or more between successive 5-min average speeds (over the time period 6:50 a.m. to 8:00 a.m.) was taken to define the transition to congested flow. For termination of congestion, the last increase in speeds of more than 15 km/hr (between 7:30 a.m. and 9:10 a.m.) defined the recovery transition. This test was run for each day's data separately. Then the data were averaged across days on the basis of 5-min intervals before or after the identified speed shift. The results of these event-based calculations are shown in Figure 7 for Station 9 and in Figure 8 for Station 4. Both the increase and the

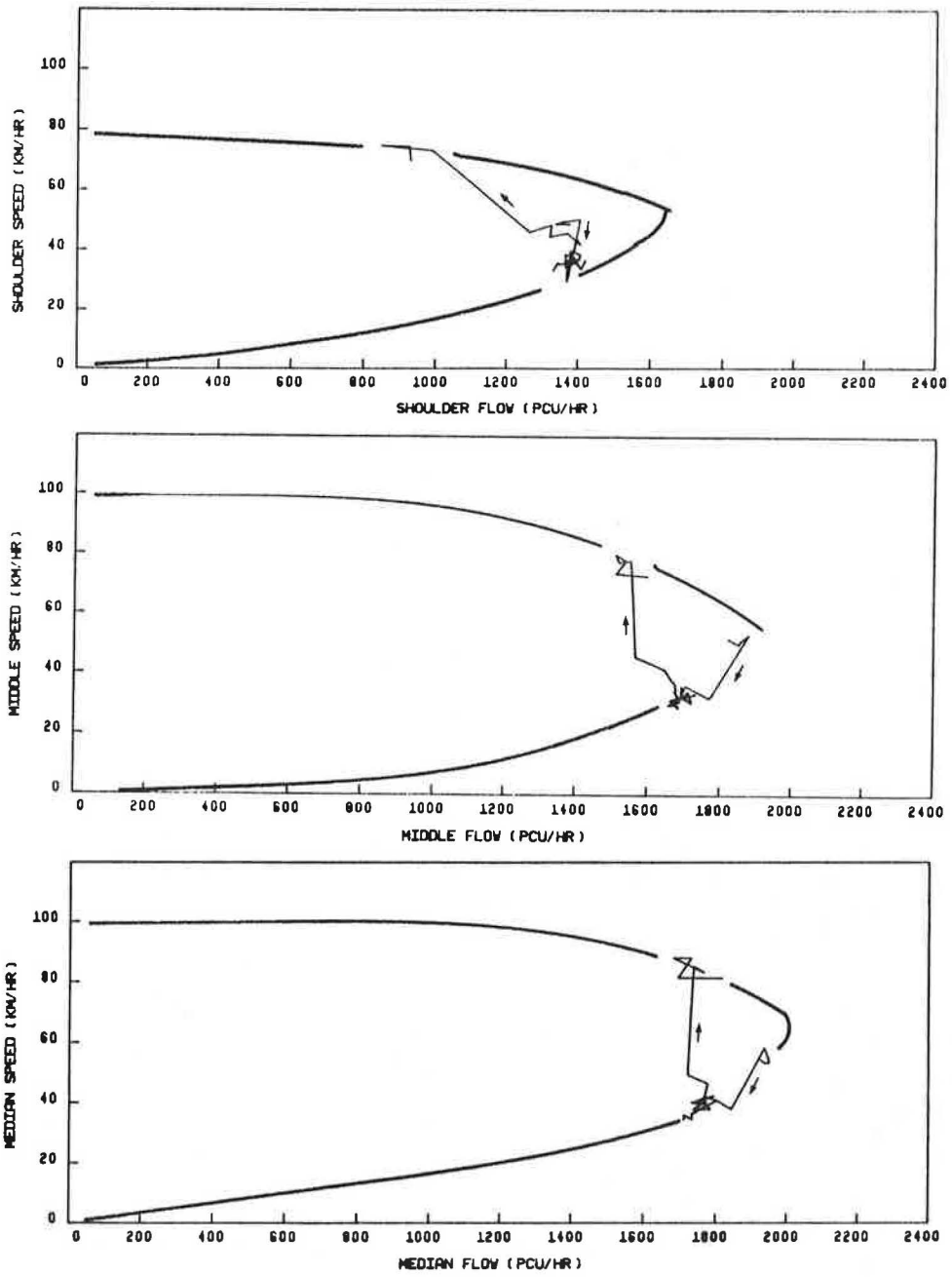


FIGURE 7 Station 9 event-based analysis of mean speeds and flows before and after transitions between congested and uncongested flows.

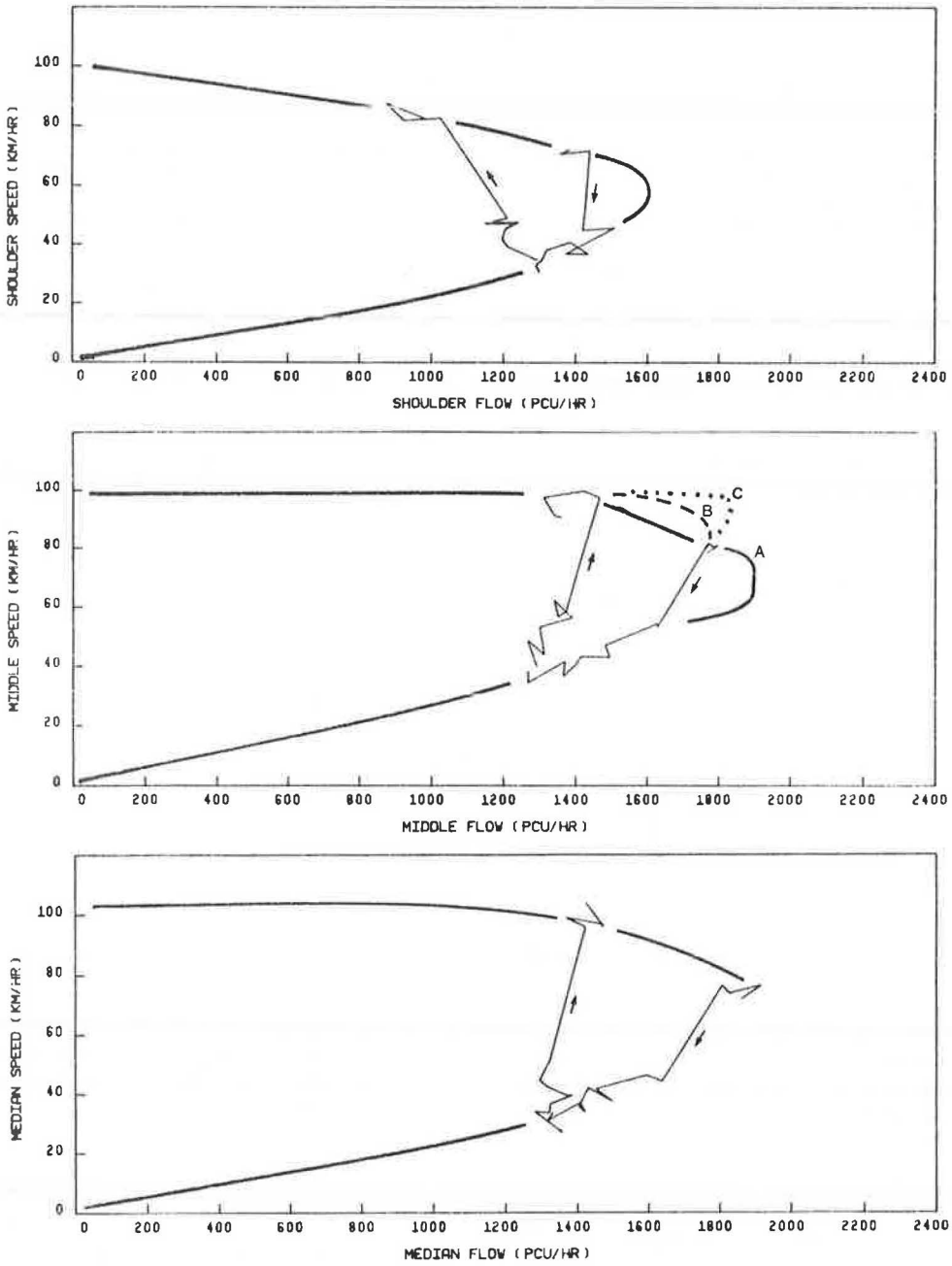


FIGURE 8 Station 4 event-based analysis of mean speeds and flows before and after transitions between congested and uncongested flow; for the middle lane, Line A represents HCM-type operation, Line B shows the simplest curve to put the data all on the lower branch, and Line C is suggested by Figure 6.

decrease are shown on the same figure to facilitate discussion, but each was calculated independently of the other.

These figures appear to resolve two of the unexpected results of Figures 4 and 6. First, both the increase and the decrease in speeds occur extremely rapidly. Within one 5-min interval, speeds have increased by nearly 40 km/hr (middle and median lanes, Figure 8), and in all cases by at least 25 km/hr. The speed decreases are not so large at Station 9, averaging only 20 km/hr, but the drop is still much quicker than was suggested by the time-traces in Figures 4 and 6. Thus conventional wisdom, which has the changes in condition occurring rapidly, appears to be supported.

Second, the increases in speed now appear to occur at roughly the same flow rate (shown most clearly in Figure 7, middle and median lanes) instead of with decreasing flow rates as shown in Figures 4 and 6. This is not so clearly the case at Station 4, or in the shoulder lanes, where the return to the upper branch is less direct. The results for Station 4 show flow rates increasing during the transition (for the middle and median lanes). Such a result suggests either that this analytical approach has not yet resolved all the difficulties or that the timing of main-line demand does not permit a consistent recovery in operations. To check the validity of the result shown in Figure 7, individual plots of daily time-connected speed-flow curves were

examined to determine what kinds of recoveries actually occurred. In general (in the middle and median lanes), the large increase in speed occurred at roughly similar volumes, with some having slightly greater and some slightly lower volumes after the shift. The vertical lines in Figure 7, therefore, appear to be good representations of the average tendency.

Thus this event-based analysis appears to have resolved the two questions raised by the first analysis and has reinforced conventional understanding in several respects. However, one aspect of the results in Figures 7 and 8 is still unexpected because the downward shift is not located where it would be expected in two important respects: it starts from a lower speed than anticipated, and it is accompanied by a decrease in flow rates. It is also noteworthy that the transition ends at lower speeds (35-40 km/hr) than are suggested by the speed-flow curves of Figures 4 and 6 (45 km/hr). This last point confirms the potential hazards of averaging with respect to clock time, which in Figures 4 and 6 probably mixes congested and uncongested flow rates between 6:50 a.m. and 7:20 a.m., resulting in unrepresentative averages.

Three possible explanations for these differences can be suggested. The first is that the transitions observed occur between two points that are already on the lower branch of the curve as discussed earlier. This is shown in Figure 7 for the middle and median lanes. The resultant implied capacity (1,900-2,000 pcu/hr) is within the normally accepted range, and it is plausible that this particular location experiences capacity flows at or before 7:00 a.m. each day. If this is the explanation, then there is no surprise in these results.

Unfortunately, this explanation is not plausible for the Station 4 data (Figure 8). For one thing, observation suggests that Station 4 is not experiencing capacity flows by 7:00 a.m. Equally important, for the transitions to be wholly on the lower branch, it is necessary to accept average speeds of 80 km/hr as representing congested flow operation (for the middle and median lanes), which in itself contradicts conventional wisdom. Taking the middle lane as an example, there are three possible ways to sketch a speed-flow curve, each of which creates a problem. Line A in Figure 8 represents the conventional curve but leads to a problem in interpretation: there is a significant decrease in volume accompanying the transition from upper- to lower-branch operation. Line B is close to the conventional wisdom, but an attempt to put the start of the transition on the lower branch results in an abnormally low capacity for the section. Such low capacity appears to be unlikely. Line C is suggested by Figure 6, which shows quite high speeds at high volumes (which logically cannot be a false result of the averaging procedure). The problem with this representation is that the speeds at the start of the transition (85 km/hr) are then considerably lower than would be expected (from Figure 8, 100 km/hr).

These difficulties lead to the third explanation: this event-based averaging is still averaging dissimilar events and, therefore, producing misleading results. For example, on one day, flow just before the transition event may be 1,800 pcu/hr, and on another it may be only 1,600 pcu/hr. Close inspection of daily data may be the only way to resolve the problems. For now, there are three reasonable explanations for the observed results, none of which is fully satisfactory.

CONCLUSIONS

Three types of conclusions can be drawn from this work. The first relates to analytical procedures, the second to speed-flow relationships, and the third to further work.

The analytical procedures used here differ in four important respects from those of most previous work. First, the analysis has been conducted on a lane-by-lane basis instead of for the freeway as a whole. Mahabir's work first showed the importance of such an approach; it is believed that the current analysis provides further support for this approach.

Second, the analysis has not proceeded by the normal curve-fitting approaches, such as regression analysis. These are the obvious approaches to interpreting a large data set, but such curve fitting is inappropriate in the present context for two reasons. Fluctuations in the data make it impossible to tell from normal scatterplots what is upper-branch and what is lower-branch operation, with the result that the curve may well be fitted to inappropriate data. Also, if the focus of the analysis is on changes in operation (i.e., transitions from one branch to another) the "lines" of interest are not those represented by the average curve.

The third difference in the analytical procedures is that an event-based averaging procedure has been used instead of clock-time averaging in order to isolate the changes in operations. This procedure is not the final answer, however, as evidenced by the confusing results for the middle lane shown in Figure 8, but it does provide better answers than do other procedures about what actually occurs on freeways. Additional efforts are needed to develop more appropriate analytical methods for such data.

The fourth difference is the size of the data set. For an analysis of the type conducted here, data for a few days or a week would be inadequate. Examination of the daily speed-flow traces made it obvious that there is considerable fluctuation in operations both within a single day and between days, such that results based on only a few days' data should be interpreted cautiously. For normal curve-fitting approaches, a week's data of 5-min flows would appear to be sufficient, but for identifying the nature of shifts in operation, considerably more days of data are required. The size of the data set used here permits reasonable confidence in the trends that have been observed.

The second type of conclusion pertains to speed-flow relationships. Some of the present conclusions support conventional wisdom; others are simply not addressed by published results and probably warrant more investigation. There are three main conclusions in support of current understanding:

1. Both the breakdown and the recovery of speeds occur fairly rapidly, with major changes occurring during a 5-min interval;
2. Recovery from lower-branch to upper-branch operation appears to take place at an approximately constant volume; and
3. Different speed-flow relationships are needed to describe operations at different points along a highway (e.g., those sketched in for Stations 9 and 4 in Figures 4 and 6).

Two main observations have not been addressed, at least in published versions of the conventional understanding:

1. The location of the line representing transitions from upper- to lower-branch operation is not easily explained and
2. There are distinctly different speed-flow relationships on the different lanes, even well upstream of any entrance ramp (e.g., Station 4).

The third type of conclusion focuses on future work. These results come from only one section of

highway, with a particular pattern of demand over time on the mainstream and on the entrance ramps, and with a pattern already affected by the freeway control system. Some of the results can be explained plausibly on the basis of this specific demand pattern. Work in other places is necessary to see if those explanations are sound. Alternatively, more detail on flow rates entering the section could be used to test the proposed explanations. In particular, the following points require investigation:

1. The recovery in speeds takes place at a lower flow rate than that at which the drop occurred. It is clear that demand flow rates are not constant over the hour for either the mainstream or the entrance ramps: a steady-state analysis is inappropriate. If flow rates on the entrance ramp downstream of Station 9 increase after Station 9 has begun lower-branch operation, then the observed flows at Station 9 must decrease. If subsequently mainstream demand decreases (because of smart drivers who time their trips well, for example), then the recovery to upper-branch operation will be from these reduced flows. If this is a plausible explanation, there may well be other expressway systems for which recovery flows are equal to or greater than flows at the onset of congestion.
2. The HCM-type speed-flow curves sketched in on many of the figures have been shaped to fit the present results, consistent with current theory, and in particular with current ideas of plausible values for capacity flow. If capacity were much higher (such as 2,400 or 2,500 pcu/hr), some of the interpretation would change. There have been recent analyses that suggest that capacity is at least 2,200 pcu/hr (4,12), but those analyses were based on conventional curve-fitting approaches and may well estimate capacity incorrectly. More work needs to be done using different analytical techniques to identify capacity flow rates.
3. Density (or lane occupancy) data have been ignored in this analysis in order to maintain a consistent discussion. These data also need to be incorporated using the types of techniques tried here instead of conventional curve fitting.
4. The present analysis has looked at each station in isolation, but clearly these data come from a system in which observed main-line flow rates are a function of downstream ramp volume (which is determined by the freeway control system) as well as of main-line demand. A complete understanding of the system requires information on how each of these volumes varies. Such an analysis would also help to make clear to what extent these results are or are not representative of freeway operation generally.

Overall, this analysis has contributed several useful points to understanding speed-flow relationships on freeways as they move to and from congested conditions. Further work is definitely needed to clarify several points. What appeared at first to be an abundance of data is in the end inadequate to resolve all the questions raised. Some insights have been gained. More are needed if the dissatisfactions with current depictions of freeway operating characteristics are to be resolved.

ACKNOWLEDGMENTS

The analysis described in this paper would not have been possible without the extensive efforts of Geddes Mahabir in making the data set manageable on the McMaster computer facilities several years ago. The assistance of the Ontario Ministry of Transportation and Communications in providing the data and their continued interest are also appreciated. Financial support came from the Natural Sciences and Engineering Research Council, through their Undergraduate Summer Research Award program, and from McMaster University. The authors would also like to acknowledge the helpful comments of two of the reviewers.

REFERENCES

1. Highway Capacity Manual 1965. Special Report 87, HRB, National Research Council, Washington, D.C., 1965, 411 pp.
2. R.P. Roess, E. Linzer, W.R. McShane, and L.J. Pignataro. Freeway Capacity Procedures. In Transportation Research Circular 212: Interim Materials on Highway Capacity, TRB, National Research Council, Washington, D.C., 1980, pp. 151-266.
3. B.L. Allen and S.M. Easa. Toward Improved Capacity and Level of Service Procedures for Ontario Highways. Final Report 2312-7814. Traffic Research Group, McMaster University, Hamilton, Ontario, Canada, 1978.
4. B.L. Allen and G.P. Mahabir. Speed/Flow and Capacity Relationships on Multilane Freeways--A Summary Report. Summary Report 2312-8311. Traffic Research Group, McMaster University, Hamilton, Ontario, Canada, Feb. 1983.
5. Highway Capacity Manual. Bureau of Public Roads, U.S. Department of Commerce, 1950.
6. R.P. Roess. Freeway Capacity Analysis Procedures: Literature Review and Annotated Bibliography. Office of Research and Development, FHWA, U.S. Department of Transportation, 1977.
7. M. Koshi et al. Some Findings and an Overview on Vehicular Flow Characteristics. Proc., 8th International Symposium on Transportation and Traffic Theory, 1981, V.F. Hurdle, E. Hauer, and G.N. Steuart eds., University of Toronto Press, Toronto, Ontario, Canada, 1983.
8. V.F. Hurdle and P.K. Datta. Speeds and Flows on an Urban Freeway: Some Measures and a Hypothesis. In Transportation Research Record 905, TRB, National Research Council, Washington, D.C., 1983, pp. 127-137.
9. S.M. Easa and A.D. May. Generalized Procedure for Estimating Single- and Two-Regime Traffic Flow Models. In Transportation Research Record 772, TRB, National Research Council, Washington, D.C., 1980, pp. 24-37.
10. A. Cedar. The Accuracy of Traffic Flow Models: A Review and Preliminary Investigation. Traffic Engineering and Control, Vol. 19, No. 12, Dec. 1978.
11. S.M. Easa. Selecting Two-Regime Traffic-Flow Models. In Transportation Research Record 869, TRB, National Research Council, Washington, D.C., 1982, pp. 25-36.
12. G.P. Mahabir. Speed, Flow, and Capacity Relations on Multilane Highways. M.Eng. thesis. McMaster University, Hamilton, Ontario, Canada, 1981.
13. E.R. Case and K.M. Williams. Queen Elizabeth Way Freeway Surveillance and Control System Demonstration Project. In Transportation Research Record 682, TRB, National Research Council, Washington, D.C., 1978, pp. 84-93.
14. D.P. Ryan and S.M. Breuning. Some Fundamental Relationships of Traffic Flow on A Freeway. Bull. 324. HRB, National Research Council, Washington, D.C., 1962, pp. 73-84.

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

Investigation of the Effect of Change in Vehicular Characteristics on Highway Capacity and Level of Service

A. ESSAM RADWAN and SYLVESTER A. F. KALEVELA

ABSTRACT

Statistical analyses were conducted on a set of traffic flow data to investigate the effect of change in vehicular characteristics on maximum volumes for various levels of service. The results obtained from analysis of traffic flow models and time headways showed that despite the change in vehicular characteristics there has not been a discernible corresponding change in highway capacity parameters.

Capacity and level of service are major objectives in traffic analysis and design of highway elements. Under given traffic conditions, a design engineer is concerned with the production of a design proposal that meets certain demand criteria including capacity and the level of service at which the facility in question will, or is intended to, operate.

The procedures for analysis and design are, in general, facilitated by the use of tables and graphs from manuals that have been developed with reference to certain base conditions. What is often needed is a set of inputs associated with the specifics of a particular analysis or design problem. The specifications normally include, among other things, traffic characteristics and demand conditions.

One of the base conditions used in discussing vehicular traffic is the passenger car unit (pcu). For many years now it has been possible to deal with problems of traffic analysis and design with reasonable efficiency and accuracy. However, in recent years changes in vehicular characteristics have been observed. These changes involve both physical and operating characteristics. The average size of today's passenger car, for instance, is smaller than that of the average car 20 years ago. In addition, technology and scientific development have provided today's car with superiority in operational qualities, such as acceleration capabilities at low speeds, and in safety provisions.

The question that is being asked in connection with the changes in passenger car dimensions and operating characteristics is how these changes affect the traffic stream characteristics with regard to

- Mean speeds at free flow,
- Jam density,
- Highway capacity under ideal conditions, and
- Maximum possible volumes under conditions specified for various levels of service.

Literature Review

Previous research developed speed-density models and flow-density models; some of these models were single-regime and others were multiregime models. Tests have shown that nearly all single-regime traffic stream models may be accurately fitted with field data only within limited ranges of flow levels(1-3; 4,pp.175-216).

Huber (5) and Haynes (6) employed a second-degree polynomial to fit volume-density traffic data. A field study carried out by Drake et al. (7) was aimed at comparing important parameters of the traffic stream models as a basis for determining their relative predicting capabilities. In this study traffic flow data were collected on the Eisenhower Expressway in Chicago and the tests were performed by regressing mean speeds versus their associated mean densities. Seven alternative speed-density hypotheses suggested by this study and their corresponding equations are given in Table 1.

Accordingly, except for the two-regime linear hypothesis, it was concluded that from the standpoint of application all models performed reasonably well and there was no basis for discontinuing the use of any. In respect to logical theoretical consistency, Drake et al. thought that the Edie hypothesis excelled in comparison with the three-regime linear form.

It was found in some studies (2) that, generally, drivers use the criterion of potential time to collision point, with average minimum headway a constant, regardless of speed. Minimum headways vary from 0.5 to 2.0 sec depending on the driver and traffic conditions, with an average of about 1.5 sec. But the value of 1.5 sec corresponds to a rate of flow per lane of 2,400 passenger cars per hour that, according to the Highway Capacity Manual (2), occurs under extremely specialized conditions.

The maximum rate of flow for a given highway is at the point of critical density, and the critical density depends on the minimum headways that drivers find tolerable at particular speeds. Generally, the higher the design standards of a highway, the shorter these headways may be.

OBJECTIVES

Speed, density, and flow and their relationships are the major subjects considered in the analyses of traffic stream characteristics. Because capacity and level of service are defined by these variables, an attempt has been made to find how changes in passenger car characteristics have affected speeds and densities and their relationship in the traffic

TABLE 1 Summary of Results Obtained by Drake *(7)*

Hypothesis	Equation	r^2	F	U_f	K_j	K_m	U_m	Q_m
Greenshields	$U = 58.6 - 0.468K$	0.896	1,005	58.6	125	62.5	29.3	1,830
2-Regime linear	$U = 60.9 - 0.515K$ $(K \leq 65)$	0.685	250	60.9	151	59.2	30.4	1,800
	$U = 40 - 0.265K$ $(K \geq 65)$							
3-Regime linear	$U = 50 - 0.098K$ $(K \leq 40)$	0.590		50.0	151	44.6	40.7	1,815
	$U = 81.4 - 0.913K$ $(40 \leq K \leq 65)$		167					
	$U = 40.0 - 0.265K$ $(K \geq 65)$							
Modified Greenberg	$U = 48.0$ $(K < 35)$	0.866	745	48.0	146	53.7	32.8	1,760
	$U = 32.8 \ln(145.5/K)$ $(K \geq 35)$							
Underwood	$U = 76.8e^{-K/56.9}$	0.901	1,050	76.8		56.9	28.3	1,610
Edie	$U = 54.9e^{-K/163.9}$ $(K \leq 50)$	0.681	245	54.9	162	50.0	40.5	2,025
	$U = 26.8 \ln(162.5/K)$ $(K \geq 50)$							
Bell curve	$U = 48.6e^{-0.00013K^2}$	0.884	872	48.6		62.0	29.5	1,830

stream and how these changes will affect capacity and level of service. To this end, the purpose of this research is

- To investigate the relationship between speed, density, and flow under today's traffic characteristics and composition;
- To compare the results of this investigation with those represented by the existing models of the traffic stream; and
- To reevaluate maximum passenger car volumes possible under operating conditions applicable to specified levels of service.

DATA SOURCE

The traffic data initially obtained for this study were collected from two selected sites: Kingery Expressway in Chicago, Illinois, and LaPorte Freeway in Houston, Texas. Both sites are three-lane basic freeway segments. This set of data is part of the field data collected in 1980, using the Federal Highway Traffic Evaluator System (TES), in order to determine passenger car equivalents (pce's) of trucks and other vehicles on urban freeways (8,9).

SITE DESCRIPTION

The LaPorte Freeway section, the focus of the study, is a 1,500-ft tangent section. The tangent is part of a ramp-bounded section of a roadway, about 2,600 ft long measured from the end of the on-ramp taper to the beginning of the exit-ramp taper.

The details of the experimental setting at the sites are beyond the scope of this paper, but the general site layout includes a system of tape-switches that are affixed to the road surface. The passage of a tire over a tapeswitch activates a high-precision event recorder.

Tapeswitches are configured within travel lanes to form traps. The strategic location of these traps and the subsequent record processing of data via TES software make possible reconstruction of all vehicle trajectories throughout the deployment area (9).

DATA EXTRACTION

To perform analyses for traffic stream models, it was deemed necessary to first obtain records of vehicular flow, speed, and density. But the data base, in the tape files from the pce study, contains records stored in a form that required sorting out to obtain the desired data. For this reason seven record items were extracted by a computer program that was tailor written for this purpose. The items are

- Lane of entry,
- Vehicle type,
- Mean speed across the deployment traps,
- Mean time headway (calculated by the program from individual headway from each trap) across the deployment area,
- Data quality flag,
- Data reasonableness flag, and
- Time of entry.

DATA PROCESSING

The term processing is employed here to mean the mathematical treatment applied to prepare the extracted data for statistical analysis. In this case processing involved

- Calculation of mean speeds for specified speed ranges. The speed ranges from which the mean speeds were calculated are 0-10, 10-20, 20-30, 30-40, 40-50, 50-60, 60-70, 70-80, 80-90, 90-100, 100-110, 110-120, and 120-130 feet per second.
- Conversion of the mean speeds from feet per second to miles per hour.
- Determination of mean rates of flow corresponding to the mean speeds. At each trap, flow was computed from time headways for each vehicle pair. The mean rate of flow for any mean speed was then obtained by the arithmetic average of all values of rate of flow within the speed range from which the pertinent mean speed was calculated. The speed distribution within each speed class was observed to be normal, and use of the mean speed is considered acceptable.
- Calculation of vehicular densities corresponding to the flow rates and speeds. Because it is known that rate of flow (Q) is equal to the product of density (K) and speed (U), ($Q = KU$), the values of density were calculated as $K = Q/U$ for all values of U greater than zero.

In connection with the data processing, a condition of data reasonableness was imposed so that only

reasonably doubt-free data were processed. This was achieved by selecting only those records for which the data reasonableness flag was zero, which is, according to the data base, the indicator for records free of all discernible indicators of doubt about data reasonableness.

PILOT ANALYSIS

In all there were seven tape files of which five contained data collected on LaPorte Freeway in Houston and two contained data obtained on Kingery Expressway in Chicago.

To establish the characteristics of these data a pilot analysis was carried out wherein the scatter of density in each file was observed. Density was picked as a suitable variable for preliminary analysis because its distribution can be used to predict performance of the traffic stream models and to estimate operational parameters of traffic flow. Previous researchers (10-12) have shown that speed-density and volume-density relations are quite characteristic of density zones--low, medium, and high density.

STATISTICS OF PILOT ANALYSIS

Both files of data from Kingery Expressway and two files of data from LaPorte Freeway are characterized by densities of less than 10 vehicles per mile. The remaining three files of data from LaPorte Freeway have density figures ranging between 4 and 144 vehicles per mile.

Because of the observed density characteristics of the sets of data, it was decided that all data from Kingery Expressway and two files of LaPorte Freeway data could not be of much practical use to this research, and those files were thereupon dropped from further analysis.

TRAFFIC CLASSIFICATION AND COMPOSITION

The typological classification of data includes 10 categories of vehicles given in Table 2. For the purpose of this research, Category 4, motorcycles, was excluded and its effect was neglected in order to simplify the work of analysis and, above all, because its overall proportion in the traffic stream was less than 1 percent (8,9). The three categories of automobiles (small, medium, and large) were collapsed into one class of automobiles and all categories (except Category 4) were combined to form a second class of mixed traffic.

TABLE 2 Vehicle Topology *(9)*

Vehicle Category	Vehicle Description	Wheelbase Length (ft)
1	Small automobile	6.0-8.8
2	Medium automobile	8.8-9.5
3	Large automobile	9.5-10.3
4	Motorcycle	3.5-6.0
5	Pickup/van/utility vehicle	10.3-13.0
7	2-axle truck	13.0-20.0
8	3-axle truck	Less than 25.0
9	Bus	Greater than 25.0
10	Combination tractor semitrailer truck	
	3 axles	Greater than 25.0
	4 or 5 axles	Any
12	Combination truck (6 or more axles)	-[a]

[a]No data available.

SIZE OF DATA BASE

By discarding all data from Kingery Expressway and in the two files from LaPorte Freeway the size of the data base was reduced from the original total of approximately 98,000 to about 49,000 vehicle counts. Nonetheless, the 49,000 vehicle counts finally used represent more than 75 percent of all vehicle counts taken at the LaPorte Freeway site. In addition, each vehicle was involved in multiple records across the deployment traps (9) so that actual data calculations were based on nearly 240,000 records. A statistical test showed no significant difference among the three files and, therefore, they were pooled into one data set (Tables 3 and 4). Table 5 gives the pooled mixed traffic classified by the three lanes.

TIME HEADWAYS

Because of the dependence of maximum rates of flow and density on minimum tolerable headways, it was considered that results of an independent statistical analysis of headways should complement the parametric results from the analysis of traffic stream models. In view of this anticipation, a sample of headway data was obtained for headway analysis as described in the paragraphs that follow.

The time headway between any two consecutive vehicles was calculated as the difference between the time of entry of the follower and the time of entry of the leader.

The source of headway data is a set of data from one file of traffic flows observed before 8:00 a.m. The choice of time was so made in order to capture the morning peak of inbound traffic flow for the analysis of headways at near-to-capacity flow. The set of data was reduced to a manageable form by classification. Nine headway classes were defined: 0-1, 1-2, 2-3, 3-4, 4-5, 5-6, 6-7, 7-8, and 8-9 sec.

TABLE 3 Speed, Density, and Volume of Automobile Data from LaPorte Freeway

Speed (mph)	Density (vpm)	Volume (vph)
12.00	125.13	1,505.60
18.13	92.74	1,669.08
24.20	73.90	1,781.60
30.84	60.29	1,857.39
36.99	48.33	1,774.49
45.36	21.14	953.07
52.09	14.41	751.37
57.88	15.19	879.66
63.61	12.73	808.52
70.09	9.85	691.02
76.96	4.29	330.23
84.13	7.62	639.39
18.90	93.00	1,760.30
24.90	60.90	1,496.40
30.40	41.70	1,267.40
37.20	42.40	1,570.00
45.10	30.80	1,383.50
51.60	26.40	1,360.50
56.90	28.50	1,614.60
63.40	24.10	1,528.10
75.70	11.90	903.10
10.80	144.00	1,480.70
18.00	88.90	1,597.00
24.10	71.50	1,720.40
29.80	58.90	1,749.80
37.30	49.10	1,811.70
45.40	22.30	1,008.30
52.10	16.30	851.40
58.00	17.00	984.00
63.90	13.70	873.50
70.30	10.30	726.50
77.10	6.90	527.80

TABLE 4 Speed, Density, and Volume of Mixed Traffic Data from LaPorte Freeway

Speed (mph)	Density (vpm)	Volume (vph)
18.20	84.70	1,531.70
24.60	63.10	1,543.70
31.10	52.60	1,631.80
37.10	43.70	1,610.60
45.20	16.70	750.70
51.90	13.50	700.10
57.80	15.00	869.20
63.60	12.60	799.60
70.10	9.30	654.30
77.00	8.50	666.50
84.10	7.60	639.40
89.40	5.50	491.10
18.90	83.80	1,566.90
24.90	59.20	1,453.70
30.60	42.20	1,288.90
36.90	41.60	1,532.40
45.90	29.90	1,343.80
51.50	28.30	1,461.00
56.80	30.50	1,723.90
11.80	91.10	1,027.60
18.10	83.10	1,496.10
24.40	66.30	1,609.70
30.20	56.60	1,709.50
37.30	44.80	1,652.80
45.20	19.70	887.40
52.00	15.40	802.10
58.00	16.80	974.90
63.90	15.10	964.40
70.30	13.70	920.80
77.10	8.50	650.90
83.30	6.80	570.40

TABLE 5 Lane Speed-Volume Data for Mixed Traffic by Lane

Median Lane		Middle Lane		Shoulder Lane	
Speed (mph)	Flow (vph)	Speed (mph)	Flow (vph)	Speed (mph)	Flow (vph)
12	1,500	12	1,503	NA	NA
18.8	1,886	18.8	1,715	19	1,763
24.1	1,829	24.1	1,623	24.8	1,638
30.2	1,540	30.8	1,637	31.0	1,735
37.9	1,386	37.0	1,802	36.4	1,692
45.2	1,052	45.1	1,332	45.4	1,515
50.8	1,030	52.1	1,080	52.4	1,076
56.4	1,046	57.3	1,018	57.6	1,175
62.9	741	63.1	908	63.7	1,187

Note: mph = miles per hour; vph = vehicles per hour; NA = not available.

Frequency tallies of these classes for four different headway types are given in Table 6. The four headway types are

- Automobile following automobile (AA),
- Automobile following truck (AT),
- Truck following automobile (TA), and
- Truck following truck (TT).

A summary of calculated values of mean headways and their variance under traffic conditions of near-to-capacity flow is given in Table 7. The rate of traffic flow during the period from which the data sample was taken was about 1,800 vehicles per hour (vph).

It can be seen that the mean time headways for vehicle pairs in which the leaders are trucks are longer than the mean headways in which automobiles lead, other factors being the same. But an analysis of variance shows that at the 95 percent level of significance the four types of headways are not significantly different from each other at the pertinent rates of flow. This observation is in agreement with recent findings by Cunagin and Chang (13) who observed that the difference among headway types diminishes with increasing rate of flow per lane-hour.

Table 6 Frequency of Headways by Type

Headway (sec)		Headway Types[a]			
Class Interval	Middle Point	1	2	3	4
0-0.99	0.5	34	14	21	10
1-1.99	1.5	122	22	28	19
2-2.99	2.5	65	25	12	9
3-3.99	3.5	12	8	6	5
4-4.99	4.5	4	2	2	3
5-5.99	5.5	2	1	2	0
6-6.99	6.5	2	1	1	0
7-7.99	7.5	0	0	0	0
8-8.99	8.5	0	0	0	0

[a] 1 = automobile following automobile, 2 = automobile following truck, 3 = truck following automobile, and 4 = truck following truck.

TABLE 7 Mean Time Headways for Headway Types 1-4

Headway Type	Mean Headway (sec)	Variance (sec^2)	No. of Observations
1	1.85	0.98	241
2	2.08	1.45	73
3	1.81	1.76	71
4	1.89	1.31	46

The overall mean value of time headway is 1.89 sec. A t-test at the 95 percent level of confidence shows that this value is significantly greater than 1.8 sec and smaller than 2.0 sec.

ANALYSIS OF TRAFFIC STREAM MODELS

Curve fitting by the method of least squares (14) was conducted on five known speed-density traffic stream models, namely, Greenshields linear, Greenberg logarithmic, Underwood exponential, Drake bell curve, and Drew parabolic model. The models were fitted on two sets of data, automobiles and mixed vehicles. Goodness of fit of each regression model was checked by the coefficient of determination (r^2) and the F-ratio test, and the comparison of parameters was done in two stages:

1. Parametric comparison of models: This involved the comparison of values of U_f, K_j, K_m, U_m, and Q_m obtained by different models analyzed with the same data [from LaPorte Freeway, collected in 1980 (see Tables 8 and 9)].
2. Parametric comparison of data: For the linear model the values of U_f, K_j, K_m, U_m, and Q_m obtained from LaPorte Freeway data were compared with values previously obtained by Drake et al. (7). A summary of the comparison is given in Table 10.

VOLUME-DENSITY MODEL

Haynes (6) investigated the volume-density relationship from data in which both flow and density were measured quantities. Despite the difference in the methods used to collect data it was thought that there was adequate basis for comparison of Haynes'

TABLE 8 Stream Equations and Parameters for Automobiles

	Greenshields	Greenberg	Underwood	Drew	Drake
Equation[a]	1	2	3	4	5
F-value	130	407	579	291	119
r^2	.812	.931	.951	.907	.798
U_f	67.9	–[b]	75.6	90.8	55.1
K_j	128.6	225	–[b]	142.8	–[b]
K_m	64.3	82.8	66.2	63.5	69.0
U_m	34.0	22.2	27.8	30.3	33.4
Q_m	2,190	1,840	1,840	1,920	2,300

[a]Equations: 1, $U = 67.92 - .53K$; 2, $U = 22.2\ln(225/K)$; 3, $U = 75.6e^{-(K/66.2)}$; 4, $U = 90.8[1 - (K/142.8)^{1/2}]$; and 5, $U = 55.1e^{-1/2(K/69)}$, where U is space mean speed in miles per hour and K is density in vehicles per mile.

[b]Model does not predict this information.

TABLE 9 Stream Equations and Parameters for Mixed Traffic

	Greenshields	Greenberg	Underwood	Drew	Drake
Equation[a]	1	2	3	4	5
F-value	154	441	472	278	204
r^2	.84	.94	.94	.91	.88
U_f	74.3	–[b]	83.0	98.7	62.6
K_j	99.1	173.5	–[b]	113.7	–[b]
K_m	49.4	63.8	52.4	50.5	60.0
U_m	37.2	24.9	30.5	32.9	23.0
Q_m	1,840	1,590	1,600	1,660	1,150

[a]Equations: 1, $U = 74.30 - .75K$; 2, $U = 24.9\ln(173.5/K)$; 3, $U = 83.0e^{-(K/52.4)}$; 4, $U = 98.7[(K/113.7)^{1/2}]$; and 5, $U = 62.6e^{-1/2(K/50)2}$

[b]Model does not predict this information.

TABLE 10 Comparison of Regression Parameters for Linear Model: LaPorte Freeway Versus Eisenhower Expressway

Parameter	LaPorte (mixed)	Eisenhower	LaPorte (automobile)
Equation	$U = 74.3 - .75K$	$U = 58.6 - .457K$	$U = 67.9 - .53K$
F-value	154	1,005	130
r^2	.841	.896	.812
U_f	74.3	58.6	67.9
K_j	99.1	125	128.6
K_m	49.4	62.5	64.3
U_m	37.2	29.3	34.0
Q_m	1,840	1,830	2,190

results and those obtained from the set of data from LaPorte Freeway. The grounds for the comparison are discussed in the next two paragraphs.

The truck proportion in Haynes' data is estimated at an average of 3 percent. In LaPorte Freeway data, trucks constitute about 9 percent. Although the percentage of trucks in the two sets of data is different, the availability of the information about their difference is sufficiently helpful as a basis for reasonable inference. The only problem is that it is difficult to make an exact inference similar to that which could be made if the truck proportions were the same in the two sets of data.

Although the lengths of the two experimental roadway sections are different, a 5-mile section in Haynes' data and 0.50 mile for the LaPorte Freeway, the equality in their number of lanes provides adequate resemblance for comparison.

In this comparison the parameters of interest were the maximum volume and the density associated with the maximum volume. Haynes obtained two volume-density equations:

$$V = 75D^2 - 0.205D - 812 \quad (1)$$

$$V = 65.5D^2 - 0.179D - 80 \quad (2)$$

where V is three-lane volume and D is three-lane density. Both equations estimated maximum three-lane volumes of about 6,000 vehicles per hour and predicted the same value of three-lane density associated with maximum volumes at 183 vehicles per mile.

The parameters obtained from a similar curvilinear regression, involving a second-degree parabola, for the LaPorte data resulted in the equation:

$$V = 43.73D^2 - 0.135D + 772 \quad (3)$$

From this relationship the value of the maximum three-lane volume is 4,860 vehicles per hour, and the value of three-lane density associated with the maximum volume is 177 vehicles per mile. These values are lower than those obtained by Haynes. But they are consistent in that

- They have been estimated from data on traffic roughly 10 percent of which is trucks, and
- The values of maximum volume and critical density estimated from this model do not differ significantly from those estimated by the speed-density hypotheses.

ANALYSIS OF AUTOMOBILE FLOW BY LANE

For each of the three freeway lanes a speed-density linear relationship was estimated for automobiles. The following parameters were observed from the equations:

1. The mean speeds at free flow decreased from the median lane to the shoulder lane. These speeds are 75, 63, and 59 mph for median, middle, and shoulder lane, respectively.
2. Jam density increased from the median lane to the shoulder lane. The values of jam density were 114, 131, and 143 vehicles per mile in the median, middle, and shoulder lane, respectively.

3. The maximum volumes for the median, middle, and shoulder lane were 2,130, 2,150 and 2,100 vehicles per hour, respectively.

Table 11 gives a summary, by lane, of the linear equations and their respective parameters. A statistical t-test, performed at the 95 percent significance level to test whether these rates are different from each other, showed that the volumes 2,130, 2,150, and 2,100 vehicles per hour were not significantly different from each other.

TABLE 11 Summary of Lane Flows of Automobiles

	Lanes		
Description	Median Lane	Middle Lane	Outer Lane
Equation	U = 74.74 - .66K	U = 62.82 - .46K	U = 58.72 - .41K
r^2	.93	.92	.90
U_f	74.74	62.82	58.72
K_j	114	137	143
K_m	57	68.5	71.5
U_m	37.4	31.4	29.4
Q_m	2,130	2,150	2,100

DISCUSSION OF RESULTS

Traffic Composition and Characteristics

The maximum traffic volumes that have been estimated by the models from the 1980 LaPorte Freeway data are, to a certain extent, comparable to those estimated by Drake's equations (Table 1) developed from the 1965 data. For the case of mixed traffic, of which 10 percent is trucks, the maximum volumes predicted from LaPorte data are practically equal to the corresponding values estimated by Drake with the Greenshields, Greenberg, and Underwood hypotheses. For automobiles, all maximum volumes estimated for the LaPorte data are higher than Drake's.

Incidentally, Drake did not explicitly report the composition of the vehicular traffic data that were used in the comparative statistical analysis of the traffic stream models. Because of the lack of information about traffic composition it, unfortunately, becomes difficult to make direct comparison between the parameters obtained by Drake and those estimated from the data collected on LaPorte Freeway in 1980. Consequently, no inference can be made with certainty regarding between-data comparison.

The results of comparison of the various models for LaPorte data exhibited a trend that is in agreement with previous observations made by Drake wherein

1. The value of U_f predicted by the Underwood curve has been observed to be considerably high,
2. The values of critical speed estimated by most of the models are much lower than the critical speeds obtainable by inspection of speed-volume data, and
3. The linear model tends to underestimate jam density but predicts high values of maximum volumes.

Capacity and Maximum Service Volumes

In both the 1965 Highway Capacity Manual (2) and the Interim Materials on Highway Capacity (3), six levels of service, A through F, are defined. For each level of service performance criteria based on threshold speeds and ceiling densities are given.

With the speed-density linear equation estimated for automobiles on LaPorte Freeway, estimates of maximum service volumes for levels of service A through E were calculated according to the criteria specified in the 1980 Interim Materials on Highway Capacity. The volumes so obtained agree with those estimated in the Interim Materials for levels of service B, C, and D. The Interim Materials give rather conservative values of maximum volumes for levels of service A and E (see Table 12).

TABLE 12 Service Volumes for Basic Freeway Segment

	Performance Criteria		Maximum Service Volume (vph)[a]		
Level of Service	Speed (mph)	Density (vpm)	1 Lane	2 Lanes	3 Lanes
A	> 50	< 15	2,700	2,400	2,400
B	> 50	< 25	4,100	3,900	3,500
C	> 48	< 35	5,180	5,100	4,800
D	> 40	< 47	6,060	5,775	5,400
E	> 30	< 67	6,510	6,000	6,000
F	< 30	> 67	-[b]	-[b]	-[b]

[a]Maximum service volumes: 1 lane, estimated from linear equation U = 67.9 - .53K (LaPorte data); 2 lanes, given in Interim Materials on Highway Capacity, 1980, *(3)*; and 3 lanes, old figures from the 1965 Highway Capacity Manual, 1965 *(2)*.

[b]Not applicable.

Time Headways at Near-to-Capacity Flow

On the basis of the overall average value of time headway, which was 1.89 sec for peak flow, it appears that the maximum uniform rate of flow that can be expected under the traffic conditions appertaining to the experimental data is around 1,900 vehicles per hour. This value is an estimate for mixed traffic.

From the mean headway estimated for headway Type 1 (automobile following automobile), the 95 percent confidence interval for maximum volumes is estimated to be 1,830 to 2,080 passenger cars per hour (pcph), with a mean of 1,955 pcph. This estimate is not unreasonable considering that it is not based on homogeneous automobile traffic. It should be noted that although the volume estimate of 1,955 pcph is based on the mean of Type 1 headways (automobile following automobile) these passenger cars experience a wave of turbulence caused by the trucks that are in the traffic stream.

PROBLEMS ENCOUNTERED IN COMPARATIVE ANALYSIS

The most stubborn problem in this research was that of lack of exact comparable conditions under which data were collected for the various previous traffic stream analyses. Some of the problems associated with the incomparability between data from LaPorte Freeway and data from earlier studies are discussed next.

It was difficult to compare model parameters estimated on the basis of data from LaPorte Freeway with those obtained by Greenberg (15), Edie (16), and Gazis (17) and data from Lincoln Tunnel because traffic flow in tunnels resembles single-lane conditions in which, unlike in three-lane flow, overtaking is not possible.

Huber (5) studied the performance of the Merritt Parkway in Norwalk, Connecticut, operating over a temporary bridge. The study conditions on the parkway were dissimilar to those of LaPorte Freeway in that

- LaPorte Freeway carried mixed traffic with 10 percent trucks;

• LaPorte Freeway study site was a three-lane section; and
• Overtaking was prohibited at the Merritt Parkway study site so traffic flow was, to a large extent, similar to that under single-lane conditions.

Studies carried out by Haynes (6), in Houston, and Drake (7), in Chicago, are the two that involved, to the author's knowledge, data with the greatest similarities to the 1980 LaPorte Freeway data. But here too there was a problem of dissimilarities in truck proportions.

SUMMARY

Due to the problems encountered because of lack of exact correspondence between the conditions associated with data from LaPorte Freeway and those associated with data previously collected elsewhere it was not possible to make numerical comparisons of the parameters obtained in this analysis and those previously obtained. However, two of the three objectives that were defined at the outset of the research were achieved. These objectives were

1. Investigation of speed-density-flow relations under current traffic conditions. The set of traffic flow data collected in 1980 was taken to represent current traffic conditions. This analysis involved estimation of traffic flow parameters obtainable by speed-density models, the predicting capabilities of which were compared in terms of results from statistical tests and inspection of flow data.
2. Evaluation of maximum possible passenger car volumes under operating conditions applicable to specified levels of service. The volumes estimated were calculated in accordance with the threshold speeds and ceiling densities specified in Interim Materials on Highway Capacity.

In addition, the analysis of time headways at near-to-capacity flow was conducted to supplement the efforts to estimate maximum possible values of capacity flow on freeway sections under traffic conditions similar to those of the analysis data.

CONCLUSIONS

On the basis of the results obtained from the analysis, the following conclusions can be drawn:

1. Among the speed-density models, the Underwood exponential curve has the highest predicting capability. The Greenberg logarithmic curve follows with the second highest predicting capability.
2. The value of jam density estimated by the Greenshields linear model is lower than those predicted by other models and than would reasonably be expected on inspection of traffic flow data.
3. The Greenshields relation is the most accurate in predicting critical speeds, and despite its tendency to underestimate jam density it out performs the other models in predicting reasonable high values of maximum volume. Its merit in estimating critical speeds and maximum volumes apparently outweighs its shortcoming in relation to jam density.
4. The values of maximum passenger car volumes obtained in the analysis of data from the LaPorte Freeway suggest that conditions affecting the characteristics of traffic flow are probably compensating so that the macro-characteristics of traffic flow ultimately remain essentially unchanged.

REFERENCES

1. D.L. Gerlough and M.J. Huber. Traffic Flow Theory, A Monograph. Special Report 165. TRB, National Research Council, Washington, D.C., 1975, Chapter 4.
2. Highway Capacity Manual 1965. Special Report 187. HRB, National Research Council, Washington, D.C., 1965, 411 pp.
3. Interim Materials on Highway Capacity. Transportation Research Circular 212. TRB, National Research Council, Washington, D.C., 1980, pp. 151-171.
4. L. Pignataro. Traffic Engineering. Prentice Hall, Englewood Cliffs, N.J., 1973.
5. M.J. Huber. Effect of Temporary Bridge on Parkway Performance. Bull. 167. HRB, National Research Council, Washington, D.C., 1957, pp. 63-64.
6. J.J Haynes. Some Considerations of Vehicular Density on Urban Freeways. In Highway Research Record 99, HRB, National Research Council, Washington, D.C., 1965, pp. 59-80.
7. J. Drake, J. Schofer, and A.D. May, Jr. A Statistical Analysis of Speed Density Hypotheses. In Vehicular Traffic Science, American Elsevier, New York, 1967, pp. 112-117.
8. E.L. Sequin, K.W. Crowley, and W.D. Zweig. Passenger Car Equivalents on Urban Freeways. Report FHWA/RD-81/156. FHWA, U.S. Department of Transportation, Aug. 1982.
9. E.L. Sequin. Urban Freeway Truck Characteristics, Data Base Documentation. Interim Report FHWA/RD-82. FHWA, U.S. Department of Transportation, 1982.
10. L.C. Edie and R.S. Foote. Experiments on Single-lane Flow in Tunnels. In Theory of Traffic Flow, American Elsevier, New York, 1961, pp. 181-185.
11. R. Herman and R.B. Potts. Single-Lane Traffic Theory Experiments. In Theory of Traffic Flow, American Elsevier, New York, 1961, pp. 120-146.
12. D.C. Gazis, R. Herman, and R.W. Rothery. Nonlinear Follow-the-Leader Models of Traffic Flow. Operations Research, Vol. 9, 1961, pp. 545-567.
13. W.D. Cunagin and E.C.-P. Chang. Effects of Trucks on Freeway Vehicle Headways Under Off-peak Flow Conditions. In Transportation Research Record 869, TRB, National Research Council, Washington, D.C., 1982, pp. 54-59.
14. R.E. Walpole and R.H. Myers. Probability and Statistics for Engineers and Scientists. Macmillian Publishing Co. Inc., New York, 1978.
15. H. Greenberg. An Analysis Traffic Flow. Operations Research, Vol. 7, No. 1, Jan.-Feb. 1959, pp. 79-85.
16. L.C. Edie. Car-following and Steady State Theory for Non-congested Traffic. Operations Research, Vol. 9, 1961, pp. 66-76.
17. D.C. Gazis, R. Herman, and R.B. Potts. Car-following Theory of Steady State Flow. Operations Research, Vol. 7, 1959, pp. 499-505.

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

Indo-Swedish Road Traffic Simulation Model: Generalized Traffic System Simulator

S. P. PALANISWAMY, G. GYNNERSTEDT, and Y. R. PHULL

ABSTRACT

A stochastic discrete-event simulation model system for heterogeneous road traffic, which prevails on the Indian highway network, is presented. The original Swedish Road Traffic Simulation Model system, designed for motorized traffic, has been generalized to cover heterogeneous traffic that includes slow-moving nonmotorized traffic. The model has also been extended for narrow roads and covers roadwidths from 3.75 to 13 m with different shoulder types and alignments in flat, rolling, and hilly terrains. Submodels for the basic desired speed, power-to-mass ratio, overtaking gap acceptance and yielding probability distributions, passing speeds, and fuel consumption rates have been calibrated with extensive field data for subsequent validation of the simulation model. The output of the model has been intended primarily (a) to furnish relevant vehicle operating data for appraisal of individual road projects, (b) to constitute background data for a relevant level of service concept, and (c) ultimately to provide a basis for an appropriate policy on geometric design of rural roads in general. The model has been programmed in SIMULA-67 language using Jackson Structured Programming concepts. The results of validation exercises have been convincing; the model should be useful to decision makers for obtaining reliable data for investment analysis.

There have been various attempts in the past to simulate rural two-lane road traffic and the most successful among them is the model developed at the Swedish National Road and Traffic Research Institute (VTI) (1-4). This is the first discrete-event simulation model with a long history of development and rigorous validation. The objective of the research reported was to adapt, modify, and extend the basic structure of the VTI model to the heterogeneous traffic prevailing on the Indian road network. The modified Swedish Road Traffic Simulation Model (INSWERTS) simulates eight different vehicle types including the slowest moving bullock carts. Roadwidths considered varied from 3.75-m-wide single lane to two lane with auxiliary lanes. Multilane highways have also been treated.

Many submodels such as the basic desired speed (BDS), power-to-mass ratio, gap acceptance probabilities, passing speed, yielding probabilities of vehicles for passing opportunities, and fuel consumption have been calibrated with extensive field data (5). These submodels along with the main simulation model have been validated for traffic on flat, rolling, and hilly terrains for different road categories. Measures used for the validation are travel time, spot speed, time headway, and number of overtakings over road stretches of about 5 km.

INDIAN HIGHWAY TRAFFIC SYSTEM

Bidirectional single-lane roads (3.5 m wide) with soft shoulders constitute the largest percentage of the rural road network in India. Traffic on these roads is subjected to significant delay due to high levels of overtaking and passing impedance. The second largest road type is known as intermediate-lane road, which is 5.5 m wide and has fewer impedances than single-lane roads. Two-lane roads (7 m wide) constitute a small percentage and traffic on these roads also experiences significant delay due to the behavior of heterogeneous traffic.

Traffic is composed of a spectrum of vehicle types that share the available road space. Vehicles differ considerably in their physical size, motive power, and control and guidance as well as performance capabilities. Passenger cars, buses, and trucks operate at higher speeds, and animal-drawn vehicles (ADVs) move at speeds of about 5 km/hr. A large number of motorcycles and scooters also operate on these roads. The interaction among these vehicles takes place in complex ways that result in congestion even under low to medium flows due to frequent bottlenecks caused by slow-moving vehicles.

Traffic Behavior on Indian Roads

On two-lane roads, vehicles in opposing streams do not interact when they pass. However, passing impedance increases as the roadwidth decreases and, in the extreme, passing vehicles are forced to stop because there is room for movement of vehicles in only one direction. Especially on narrower roads vehicles normally operate in the middle of the roadway. Drivers of slower vehicles do not yield when faster ones catch up from behind resulting in delays to the latter.

Narrow roads present a formidable impediment to the catching-up vehicles because they have to follow slower ones that have to reduce their speeds and then move partly onto the shoulder to enable the faster ones to overtake. Often slower vehicles do not yield when a faster one catches up. A queued vehicle is normally allowed to overtake after a period, which is rather random. However, if the shoulder ahead is in good condition, the probability of the slower vehicle yielding is higher. Other roadway parameters affecting this yielding probability are the roadwidth and geometric details (6).

On wider two-lane roads, slower vehicles do not move to the shoulder to enable faster ones to get by because there are overtaking opportunities for the latter (7). However, on narrow roads, the operating rules are such that slow-moving vehicles are virtually forced to use the shoulder, which results in reduced operating speed. Thus, on narrow roads, the shoulders play a significant role in determining overtaking and passing speeds.

This discussion establishes that the decision maker in overtaking situations on narrow roads is the leader not the follower, which complicates the decision logic of drivers on these roads. In addition, there are no flying or accelerative overtakings on narrow roads. The overtaking sequence is akin to that in the inner lane of multilane unidirectional traffic (6).

SIMULATION MODEL

The model has been programmed in SIMULA-67 language using Jackson Structured Programming (JSP) concepts that permit lucid organization of vehicle flow logic (8-10). The entities considered in the modified Indo-Swedish model are the roads, the vehicles, and the drivers. The following sections contain a brief explanation of how the VTI model has been successfully adapted and modified for Indian traffic.

Roads

Road sections with varying width, roughness, curves, grades, and speed limits can be simulated. The road is represented by homogeneous blocks with the same geometry and traffic regulations. Sight distance along the road stretch, presence of good shoulders as well as auxiliary lanes, overtaking restrictions, and presence of overtaking and no-overtaking zones are also included. Each block is associated with a median speed about which vehicle speeds are distributed.

Vehicle Types

The significant factors contributing to the performance of traffic are the physical size and power-to-mass ratio distributions of vehicles. Each vehicle type has been identified within this framework:

- Passenger cars, pickups, and jeeps;
- Trucks and buses [heavy motor vehicles (HMVs)];
- Farm tractors and ADVs; and
- Motorcycles and scooters.

These four vehicle groups constitute the majority of traffic on Indian roads. In addition, bicycles, pedaled cycle rickshaws, and other slower vehicles also form part of the traffic, but their effect on traffic flow has been taken as noise on the system and calibrations of submodels have been adjusted for this noise.

Basic Desired Speed and Passing Speed Distribution

The notion of basic desired speed (BDS) distribution is pivotal to the modeling of a traffic system. On ideal roadways drivers are assumed to travel at a speed restricted only by the characteristics of their vehicles; this distribution is the BDS. BDS is the starting point for modeling speed reductions caused by roadwidth and alignment. Figure 1 shows the BDS distribution for passenger cars, trucks and buses, and two wheelers.

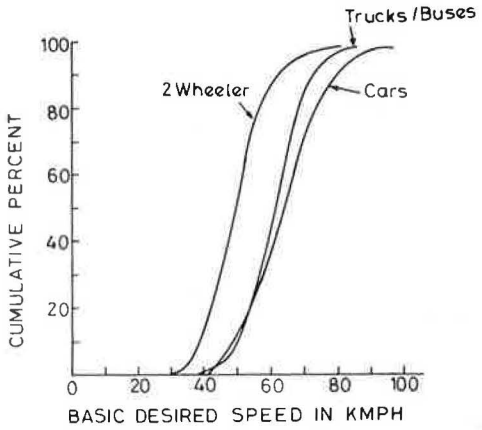


FIGURE 1 Basic desired speed distribution.

Submodels Describing the Effect of Roadwidth, Curvature, and Speed Limit on BDS

The effects of roadwidth, curvature, and speed limit on BDS have been considered in a recursive model form. The effect of roadwidth is modeled first. Then individual curves that cannot be taken at the speed limit are modeled. Grade has been taken into account in the simulation model by using the force equation.

For roads that have straight horizontal alignment and are 12 m or more wide, there is a BDS distribution with a given median speed (V_o). For roads that have straight horizontal alignment but are less than 12 m wide, V_o is reduced to V_1 where V_1 is a function of roadwidth. At this stage, V_1 is reduced by accounting for the horizontal alignment. Thus a new median speed (V_2), which is a function of roadwidth and its curvature, is obtained. Curves with a mean radius of 1000 m or less are considered. The new median speed (V_2) is adjusted for the speed limit to V_3. After the median speed V_3 has been calculated a resulting new speed distribution is calculated. This is accomplished by moving the BDS from the median value V_o to the new median V_3 and at the same time rotating it about V_3 so that the dispersion of the distribution decreases.

A transformation measure (Q-value) is used to indicate how far the BDS must be rotated about V_3. The Q-value, which is a function of the median speeds V_o, V_1, V_2, and V_3, is expressed as

$$V_o^Q - V_3^Q = V_{oi}^Q - V_{3i}^Q$$

where V_{oi} and V_{3i} are the speeds at an arbitrary percentile in the respective distributions. The distributions are thus shifted parallel to each other in the space V^Q as shown in Figure 2. For $Q = 1$ the relation implies a purely parallel shift of the two distributions and no rotation about the median value V_3. For $Q < 1$ the desired distribution is rotated counterclockwise about V_3. This rotation results in smaller variation between different percentiles compared with the BDS distribution. These rotations imply that a vehicle with a higher BDS reduces its speed more than those with a lower speed when influenced by speed-reducing factors. Furthermore, the smaller the value of Q, the larger will be the rotation (8). Individual road factors have their own speed reduction measure (q_i).

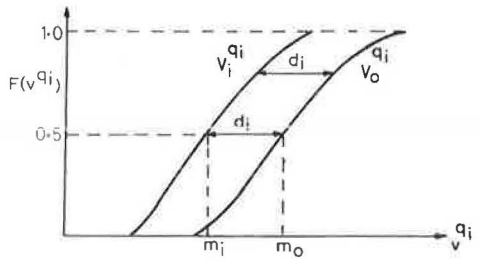


FIGURE 2 Effect of roadway factors on basic desired speed.

Model for Speed Sequence on Grades and Acceleration Sections

Consider the free-body diagram of forces acting on a vehicle on an upgrade:

$$m\,(dv/dt) = F - F_L - F_r - mg \sin i$$

where m is mass of the vehicle, v is speed, i is slope, and t is time. The tractive force at the wheel is

$$F = P/V$$

where P is the power of the vehicle in Watts measured at the wheels. The air resistance is

$$F_L = C_L\,Av2$$

where C_L is the air resistance coefficient in kilograms per cubic meter and A is the frontal exposed area in square meters. The rolling resistance is

$$F_r = m \cos i\,(C_{r1} + C_{r2} \cdot v) \approx m\,(C_{r1} + C_{r2} \cdot V)$$

where C_{r1} and C_{r2} are rolling resistance coefficients. Force due to gravity is mg sin i ≈ mgi. These expressions are substituted in the force equation and the following is obtained:

$$(dv/dt) = (p/v) - [(C_2A/m)\,v^2] - (C_{r1} + C_{r2} \cdot v) - gi$$

where p = P/m is the power-to-mass ratio of the vehicle. Each vehicle is allotted a p-value, which is the maximum power-to-mass ratio that the driver desires to or can use. Figure 3 shows the p-distribution for different vehicle types obtained from the field studies. Each time a vehicle passes a block limit, its ability to maintain its desired speed

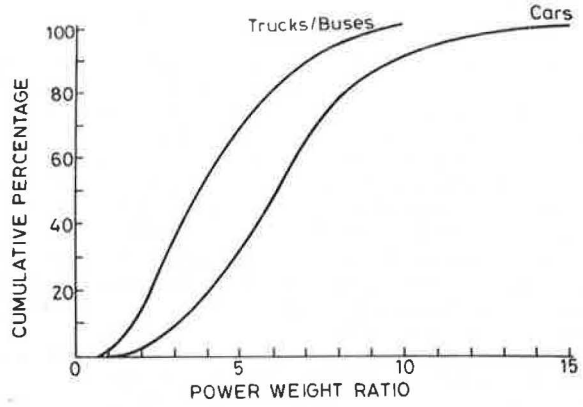


FIGURE 3 Power-to-weight ratio distribution.

(V_{3i}) is tested. If a < 0 or if the vehicle's current speed is lower than V_{3i}, the speed is calculated by solving the force equation numerically. In this way a free-running speed profile along the road is created. This procedure is used only for fast traffic. Animal-drawn vehicles and farm tractors move at constant speeds that are independent of roadwidth, curves, and gradient.

Speed Reduction for Heavy Vehicles Traveling Downhill

For faster vehicles traveling downhill, the additional gravitation force results in higher acceleration ability and hence such vehicles could attain the desired block speed (V_{3i}) quite rapidly. However, observations of vehicles on hills indicate that they descend in lower gears and thus maintain a lower rate of acceleration in reaching V_{3i}. For this reason the p-value available to them is not fully used. This is more pronounced in the case of heavy vehicles. To account for this downhill travel behavior, adjustment has been made to reduce V_{3i} (7). The block speed reduction for heavy vehicles on grades with slope ≤ -1 percent has been empirically determined to be 4.24 m/sec for p ≤ 3.3, 11.39 - 2.166p for 3.3 < p < 5.25, and 0 for p ≥ 5.25.

Model for Vehicle Interactions

The modeling of vehicle behavior that constitutes the core of the simulation model considers critically different kinds of vehicle interactions and develops concepts necessary for a dynamic sequence of vehicle movement. A free-moving vehicle interacts with the road and its speed is conditioned by roadwidth, curves, hills, speed limits, and other road conditions. However, as traffic flow increases, interactions among vehicles are to be considered as well. For example, when a faster vehicle catches up with a slower vehicle the interaction is viewed in terms of the catching-up vehicle that reduces its speed and prepares to follow if the slower one decides not to yield space by moving to the hard shoulder. On narrow roads, when two vehicles traveling in opposing streams meet, both vehicles reduce their speed to a level at which they can pass each other. As the flow increases, a multiplicity of events occurs; such events require specific decisions that are discussed elsewhere (8). Most of these events are common to vehicles traveling on narrow as well as wider roads. On narrow roads there are more stringent rules to be considered for event prediction. As roadwidth increases, narrow lane traffic behavior can be logically extended to wider two-lane roads as well as to four-lane divided highways. These are discussed later.

Traffic Interactions on Narrow Roads

For narrow roads a distinctive behavioral logic has been developed to reflect the peculiar conditions prevailing on the Indian road network. This is one of the major achievements of the research reported in this paper.

Overtaking Decisions

Consider what happens when two vehicles with different desired speeds travel in the same direction. The road stretch is assumed to be free of any oncoming vehicle. Let e_i be the faster vehicle and e_j the slower one in front. Eventually e_i catches up

with e_j. The decision situation for e_i is whether to continue to travel at the same speed. The point at which this decision has to be made is similar to that at which the decision about flying overtaking is made in the case of two-lane roads except that e_j is the decision maker. Vehicle e_i could continue to travel at its own desired speed if e_j decides to move partly to the shoulder; otherwise e_i becomes constrained. Observations in the field can be used to assess the yielding probability distribution of the slower vehicle, which is a function of roadway and vehicle parameters. The Monte Carlo method is used to decide whether e_j will yield space so that e_i can travel unimpeded (6).

When e_j does not yield, e_i must decelerate and follow e_j until e_j does yield. If the shoulder ahead is equivalent to an extra lane, e_j will yield with a specified probability and change to the shoulder to travel with a speed consistent with shoulder conditions. Other sections at which e_j could yield are where there is adequate shoulder (beginning at a block border) and the maximum sight point. This overtaking opportunity is accepted if the obstructing vehicle (e_j) yields way by moving to the hard shoulder or climbing lane and if the trailing vehicle (e_i) has sufficient ability to overtake. Three different conditions must be fulfilled for an obstructing vehicle to yield for overtaking:

1. There must be a hard shoulder or extra lane at least a certain number of meters in front, and this model constant is set at 200 m;
2. Surrounding traffic must be such that space is available; and
3. A stochastic function for yielding must be true.

The trailing vehicle is considered to be able to overtake the lead vehicle if the driver estimates the distance available for overtaking to be $\leq x$ m. Vehicle e_j returns to its normal lane when e_i has overtaken it and space is available. The importance of alignment (in terms of the frequency with which maximum sight distance is provided), shoulder type and condition, and their effect on traffic flow on narrow roads are captured in this decision submodel (6).

Meeting and Passing Decision Process

Consider the trajectories of two vehicles that meet and pass each other. Let e_1 be the vehicle traveling in direction 1 and e_2 the vehicle traveling in direction 2. Let v_1 be the speed of e_1 and v_2 the speed of e_2. Scanning the road ahead the driver of each vehicle identifies the oncoming vehicle. Both e_1 and e_2 are required to decelerate and pass each other at their desired passing speeds. Let the passing speeds of these vehicles be v_1' and v_2'. Both vehicles then decelerate at acceptable rates to reach their passing speeds. The decision point at which e_1 initiates its deceleration need not be the same as that of e_2. For instance a bullock cart never reduces its speed when passing. Deceleration distances (D_1 and D_2) are determined and are known as passing head lengths for e_1 and e_2, respectively. It is now possible to identify the start of deceleration decision points for e_1 and e_2. When coordinates of their current positions in space and time are given, the corresponding coordinates at which their passing head lengths overlap can be determined. This point of overlap is defined as the meeting point. Therefore, meeting point is the decision point for vehicles to start deceleration. There are three phases in a passing sequence. First, a vehicle decelerates to its desired passing speed. Second, the passing vehicle travels at a constant speed during the passing mode. Finally, the vehicle accelerates to attain its desired operating speed. This sequence involving deceleration, lower speeds during passing, and then acceleration results in higher levels of acceleration noise and considerable delay to vehicles. In addition, they are exposed to severe conflict situations that lead to high accident rates on these roads.

Meeting and Passing Involving Platoons

A more complex passing situation involving a two vehicle platoon and an oncoming vehicle can be modeled now. Let e_{11} and e_{12} be the platoon leader and constrained vehicle traveling in one direction and e_2 the vehicle traveling in the other direction. To start with, assume that e_{12} is not constrained and that e_{11} and e_2 are separated by a large gap. Eventually e_{12} catches up with e_{11}. At this time, it has also been assumed that e_{11} and e_2 are close to their meeting point. In this situation, it is important to know how the decisions are made by e_{11}. There are two possibilities: In the first case e_{11} allows e_{12} to overtake and then passes e_2. In the second case e_{11} passes e_2 and then allows e_{12} to overtake while it is still on the shoulder lane.

In the first case the speed of e_{11} depends on the shoulder condition as well as the speed of the overtaking vehicle and e_{12} will have to decelerate to pass e_2 immediately after overtaking e_{11}. In the second case e_{12} follows e_{11} and both move to the shoulder and then pass e_2. When e_{11} and e_{12} have passed e_2, e_{11} allows e_{12} to overtake by accelerating (see Figure 4). The second case has been chosen for use in the simulation logic so that account is taken of the preferential structure of drivers; namely, that priority is given to passing and only then are other events considered. This is more often true of fast-moving vehicles except motorcycles.

The procedure involving the prediction of event times during passing is unique to this model. It must be emphasized that when the meeting of vehicles in opposing streams is imminent, this meeting event is given top priority in obtaining vehicle trajectories (6).

Application of Overtaking and Passing Behavior on Indian Roads

Passing is a totally unavoidable event on narrow Indian roads. Overtaking can be avoided or deferred. Therefore, it is passing that is given top priority in obtaining vehicle trajectories on narrow roads, even if overtaking is imminent. The behavior of the platoon leader is not irrational in that it is he who is in the best position to judge passing safety in light of the absence of forward visibility to the following vehicle.

It has been shown that overtaking and passing are highly interrelated on narrow roads. The degree to which this interaction is reduced is not dependent on road designation but on effective pavement width. Due to the ubiquity of narrow vehicles on Indian roads, even in the case of two-lane roads effective roadwidth is often diminished to the extent that narrow road behavior is required. However, the number of events will be extremely large if passings have to be taken into account. On wider two-lane roads, if it can be assumed that the medial friction due to passing is negligible, relevant events could be modeled appropriately as discussed next.

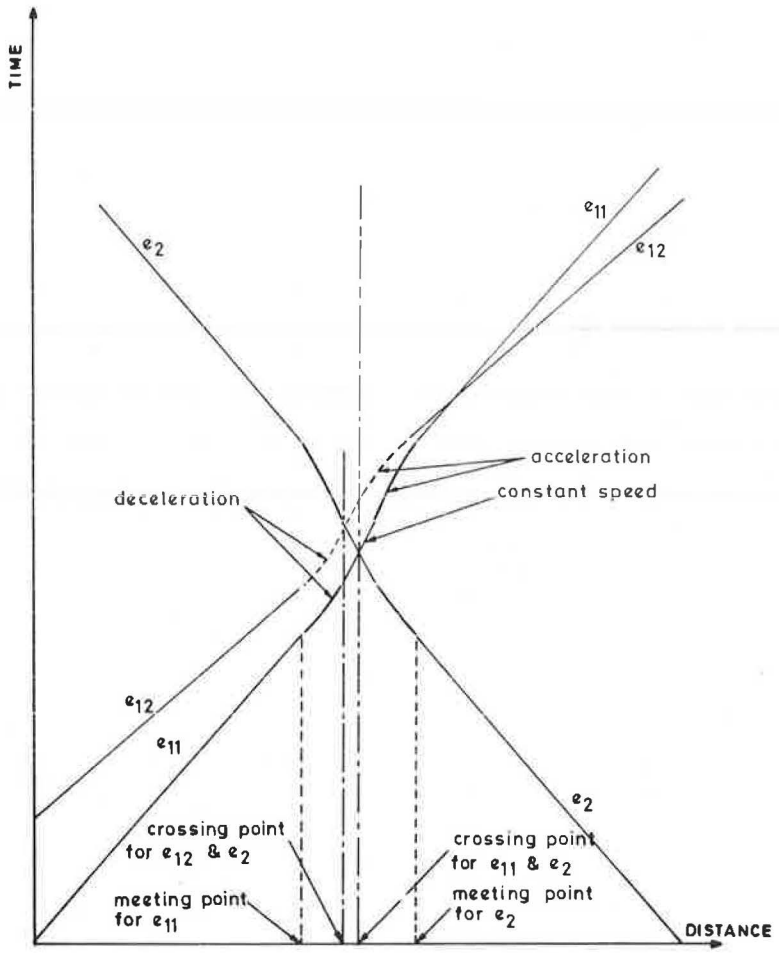


FIGURE 4 Two vehicles with same speed and one with slower speed in passing situation.

Traffic Interactions on Wider Two-Lane Roads

On wider roads overtakings and passings are considered in detail. In considering overtaking situations the VTI model has been adjusted for the reduced impedances of narrow vehicles and ADVs. Gap acceptance functions obtained from the field studies on flat terrain have been adjusted for hilly terrain conditions. Detailed accounts of various events on two-lane roads have been documented elsewhere (6-8). Only the salient modifications proposed for Indian traffic are given here.

Gap Acceptance Function and Modification for Hilly Roads

Overtaking gap acceptance functions have been developed for 32 different situations using test vehicles fitted with video and radar speedmeter instrumentation identical to that used by the Australian Road Research Board (11). They are applicable for roads in flat and rolling terrain. The general form of the functions is expressed as

$$P(x) = \begin{cases} 0 & \text{if } x \leq S_1 \text{ m} \\ (x - S_1)/(S_2 - S_1) & \text{if } S_1 < x < S_2 \text{ m} \\ 1 & \text{if } x \geq S_2 \text{ m} \end{cases}$$

where x is the sight distance and S_1 and S_2 are the calibration constants for each of the 32 overtaking situations (7). On hilly roads, vehicles travel slower and therefore distance required for overtaking is smaller. Thus, the probability of a given gap being accepted is higher on hilly roads. The values obtained for S_1 and S_2 on level roads have been modified as follows:

$$S_1^* = S_1 \, (\bar{V}_{hill}/\bar{V}_{level})^{n}$$

and

$$S_2^* = S_1^* + (S_2 - S_1) \, (\bar{V}_{hill}/\bar{V}_{level})$$

where S_1^* is the minimum gap for overtaking on hilly roads, S_2^* is the distance with highest probability, and $\bar{V}_{hill}$ and $\bar{V}_{level}$ are the mean free speeds of vehicles on hilly and level roads, respectively; n is a calibration constant that is equal to unity in this case.

Modifications for ADV and Narrow Vehicles for Interaction Decisions

In the VTI model decisions regarding overtaking and following are made when a vehicle is close to another vehicle in the same track. This logic holds as long as wider vehicles, which occupy the full lane, are considered but needs modifications to adjust for the interactions involving narrow vehicles such as motorcycles. For example, two motorcycles may be

found traveling beside each other in the same lane at their free speeds. Thus, in decisions involving interactions, distinction should be made in terms of whether a vehicle interacts with a vehicle immediately in front or behind. If vehicles in the immediate vicinity do not interact, the vehicle with which interaction is likely should be identified. The modified model calculates a reference to the interacting vehicle in front of (or behind) the vehicle in question. If the vehicle so identified is not likely to interact, the next vehicle in the lane is tested for interaction and the procedure is repeated until an interacting vehicle is found (if one exists). When the interacting vehicle has been identified, all decisions are made with reference to this vehicle only (7).

Extension of the Model to Four-Lane Divided Highways

In the previous sections traffic on narrow roads and wider two-lane roads was discussed. A logical extension of these models to the case of four-lane divided highway has been achieved by invoking the auxiliary lane concept embedded for overtaking situations on two-lane roads and especially in the case of narrow roads if passing is eliminated (6,7). In unidirectional flow situations the logic used for vehicle movement in the opposing lane and traffic from that direction are omitted while vehicles in a normal lane yield and allow catching-up vehicles to overtake as in the former category of roads. The yielding probabilities are obtained from field studies. The highest probability of a slower vehicle yielding by moving to the outer lane has been set to one with all other conditions having been met. This way, the proposed extension is much simpler than the complex multilane traffic process.

Input Data to the Model

Two types of data are used in the validation of the model. The first consists of data structure describing the road alignment and traffic regulations. The road is described for each direction as a series of homogeneous blocks and the data required are (a) space coordinate for the beginning of each block, (b) carriageway width, (c) hard shoulder width, (d) speed limit, (e) slope, (f) curvature, and (g) roughness. In addition, space coordinates for each sight section and the sight distance available at this section are also required. Codes are used for stretches having overtaking restrictions and hard shoulders or climbing lanes. The second data type consists of data on traffic to be simulated over the defined road stretch. They are, for each vehicle, (a) identity number, (b) vehicle type, (c) basic desired speed, (d) p-value, (e) direction of travel, (f) coordinates for entry and exit from the road, and (g) time and spot speed at entry to the road stretch. Individual vehicle data are collected for both directions of traffic at three points along the road: at the ends and midway. Because vehicles have to be associated with their basic desired speed, an inverse transformation is made on the Q-model on the basis of their entry speed. The p-value for each vehicle is obtained from the known distributions for each vehicle type.

Output Data from the Simulation Model

The resulting events from a traffic simulation are obtained in chronological order for statistical analysis of time headway, travel speed, spot speed, number of overtakings, and so forth. Twenty types of situations have been defined that are stored in the form of an event file. There are a number of post-processing programs to analyze the event file to obtain histograms of travel speeds, time headways, and spot speeds. For any stretch of road, a table of overtakings (by type of overtaken and overtaking vehicle) is obtained along with the percentage of free, constrained, and overtaking vehicles.

VALIDATION OF SIMULATION MODEL SYSTEM AND DISCUSSION OF RESULTS

In the validation exercise comparisons were made between the observed and the simulated distributions of travel speeds, spot speeds, time headways, and overtakings. In addition, the validation test considered the difference between observed and simulated means and standard deviations in order to quantify just how "good" or "bad" the comparisons were. The road stretches used in this test are classified according to topography and Indian road designation. Table 1 gives the sampling framework used. Further validation is in progress for those cells in the matrix that were not included in this study.

Comparison of Observed and Simulated Travel Speeds

Travel speed distribution for the heavy motor vehicles on the single-lane, intermediate-lane, two-lane, and four-lane divided highways are shown in Figure 5. The figure indicates consistently good fit between the observed and the simulated travel speeds. For example, on the single- and intermediate-lane roads for a number of simulations the mean percentage differences between the two distributions are 0.635 and 1.917, respectively, indicating an extraordinary fit (6). For certain simulation runs the observed and the simulated mean values were almost the same, confirming the capability of the model. Also, it can be seen from this figure that the effect of roadwidth is captured in the model system in a rather convincing manner. Similar results have been obtained for the spot speed distributions of vehicles, which further reinforces the validity of the model (6,7).

In addition, Figures 6-8 show the plot of observed versus simulated mean travel speeds for all

TABLE 1 Sample Distribution of Roads and Flows Used in Calibration and Validation

	Designation			
Terrain	Single-Lane Road	Intermediate-Lane Road	Two-Lane Road	Four-Lane Road
Flat	Five stretches with negligible gradient and curves (9 flows)	Two stretches with negligible gradient and curves (3 flows)	One stretch (5 flows)	One stretch (3 flows)
Rolling			One stretch (3 flows)	
Hilly	One stretch (3 flows)		One stretch (2 flows)	

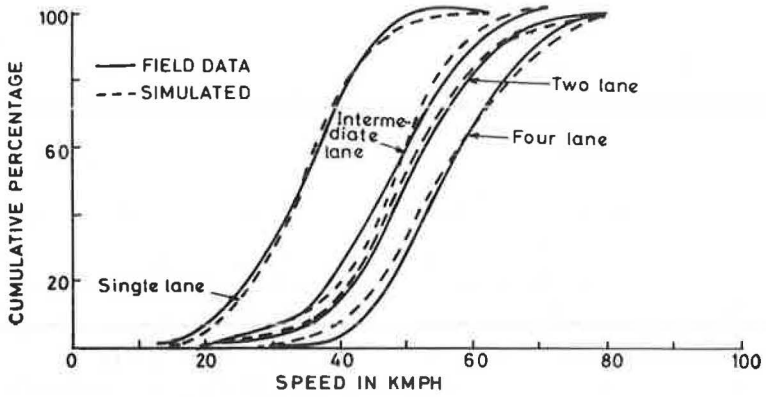


FIGURE 5 Travel speed in km/hr for heavy motor vehicles.

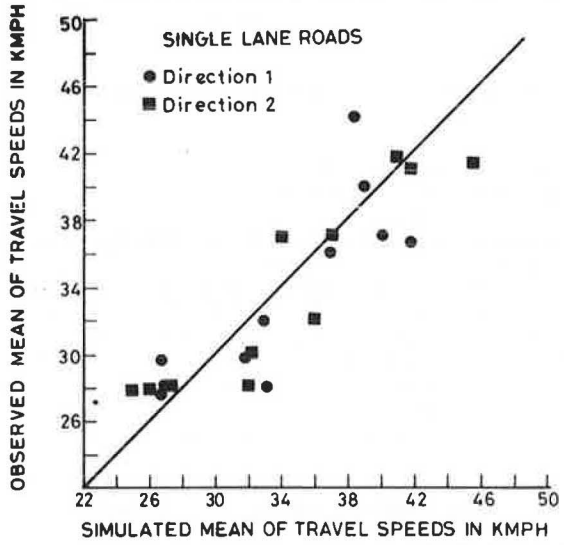


FIGURE 6 Simulated versus observed mean of travel speeds for heavy motor vehicles.

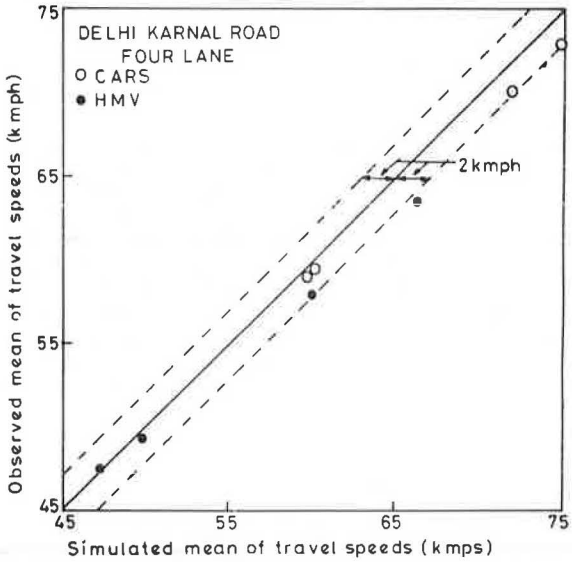


FIGURE 8 Simulated versus observed mean of travel speeds on four-lane road.

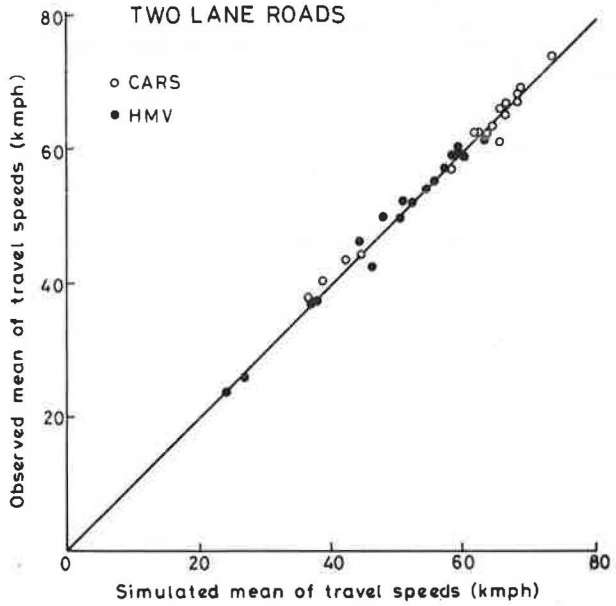


FIGURE 7 Simulated versus observed travel speeds on two-lane roads.

the simulations performed on narrow roads as well as for other, wider roads. This, too, indicates the predictive power of the model.

Comparison of Time Headways at Exits

The mean time headway for a flow of traffic is given by the inverse of the rate of flow. Because the rate of flow for each simulation run was specified from the data, the mean simulated time headway was constrained to be approximately equal to the mean observed time headway for each run. Figure 9 shows the observed versus simulated time headways at exit for the single-lane road. Because the traffic flow was low to medium on all sites, the time headway distributions were all exponential indicating random arrival process at any given point on the road and especially at the exits that have been replicated by the simulation model in close agreement with observed values.

Comparison of Observed and Simulated Overtakings

One of the stringent validation measures in traffic simulation concerns the number of overtakings by

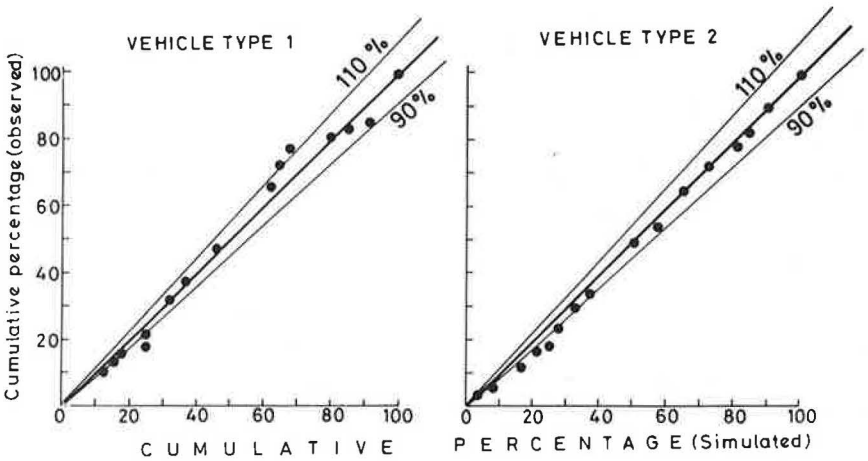


FIGURE 9 Headway distribution of vehicles at exit: G. T. Road, Kanpur, single-lane with bad brick shoulders (Direction 1).

different combinations of vehicle types. Figure 10 shows the observed and simulated overtakings for the two-lane road stretches simulated. The results for other roads are of similar accuracy (6,7). The number of overtakings increases as the roadwidth increases attaining the maximum in the case of divided multilane highways. The results also provide a measure of restraint a given road has on the free movement of the vehicles. The comparison of overtakings was made feasible by the carefully designed data collection procedure as well as by modeling the overtaking decision processes.

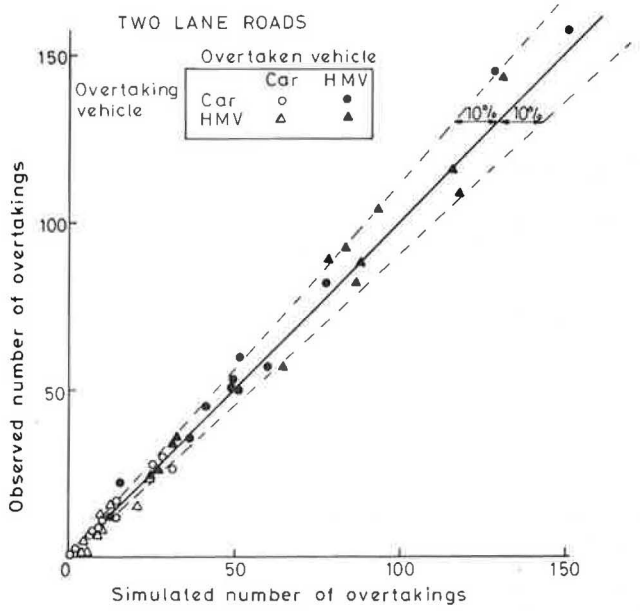


FIGURE 10 Simulated versus observed number of overtakings on two-lane roads.

SUMMARY AND CONCLUSIONS

Indian road conditions are unique in a number of ways and have provided fertile ground for research on quality of traffic flow and resulting highway capacity. Until now, highway capacities and level of service were measured effectively only under the relatively homogeneous vehicle flow conditions found in many Western countries. There, driver behavior is contingent on specific road classification (two lane, four lane divided, and so forth) where specific maneuver patterns are required and better understood by all road users. Heterogeneity in the traffic stream changes the basic premises of Western traffic flow models. On Indian roads, size, width, speed, and operational characteristics vary greatly from one vehicle to another. Individual interactions with other vehicles demand a unique decision-making process. Rigid rules of movement and segregation of vehicles by speed, classification, or direction do not exist on the majority of Indian roads. Therefore, it is the shear width of roads that determines the range of possible driver behavior and hence interaction among vehicles. The breakthrough of the adjusted Swedish model is that it allows analysts to calculate the impact of roadwidth on a myriad of episodic road encounters and the resulting travel time and highway capacity. Combined with the existing knowledge, based on conditions in Western countries, the model for heterogeneous traffic now completes coverage for all rural traffic and road conditions. This model can truly be called a generalized model for vehicle behavior under heterogeneous traffic conditions.

ACKNOWLEDGMENTS

This paper is based on an Indo-Swedish research project that is part of a comprehensive Indian road user cost study. The study has been sponsored by the Ministry of Shipping and Transport (MOST), Government of India, and the World Bank. The Swedish contribution has been funded by the Swedish International Development Agency (SIDA). The Central Road Research Institute (CRRI), New Delhi, has been the responsible executing agency assisted by the Indian Institute of Technology (IIT), Kanpur. The National Swedish Road and Traffic Research Institute (VTI) has acted as advisor.

Study Director L.R. Kadiyali (CRRI) and Research Director G. Gynnerstedt (VTI) have been responsible for the introduction of the simulation technique and the adaption of the Swedish Traffic Simulation Model system to Indian road conditions. Field studies for calibration and overall validation have been executed by CRRI, and the extensions and modifications for the two complementary versions of the model system have been made at IIT by B.R. Marwah and S.P.

Palaniswamy. The VTI team (G. Gynnerstedt, A. Brodin, and A. Carlsson) has acted as advisors and assisted in the development work. The project has received overall guidance from the World Bank group (C. Harral, P. Fossberg, and T. Watanatada). The valuable advice and encouragement of G. Singh and K.K. Sarin, directors general of road development, MOST; C.G. Swaminathan and M.P. Dhir, directors, CRRI; B. Stroem and H. Sandebring, directors general, and B. Thunberg, director, VTI; and S. Sampath, director IIT, Kanpur, are gratefully acknowledged.

REFERENCES

1. A. Carlsson, G. Gynnerstedt, and B. Westerlund. A Model for the Monte Carlo Simulation of Traffic along Two Lane Single Carriageway Rural Roads. Communication 43. VTI, Linkoping, Sweden, 1977.
2. G. Gynnerstedt. Systems Analysis of Road Traffic. Communication 44. VTI, Linkoping, Sweden, 1977.
3. A. Brodin, G. Gynnerstedt, and G.A. Levander. Program for the Monte Carlo Simulation of Vehicle Traffic along Two Lane Rural Roads. Communication 143. VTI, Linkoping, Sweden, 1979.
4. A. Brodin and A. Carlsson. VTI Trafiksimulerings-model-del I Berkrivning av modell och programsystem. Communication 321. VTI, Linkoping, Sweden, 1983.
5. Road User Cost Study in India. Final Report. Central Road Research Institute, New Delhi, India, 1982.
6. S.P. Palaniswamy. A Generalized Simulation Model for Vehicle Behavior Under Heterogeneous Traffic Conditions. Indo-Swedish Traffic Simulation Research Project, IIT, Kanpur, India, 1983.
7. B.R. Marwah. A Simulation Model of Two Lane and Four Lane Highways for Indian Conditions. Indo-Swedish Traffic Simulation Research Project, IIT, Kanpur, India, 1983.
8. G. Gynnerstedt. Indo-Swedish Traffic Simulation Research Project. Draft report. Central Road Research Institute, New Delhi, India, 1983.
9. A. Brodin. VTI Traffic Simulation Model, a Program for the Monte Carlo Simulation of Vehicle Traffic along Two Lane Rural Roads: An Application of JSP and SIMULA. Communication 322A. VTI, Linkoping, Sweden, 1983.
10. S.P. Palaniswamy and A. Brodin. A Generalized Simulation Model--A Program for the Monte Carlo Simulation of Heterogeneous Vehicle Traffic along Single, Intermediate and Narrow Two Lane Roads: An Application of JSP and SIMULA-67. Indo-Swedish Traffic Simulation Research Project, IIT, Kanpur, India, 1984.
11. R.J. Troutbeck. Overtaking Behavior on Australian Two Lane Rural Highways. Special Report 20. Australian Road Research Board, Numawading, Victoria, 1981.

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

Reduced-Delay Optimization and Other Enhancements in the PASSER II-84 Program

EDMOND CHIN-PING CHANG, CARROLL J. MESSER, and BLAIR G. MARSDEN

ABSTRACT

The development of a research study conducted by the Texas Transportation Institute entitled "Reduced-Delay Optimization and Other Enhancements to PASSER II-80" is summarized. The research was sponsored by the Texas State Department of Highways and Public Transportation (SDHPT) in cooperation with the FHWA, U.S. Department of Transportation. The brief 6-month research effort was directed toward several topic areas including development of a reduced-delay optimization procedure that could fine tune the offsets of traffic signals to reduce total arterial system delay and maximize arterial progression, development of methods that can better estimate vehicular delay in a nearly saturated traffic system, and development of methods to estimate fuel consumption for arterial traffic movements in an urban network. Significant enhancements have been made to the popular PASSER II program. This study also demonstrated the practical combination of the maximum bandwidth procedure and the minimum delay algorithm to effectively maximize progression and reduce delay, stops, and fuel consumption in optimizing arterial traffic signal operations. An enhanced version of the PASSER II-80 program, PASSER II-84, was programmed on Texas SDHPT's computer system. Program documentation and revised data-coding instructions were also prepared.

Continued demand for urban mobility requires that the highest degree of traffic service be obtained from existing urban arterial streets and intersections. The ability of signalized intersections to move traffic depends on the concurrent functioning of existing traffic control devices and proper signal timing settings on the street (1-3).

Traffic signal optimization is a complicated process that determines the cycle length, green time, phase sequence, and offsets between the signals. Optimization depends heavily on the relationships among the distances between signalized intersections, travel speed, cycle length, roadway capacity, and side friction along the arterial. Global optimization is time consuming and difficult to achieve without a thorough understanding of the interactions and sensitivities of the site-dependent variables. An alternative is to select proper independent variables, define relationships, and solve the optimization problem heuristically.

Computer techniques for off-line, fixed-time signal timing optimization have commanded widespread interest, but they can optimize merely a portion of the signal timing plan variables, one step at a time. Above all, the models involving a nonlinear formulation still cannot guarantee an optimal solution. The major development of a computerized signal timing optimization algorithm began in the early 1960s with the coordinated offsets of consecutive traffic signals for maximum throughput (1-3).

Currently, two major approaches to coordinating traffic signals along arterial streets are used: (a) the bandwidth maximization-based procedure and (b) minimization of disutility functions such as delay, stops, fuel consumption, or air pollution. The former includes Progressive Analysis and Signal System Evaluation Routine (PASSER), MILP, and MAXBAND (4-6). The latter includes TRANSYT-7F, MITROP, and SIGOP (7-12).

Because of the easily understood time-space diagram and the favorable progressive movement, several maximum bandwidth-based procedures were developed. Generally, relative progression efficiency depends on distances between signalized intersections, travel speed, cycle length, roadway capacity, and side friction along the arterial. On the other hand, delay is well recognized by traffic engineers as a useful tool for evaluating a traffic control system (13,14). However, the calculation with maximum bandwidth does not necessarily minimize the total delay due to the difference of objective functions. Appropriate traffic signal settings can help smooth the traffic flow through a street network, thereby reducing delay and stoppage (7,15-17).

Many traffic engineers still prefer maximum bandwidth settings because of the easily applicable time-space diagrams and the apparently verifiable progression along a major arterial street (4,5,7,10,18). The benefit of signal progression synchronization can be confirmed visually in the field, thereby minimizing complaints from a demanding public. In addition, several studies [e.g., Wagner (19); Wallace (20); Gerlough and Barnes, cited by Rogness (21); Rogness (21), and Cohen (14)] together with much practical user experience demonstrate that the bandwidth method does yield consistently good results on arterial progression systems.

Research by Huddart (22) indicated the possibility of arriving at a compromise between the maximizing bandwidth and minimizing delay method (using a stop penalty) in computing traffic progression performance. Wallace (12) also encouraged the use of PASSER II as a preprocessor for TRANSYT to minimize systemwide delay. Rogness (21) used a heuristic procedure to study the relative performance of PASSER II and TRANSYT programs under synthetic scenarios of cycle length, intersection spacing, and phasing sequence for single arterial street signal timing optimization. He further concluded that there is potential for obtaining good to optimal solutions by combining PASSER II and TRANSYT with some recommended enhancements. Cohen (14) suggested a similar heuristic using MAXBAND with TRANSYT.

Substantial improvement of the total arterial system operations could be achieved by combining the apparent advantages of maximum bandwidth and minimum delay (9,19-21). The maximum bandwidth solution, based on the time-space diagram calculation, is the most efficient way to provide optimal signal phasing sequences. It is less affected by travel demand fluctuation than are solutions of minimum delay.

The PASSER model was first developed by Messer et al. (4) and modified to an off-line computer program cooperatively by the Texas Transportation Institute and Texas State Department of Highways and Public Transportation (SDHPT) (23). PASSER II was designed primarily for high-type arterial streets with protected left-turn lanes and phases. It provides the timing parameters for modern eight-phase controllers, such as the phase sequences, cycle length, green splits, and offsets, to provide the minimum interference and maximum progression bandwidth efficiency. The theory, model structure, methodology, and logic of PASSER II have been evaluated and documented (4,23-25).

PASSER II is widely used because of its ability to select multiple phase sequences in an easy, understandable maximum progression solution format. However, the heavy reliance on bandwidth optimization to achieve maximum progression might somehow limit its optimal solution capability to minimize systemwide vehicular delay. To improve the PASSER II computer program, the Texas SDHPT sponsored a highway planning and research project entitled "Reduced-Delay Optimization and Other Enhancements to PASSER II-80." This study developed the fundamental procedures of fine tuning offset to minimize total arterial delay and preserve the convenience of bandwidth maximization for multiphase traffic signal timing optimization. By applying the enhanced reduced-delay algorithm and fuel consumption computations, improved signal timing can be expected for PASSER II users in the future.

STUDY PROCEDURE

This study was undertaken to find an efficient and usable delay-based search algorithm for selecting a reduced-delay, arterial signal timing plan based on a maximum bandwidth solution. Four major items were developed: delay calculation methodology, offset fine-tuning capability, offset optimization routine, and fuel consumption computations.

The delay reduction procedures have several assumptions:

1. Cycle length, green split, phase sequence, and progression speed are known for each signal;
2. The optimal time-space diagram, shown in Figure 1, a maximum bandwidth solution, is provided as the starting solution and constraint for arterial system delay-offset analysis; and
3. The interactions between two intersections depend on the signal phase pattern, traffic volumes, and offsets of neighboring intersections.

At first, specific enhancements to improve the performance of PASSER II-80 as a maximum bandwidth-based procedure were identified. Then, the existing PASSER II-80 program was extended to provide a maximum bandwidth-based minimum-delay solution and, at

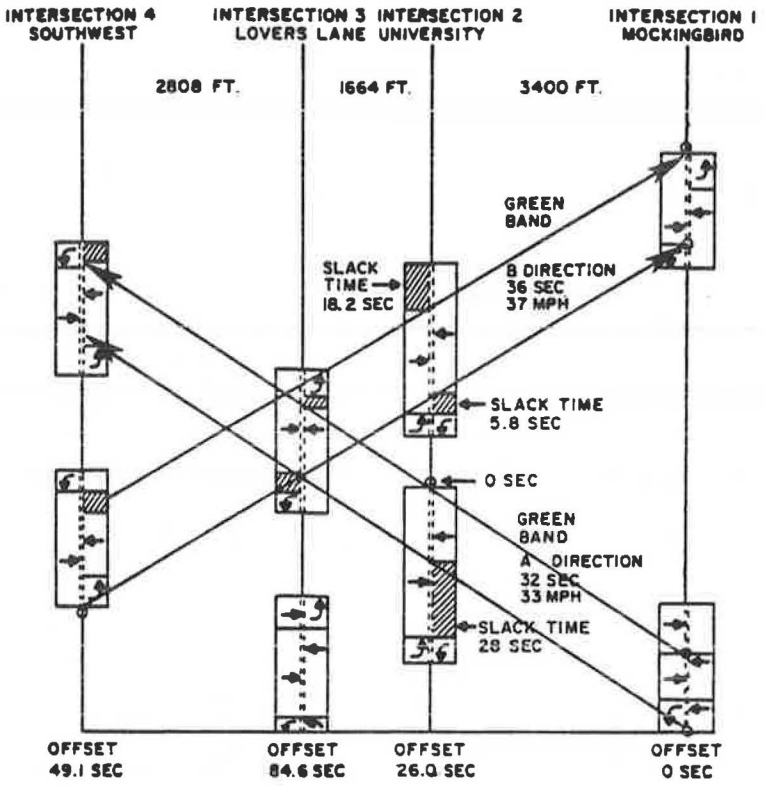


FIGURE 1 Example of slack time, slack-time allowance, and allowable offset range in PASSER II-80 time-space diagram (Skillman Avenue, Dallas, Texas).

the same time, to minimize delay, stops, and fuel consumption on arterial streets. Finally, the fuel consumption computations of TRANSYT-7F were enhanced for PASSER II-80 with the capability for future coefficient modifications by FHWA.

PASSER II ENHANCEMENTS

Efforts were made to improve the PASSER II-80 program to reflect users' working experience (24,25). The basic objective was to add a system offset fine-tuning optimization routine to PASSER II-80. The new extensions began with fine tuning the offsets starting from an existing progression solution in straightforward, deterministic, and noniterative approaches without affecting the major input and output structure.

Figure 2 shows a summary of the initial PASSER II-80 solution and enhanced PASSER II-84 program outputs. At first, the green splits are calculated by the modified Webster green split routine to equalize the specific volume-to-saturation flow ratios on critical movements. Then, optimal progression solutions are calculated by Brook's minimum interference theory to optimize phasing sequence and offset arrangements within coded preferable speed and optimal cycle length (26). The progression bandwidths are further adjusted to the sum of the total link volumes in both the A and the B directions.

After these calculations, the "best solution" and resultant time-space diagram, as in PASSER II-80, provide the initial solution to PASSER II-84. The enhanced PASSER II-84 can further provide the following capabilities:

1. Check the through progression band versus the actual green time interval. Detect any "plot through

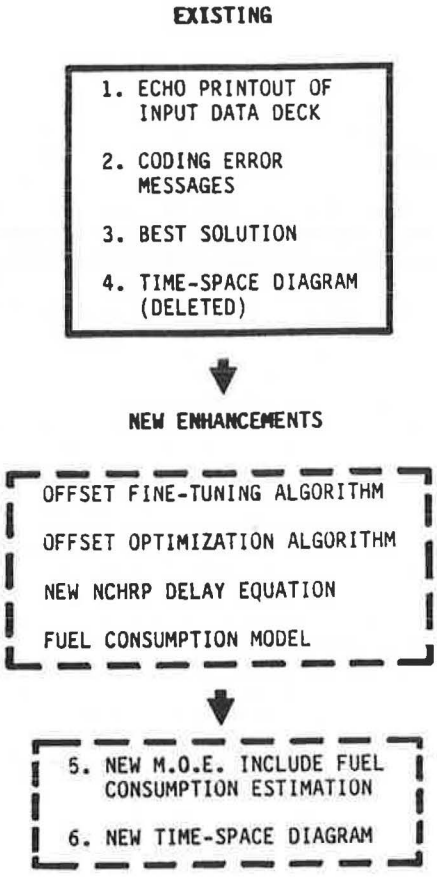


FIGURE 2 Schematic layout of elements in new PASSER II-84 output.

the red" and correct it, and then revise the available slack-time allowance in the time-space diagram.

2. Optimize offsets between intersections using delay-offset analysis. Adjust the time-space coordinate for a new time-space diagram.

3. Calculate the average delay per vehicle evaluation using the tentative NCHRP delay estimation equation to account for the oversaturated condition with saturation ratio greater than 1.0.

4. Estimate total fuel consumption (gal/hr) by the modified fuel consumption estimation model used in TRANSYT-7F.

5. As an option, compute the perfect one-way progression solution with the allowable design speed variations.

6. Provide the optional translation of the phase movement definitions of NEMA and PASSER II.

Finally, the result of the enhanced PASSER II-84 calculations will supply a reduced-delay best solution, including a fine-tuned time-space diagram and fuel consumption calculations.

Tentative NCHRP Delay Equation

Analytical delay estimates are commonly used in many computer models. The most widely used one is the Webster's model as plotted in Figure 3, which is based on Pignataro (2). However, because of the mathematical discrepancy, it is only applicable for a saturation ratio (or V/C ratio) up to about 0.95.

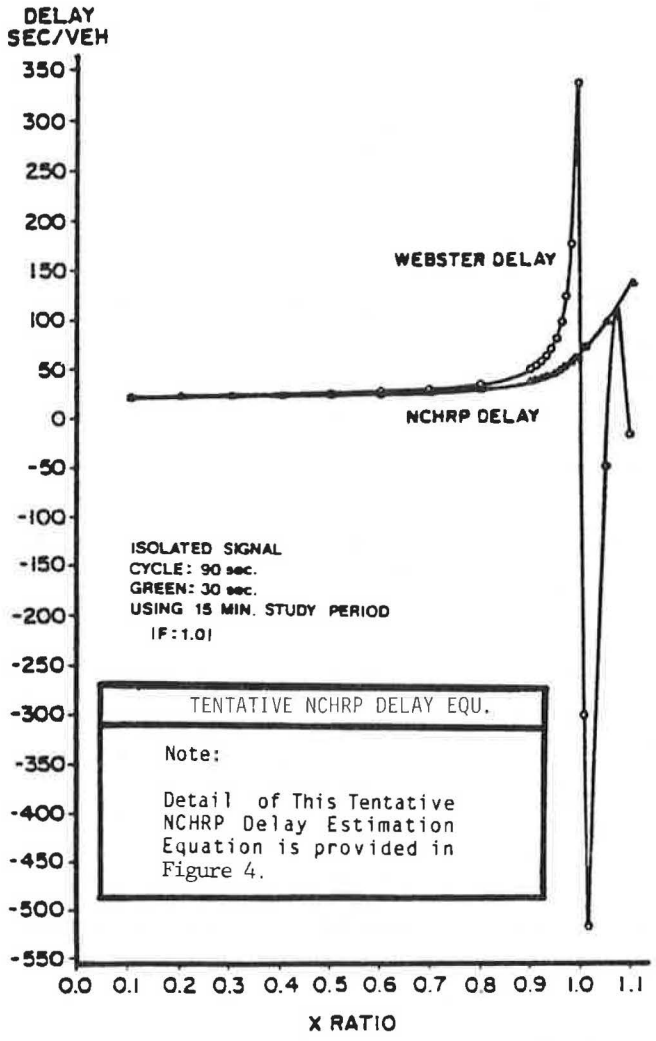


FIGURE 3 Webster and tentative NCHRP delay versus X ratio.

In updating the 1965 Highway Capacity Manual, National Cooperative Highway Research Program Project 3-28(2) developed a capacity and level of service method for urban signalized intersections. Specifically, a tentative delay estimation equation was developed to calculate delay and level of service for each lane combination (right, through, or left-turn lane), approach, and the overall intersection under normal, saturated, and oversaturated conditions. Summaries of the delay equations for the basic delay conditions are shown in Figure 4, which is based on Messer and Roess (27).

Uniform delay (UD) occurs in the period when all queued vehicles clear the approach on each cycle, assuming that none of the queued vehicles has to wait through more than one red period. The UD formula estimates the average stopped delay per approach vehicle for lane groups with a V/C ratio less than or equal to the overflow condition. The formula is based on uniform arrivals for various analysis period lengths (5 min to several hours). Overflow delay (OD) occurs when, on some cycles, all queued vehicles clear the approach while, on other cycles, some of the queued vehicles do not clear the approach due to variation of the traffic volume. Overflow delay is estimated according to the amount of the V/C ratio; that is, this empirically derived value adjusts the amount of the overflow delay according to the degree of oversaturation and randomness of traffic arrival patterns. The uniform delay component is not affected by the length of the analysis period. The overflow delay, because it is an estimate of arrival variations, is highly dependent on the analysis period. For convenience, a 15-min analysis period was assumed in PASSER II-84.

The tentative NCHRP delay estimation equation is valid for a V/C ratio above 1.0. However, the following guidelines should be noted when the V/C ratio is over 1.0:

1. Use the actual approach volumes. Check the analysis to assure that the volumes have not been adjusted to analyze the peak 15-min period or the worst lane;
2. Use the V/C ratio derived from the actual volumes, not the adjusted volumes; and
3. Use the time period that relates to the volumes and the V/C ratio.

Figure 3 indicates that the tentative NCHRP equation estimates the same or less delay than Webster's equation when V/C < 1.0 and provides a much better estimate of delay in oversaturated conditions. However, because the NCHRP delay equation was primarily designed for evaluation of uncoordinated signalized intersections, a version of this tentative NCHRP equation with a modified uniform delay term was added to the PASSER II-84 program by "platoon interconnection" adjustment as used in the existing PASSER II-80 program (28).

Offset Fine-Tuning Algorithm

To minimize total arterial system delay, efforts were made to find an efficient and applicable method to fine tune the offsets of individual intersections in an arterial system. It was determined that any new enhancements to PASSER II-80 should not conflict with the original maximum progression solution; such enhancements should, instead, fine tune this base solution by adjusting the relative offsets.

$$\text{NCHRP DELAY} = UD + OD$$

$$\text{WHERE} \quad UD = 0.5 * \left[\frac{(C) * (1 - G/C)^2}{(1 - V/S)} \right]$$

$$OD = 225 * F * (X)^2 * \left[(X-1) + \sqrt{(X-1)^2 + \frac{16X}{V * F}} \right]$$

WHERE UD - UNIFORM DELAY
OD - OVERFLOW DELAY
C - CYCLE LENGTH
G - EFFECTIVE GREEN TIME
V - DEMAND VOLUME
S - SATURATION FLOW RATE
X - SATURATION FLOW RATIO
F - STUDY PERIOD ADJUSTMENT FACTOR
F = 1.0 WHEN STUDY PERIOD IS 15 MIN.
F = 4.0 WHEN STUDY PERIOD IS 1 HOUR

FIGURE 4 Tentative NCHRP delay estimation equation.

In the offset fine-tuning algorithm of PASSER II-84, subroutines FINTUN and PUSHUP first check through band versus actual green time, detect and correct the rare "plot through the red" condition, indicate the available slack-time allowance for the offset optimization, then reconstruct the time-space coordinates by travel time and distance calculations. The available slack-time allowance is the slack time, for a particular direction at each signal, that the offsets could be adjusted without losing the optimal progression solution and bandwidth.

Then, all the slack time is identified on both sides of the progression band for both travel directions at each traffic signal. The minimum value of the slack time that can be adjusted indicates the maximum amount of allowable green time available for adjusting the existing offset at each signal without affecting the bandwidth of the correct progression solution. This algorithm further reduces the need for manual adjustments on the final time-space diagram and provides a basic range of solutions for later constrained offset-optimization without searching through the entire cycle length as is done in the ordinary delay-offset analysis.

This offset fine-tuning algorithm is illustrated using the example time-space diagram of Skillman Avenue shown in Figure 1. At first, the slack time available for through movement but not used in the existing time-space diagram is indicated by hash marks. After the comparison of the relative magnitude of the slack time, the minimum values of slack time in each direction are identified as the allowable slack-time ranges at each signal for later offset fine-tuning optimization.

For example, in Figure 1, the slack times at the second signal are identified, respectively, as 28, 0, 5.8, and 18.2 sec on either side of the progression bands in both the A and the B directions. Because the A direction progression band is constrained by the zero slack time on the upper side of the A progression band for downward offset adjustment, the only allowable slack time available for upward offset adjustment, without affecting the width of both progression bands in either direction at intersection 2, is 5.8 sec on the lower end of the progression band in B travel direction. The resultant allowable slack-time range as found by this algorithm is 5.8 sec, and the resultant allowable offset adjustment range is, therefore, from the original 26 sec to a possible 31.8 sec.

Offset-Optimization Algorithm

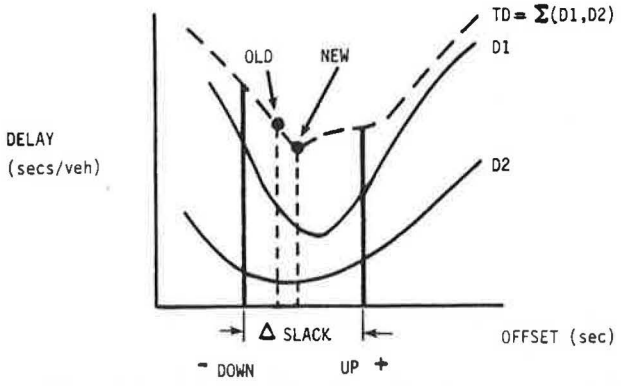
In PASSER II-84, subroutines OFSE2, OFSE3, and TSCORD provide an optimum progression offset between each signal. Further fine-tuning optimization would be obtained by adjusting the initial progression offset to some other fine-tuned offset within the allowable slack-time range to reduce the total two-way link delay from subroutines FINTUN and FKCDLY.

In this algorithm, progression remains the highest priority optimization objective and serves as the base for further optimization. The delay-offset analysis only fine tunes the offset within the allowable slack-time range. Operational performance is evaluated by average vehicle delay experienced by all vehicles. Because the total number of vehicles operated in the arterial street system during a fixed time period is a constant, adjusting signal settings can only redistribute the traffic and resultant delay on the network. Vehicular delays occurring between the intersections are calculated by a ver- sion of the deterministic delay-offset technique similar to that used in PASSER III for signalized diamond interchanges (24,29-31).

According to the combination method, where two or more links occur in parallel joining two nodes, the delay function of the individual links can be combined with reference to the same offset to yield an aggregate delay function. Then, the total delay function is calculated by combining all the individual delay-offset functions of the through links in both the A and the B directions. The average combined delay function is obtained by dividing the total delay of the adjacent signal pairs by the total link traffic volumes. An optimal offset, be-

tween the adjacent pair of signals, is obtained by searching for the minimal value of the combined delay-offset function.

Figure 5 shows the basic theory of the constrained delay-offset analysis in PASSER II-84. As indicated, the delay-offset curves D1 and D2 for each internal link are first developed. Then, the total delay-offset curve between a pair of signalized intersections is derived by accumulating the delays on each link at respective offset locations. The TD curve shown in Figure 5 is volume weighted instead of the sum of D1 and D2. The slack time allowance range (ΔSLACK) for each intersection is then identified by the range between -DOWN and UP+.



STEP 1. Calculate Average Delay-Offset Curves (D1 and D2),

STEP 2. Sum Up Individual Link Delay-Offset Curves to Obtain Total Systemwide Average Delay Per Vehicle Curve (TD),

STEP 3. Identify Slack Time Allowance Range (ΔSlack) on TD,

STEP 4. Obtain Reduced-Delay Solution (NEW) from Existing Progression Solution (OLD).

FIGURE 5 Delay offset analysis in PASSER II-84 offset-optimization algorithm.

Finally, the systemwide average delay per vehicle of the existing progression solution (OLD OFFSET) is calculated, and the offset optimization algorithm will search over the slack-time-allowance range (ΔSLACK) to obtain the reduced-delay solution (NEW OFFSET) as shown. Therefore, the optimization problem becomes:

> Find a new reduced-delay offset (NEW) within the slack-time allowance (-DOWN≤SLACK≤UP+) for a given combination of fixed cycle, phase sequence, green split, initial progression offset, and two delay curves (D1 and D2) of right-turn and through movements in both A and B directions.

In this algorithm, both the original maximum bandwidth procedure and the minimum delay algorithm are considered. The maximum bandwidth solution (OLD) in PASSER II-80 can be improved by this modified delay-offset algorithm to a reduced-delay solution (NEW) under multiphase operation in PASSER II-84. The detailed evaluation procedure and study results comparing the PASSER II-80 and the PASSER II-84 programs are discussed in another paper in this Record entitled "Minimum Delay Optimization of a Maximum Bandwidth Solution to Arterial Signal Timing."

Fuel Consumption Model

Faced with a fuel shortage and the increased fuel prices prevalent in the 1970s, traffic engineers became more and more sensitive to the consequences of delay and stops. To provide a more realistic evaluation of traffic signal alternatives, a fuel consumption estimation model was applied with the measure of effectiveness from PASSER II-84. It is capable of accepting any future modifications, by FHWA, to the fuel consumption equations.

After reviewing the available fuel consumption estimation models, the fuel consumption routine in TRANSYT-7F was modified and added to PASSER II-84. The model was developed from a series of stepwise multiple regression analyses of data collected and programmed by the Transportation Research Center of the University of Florida (8). The basic model estimates the total arterial system fuel consumption (gal/hr) as a function of total travel (veh-mile/hr), total delay (veh-hr/hr), total stops (veh/hr), and cruise (free) speed (mph).

Among these variables in the fuel consumption model, the total stops in vehicles per hour was the only variable not available in PASSER II-80. Therefore, a modified formula, developed by Akcelik and Miller, was applied in PASSER II-84 to estimate the total stops for coordinated multiphase traffic signals operated on arterial streets (8,17). In summary, the formula for estimating total number of stops per hour is calculated directly as

$$H = (3240/C) \cdot [(vr/1 - g/s) + N_o]$$

where

v = arrival flow rate (veh/sec),
C = cycle time (sec),
g = effective green time (sec),
s = saturation flow rate (veh/sec), and
N_o = average overflow queue (veh/sec) where

$$N_o = \exp\left(\{1.33 \cdot [(1 - vC/gs)/(vC/gs)] \cdot (s \cdot g)^{1/2}\}/(2 - vC/gs)\right).$$

Perfect One-Way Progression

Because of the physical restrictions and unique traffic characteristics of the urban street network, the PASSER II program may sometimes be used to provide one-way progression. This option can provide the optional time-space diagram for a one-way street or for an arterial street system with heavy directional peak-hour travel.

Subroutine ONEWAY, similar to the one in PASSER III, calculates the offsets and overwrites the time-space coordinates providing "perfect" one-way progression along a two-way arterial street (26,p.27). The "perfect" one-way progression solution in either the A or the B direction can be obtained by specifying a 1 or a 99 in the optional "min. B direction band split" of the PASSER II-84 input data set. Figure 6 shows an example of the result from the subroutine when 1 is specified for the "perfect" one-way progression option in the A direction.

As indicated, a "perfect" progression band that uses the whole amount of the available green time for progression movement in the A travel direction is provided as specified. Similarly, the "perfect" one-way progression band can be provided for the progressive movement in B travel direction by specifying 99 in this one-way progression option. In both cases, the traffic signal settings optimized

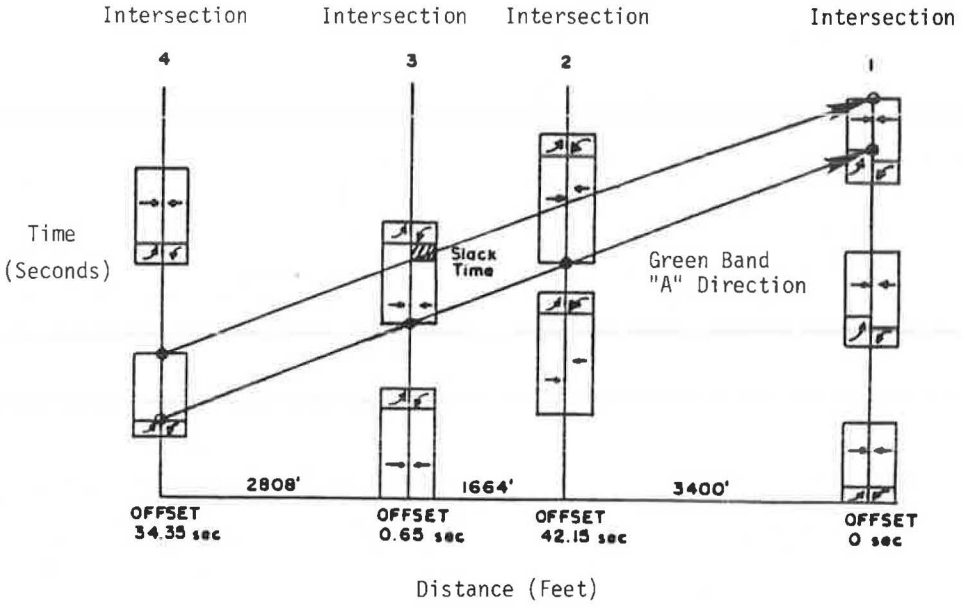


FIGURE 6 Time-space diagram of the perfect one-way progression in A direction.

for the original two-way progression solution of the PASSER II program are used with the fine-tuning offsets.

Phase Movement Translations in NEMA and PASSER

Two widely accepted phase movement designations have been used in PASSER II-80: National Electrical Manufacturers' Association (NEMA) and PASSER phase definitions. As shown in Figure 7, the NEMA designation could be considered as swapping the major street movements 3 and 4 with the minor street movements 5 and 7 compared to PASSER's designation.

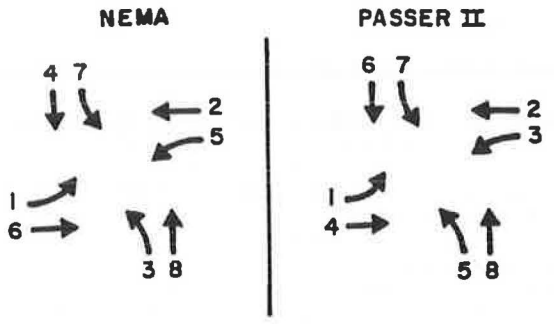


FIGURE 7 Movement definitions in NEMA and PASSER II.

As the data in Table 1 indicate, PASSER II-80 uses PASSER's phase designation as default input with the option of using the NEMA movement definition as an alternative. As indicated, PASSER II-84 still has the existing PASSER II-80 options to choose between: the NEMA (option 1) and the PASSER phase definition (option 0). In addition, if the user wants to use PASSER's phase definition as input but chooses the NEMA phase definition as output, a 2 may be entered. If the user prefers to use the NEMA phase as input but desires PASSER's movement numbering definition for output, 3 may be entered in the

TABLE 1 Movement Translation Codes for NEMA and PASSER II

INPUT OPTION / OUTPUT OPTION		INPUT P2	ECHO NEMA
OUTPUT	P2	0	3
	NEMA	2	1

data field. PASSER II-84 recognizes the options selected (or default) and provides proper phase movement designations in both the echo printout of input data deck and the final printout of the PASSER II-80 "best solution."

Summary

In summary, significant programming efforts have been completed using ANSI FORTRAN 77 standard on the Amdahl Computer System at Texas A&M University. The results comprise the revised PASSER II-84 program with the new delay calculation and offset optimization routine, as shown in Figure 8. These efforts permit the user to determine the optimal signal settings for progression operation on signalized arterial streets without having to manually adjust the offsets of the final time-space diagram for reduced-delay operation.

CONCLUSIONS AND RECOMMENDATIONS

The enhanced reduced-delay optimization in PASSER II-84 guarantees minimizing total arterial system delay within the slack-time allowance of the original PASSER II solution. However, the general improvement that can be achieved by PASSER II-84 relies mainly on the quality of the original answer. If the green times were intentionally constrained,

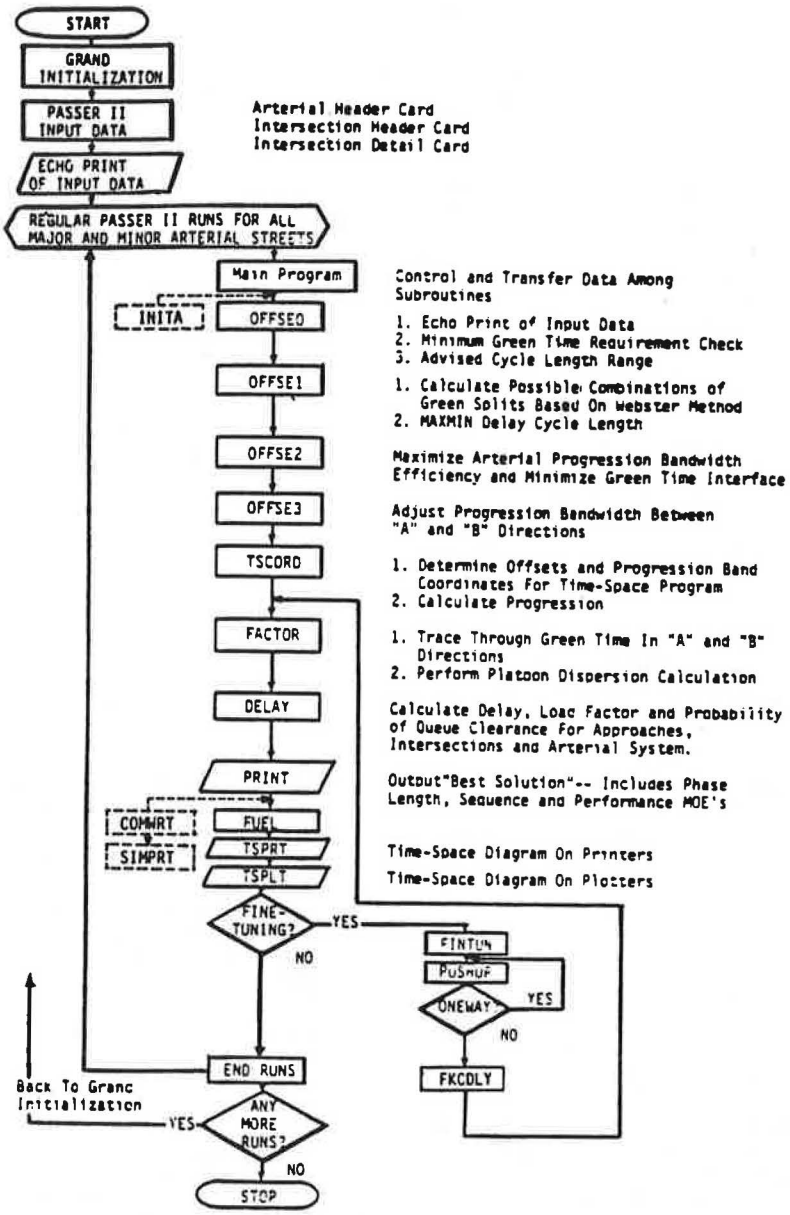


FIGURE 8 Functional flowchart for PASSER II-84.

or had been engineered in an expert manner, the improvements would not be as significant as they would be for ordinary PASSER II solutions.

This study and several other related studies have also found that trading off progression bandwidth in either arterial travel direction, instead of using the total directional traffic volume ratios and minimum green as constraints, may further improve the total system performance. Because the enhanced PASSER II-84 program does not have a microscopic simulation model to predict actual platoon dispersion effects on the downstream signals, the accuracy and estimation ability of the "platoon projection model" or the "platoon interconnection effect" is constrained by site-dependent travel behavior, vehicular mixes, and travel speeds.

Therefore, further research is recommended on field validation of the reduced-delay offset-optimization algorithm, the calibration of platoon dispersion models, alternative strategies for allocating the directional bandwidths, revision of green split routine to account for the impact of green time adjustment on overall system delay, and trade-offs of local and system optimization problems in arterial signal optimization.

Documentation of research results in the development of PASSER II-84, the latest version of arterial traffic signal timing computer model of the Texas State Department of Highways and Public Transportation, is available elsewhere (32). A comparison of the basic features in existing PASSER II-80, enhanced PASSER II-84, TRANSYT-7F, and MAXBAND computer models is given in Table 2. No modifications to the existing user's manual or data coding are required in the enhanced PASSER II-84. The basic program is currently operational on the Texas SDHPT district remote computer terminals.

ACKNOWLEDGMENTS

The paper is based on research conducted for the Texas Department of Highways and Public Transportation in cooperation with the FHWA, U.S. Department

TABLE 2 Comparison of PASSER II-80, PASSER II-84, TRANSYT-7F, and MAXBAND Computer Programs

	PASSER II (NEW)	PASSER II (OLD)	MAXBAND (MITROP)	TRANSYT-7F
Control Variables	Cycle Offset ϕ-Sequence ϕ-Length	Cycle Offset ϕ-Sequence ϕ-Length	Cycle Offset ϕ-Sequence ϕ-Length	Cycle Offset ϕ-Length
Optimization	Max. Bandwidth Min. Interference Delay-Offset	Max. Bandwidth Min. Interference	Max. Bandwidth MPCODE	Min. Delay
Solution	Local Optimum	Local Optimum	Global Optimum	Local Optimum
Objective Function	Max. Efficiency Min. Link Delay (Interior)	Max. Efficiency	Min. (Flow × Cost) Offset Split	Min. PI = (Delay + k × stops)
Delay Measurement	Mod. Webster NCHRP-TTI-PINY	Mod. Webster	Link Performance Saturation Deterrence	Platoon Representation by Flow Profile
Delay Component	1. Uniform Delay 2. Overflow Delay	1. Uniform Delay 2. Random Delay 3. Empirical Adj.	1. Deterministic Queue 2. Stochastic Overflow	1. Uniform Delay 2. Random Delay 3. Saturation Delay
Fuel Consumption	Yes	No	No	Yes
Data Input Base	Node	Node	Node	Link
Phase Selection	PASSER NEMA or Combinations	PASSER NEMA ($\frac{A}{B}$ Dir.)	Inbound Outbound	Link Movements

of Transportation. The technical input and constructive comments from Texas SDHPT and FHWA officials during the conduct of this research are greatly appreciated.

REFERENCES

1. Manual of Traffic Engineering, 2nd ed. Institute of Traffic Engineers, Washington, D.C., 1976.
2. L.G. Pignataro. Principles of Traffic Engineering--Theory and Practice. Prentice-Hall, New York, 1976.
3. P. Ross and D. Gibson. Review of Road Traffic Network Simulation Models. In Transportation Research Record 644, TRB, National Research Council, Washington, D.C., 1977, pp. 36-41.
4. C.J. Messer, R.H. Whitson, C.L. Dudek, and E.J. Romano. A Variable Sequence Multiphase Progression Optimization Program. In Highway Research Record 445, HRB, National Research Council, Washington, D.C., 1973, pp. 24-33.
5. J.D.C. Little and M.D. Kelson. Optimum Signal Timing for Arterial Signal Systems, Vol. 1: Summary Report. Operations Research Center, Massachusetts Institute of Technology, Cambridge, April 1980.
6. M.D. Kelson. Optimal Signal Timing for Arterial Signal Systems, Vol. 3: MAXBAND Programmer's Manual. Operations Research Center, Massachusetts Institute of Technology, Cambridge, May 1980.
7. D.I. Robertson. TRANSYT: A Traffic Network Study Tool. TRRL Report 253. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1969.
8. Transportation Research Center, University of Florida. TRANSYT-7F--User's Manual. FHWA, U.S. Department of Transportation, Feb. 1983.
9. J.D.C. Little, N.H. Gartner, and H. Gabby. MITROP: A Computer Program for Simultaneous Optimization of Offsets, Splits, and Cycle Time. Traffic Engineering and Control, Aug. 1976, pp. 355-362.
10. E.B. Lieberman and J.L. Woo. SIGOP II: A New Computer Program for Calculating Optimal Signal Timing Patterns. In Transportation Research Record 596, TRB, National Research Council, Washington, D.C., 1976, pp. 16-21.
11. R.O. Rogness. Possible PASSER II Enhancements. In Transportation Research Record 881, TRB, National Research Council, Washington, D.C., 1982, pp. 42-48.
12. C.E. Wallace and K.G. Courage. Arterial Progression--New Design Approach. In Transportation Research Record 881, TRB, National Research Council, Washington, D.C., 1982, pp. 53-59.
13. D.W. Whitehead, K. Ross, and B.D. Miller. The Use of Empirical Platoon Profiles to Minimize Signal Network Delay. Traffic Engineering and Control, Vol. 17, No. 1, Jan. 1976, pp. 10-13.
14. S.L. Cohen. Concurrent Use of the MAXBAND and TRANSYT Signal Timing Programs for Arterial Signal Optimization. In Transportation Research Record 906, TRB, National Research Council, Washington, D.C., 1983, pp. 81-84.
15. N.H. Gartner, J.D.C. Little, and H. Gabby. Optimization of Traffic Signal Settings by Mixed-Integer Linear Programming, Part I: The Network Coordination Problem. Transportation Science, Vol. 9, No. 4, Nov. 1975, pp. 321-342.
16. C.J. Messer, D.B. Fambro, and D.A. Anderson. Effects of Design on Operational Performance of Signal Systems. Research Report 203-2F. Texas Transportation Institute, Texas A&M University, College Station, Aug. 1975.
17. R. Akcelik. Time-Dependent Expressions for Delay, Stop Rate and Queue Length at Traffic Signals. Internal Report AIR 367-1. Australian Road Research Board, Numawading, Victoria, Australia, Oct. 1980.
18. L.P. Rach et al. Improved Operation of Urban Transportation Systems, Vol. 1: Traffic Signal Control Strategies--A State-of-the-Art. Corporation of Metropolitan Toronto, Ontario, Canada, March 1974.
19. F.A. Wagner, D.L. Gerlough, and F.C. Barnes. Improved Criteria for Traffic Signals at In-

dividual Intersections. NCHRP Report 32. HRB, National Research Council, Washington, D.C., 1967.
20. C.E. Wallace. Development of a Forward Link Opportunities Model for Optimization for Traffic Signal Progression on Arterial Highways. Ph.D. dissertation. University of Florida, Gainesville, 1979.
21. R.O. Rogness. Evaluation of a Heuristic Programming Approach to Arterial Street Signal Timing Operation. Ph.D. dissertation. Texas A&M University, College Station, 1981.
22. K.W. Huddart. The Importance of Stops in Traffic Signal Progression. Transportation Research, Vol. 3, No. 1, April 1969, pp. 143-150.
23. C.J. Messer, H.E. Haenel, and E.A. Koeppe. A Report on the User's Manual for Progression Analysis and Signal System Evaluation Routine--PASSER II. Research Report 165-14. Texas Transportation Institute, Texas A&M University, College Station, Aug. 1974.
24. C.J. Messer, D.B. Fambro, and E.C. Chang. Signal Timing Optimization to Maximize Traffic Flow--A 2-Day Workshop. Texas Transportation Institute, Texas A&M University, College Station, April 1983.
25. PASSER II-80 User's Manual. Texas State Department of Highways and Public Transportation, Austin, Dec. 1982.
26. W.D. Brooks. Vehicular Traffic Control, Designing Arterial Progression. IBM Document. Armonk, N.Y., 1966.
27. C.J. Messer and R. Roess. Signalized Intersections. Draft. NCHRP Project 3-28 to Update the 1965 Highway Capacity Manual. TRB, National Research Council, Washington, D.C., July 1984.
28. C.J. Messer and D.B. Fambro. A Guide for Designing and Operating Signalized Intersections in Texas. Research Report 203-1. Texas Transportation Institute, Texas A&M University, College Station, Aug. 1975.
29. F.V. Webster and B.M. Cobbe. Traffic Signals. Technical Paper 56. Road Research Laboratory, London, England, 1966.
30. J.C. Tanner. A Problem of Interface Between Two Queues. Biometrika, Vol. 40, 1953.
31. P.J. Tarnoff. The Results of FHWA Urban Traffic Control Research: An Interim Report. Traffic Engineering 45, April 1975, pp. 27-35.
32. E.C. Chang, C.J. Messer, and B.G. Marsden. Analysis of Reduced-Daily Optimization and Other Enhancements to PASSER II-80--PASSER II-84--Final Report. Research Report 375-1F. Texas Transportation Institute, Texas A&M University, College Station, April 1984.

The opinions and conclusions expressed are the authors' and are not necessarily those of Texas SDHPT and FHWA.

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

Abridgment

Minimum Delay Optimization of a Maximum Bandwidth Solution to Arterial Signal Timing

EDMOND CHIN-PING CHANG and CARROLL J. MESSER

ABSTRACT

This study indicated the advantages and drawbacks of combining the two major state-of-the-art traffic signal control strategies: the bandwidth maximization procedure and the delay minimization technique. The enhanced reduced-delay optimization model provided in PASSER II-84 guarantees minimum total arterial system delay within the slack-time allowance range of the original PASSER II-80 maximum progression solution. Modifications to the PASSER II signal timing plan for an arterial street system, using both a maximum bandwidth procedure and a minimum delay signal timing optimization algorithm, are evaluated. An efficient and usable delay-based search algorithm to assist traffic engineers in selecting a minimum delay arterial street signal timing plan that optimizes phasing sequence, cycle length, and offsets based on maximum bandwidth calculations in an urban network is demonstrated. The maximum bandwidth procedures are based mainly on calculations of distance, travel speed, and continuity of available green time for progressive movements without direct relationship to delay. The minimum delay algorithm minimizes total system delay endured by all traffic in the analysis network. Resulting offsets confirmed the feasibility of minimizing delay by the optimal offsets from the maximum bandwidth algorithm. When minimum delay and maximum progression are used, as calculated by the enhanced PASSER II-84, an improved level of service results thereby providing maximum progres-

sion and minimum total system delay within the offset slack-time range. This research has provided various insights into the operational characteristics of the enhanced PASSER II-84. It was found that PASSER II-84 consistently outperformed PASSER II-80. A consistent and satisfactory trend of delay reductions was found between PASSER II-84 and NETSIM evaluations. Recommendations are to implement the enhanced PASSER II-84 with possible field validation, to develop alternative strategies for allocating directional bandwidths, and to explore execution time and program efficiency.

This study evaluated the modifications to PASSER II-84 for an arterial street system using both a maximum bandwidth procedure and a minimum delay algorithm (1-6). The objectives were to devise a new delay optimization mechanism and to develop an experimental design for validating, evaluating, and comparing the enhanced PASSER II-84 and the existing PASSER II-80. Specifically, the study demonstrated an efficient and usable delay-based search algorithm to assist traffic engineers in selecting a minimum delay arterial traffic signal timing plan that optimizes the arterial offsets based on the optimal maximum bandwidth solution.

DELAY MINIMIZATION

A major task was to add a system offset optimization routine to the basic PASSER II program (1,7). The new extensions began by fine tuning the offsets from an existing progression solution using only straightforward approaches. The PASSER II-84 minimum delay optimization is designed to (8) minimize total arterial system delay as function of the offset at each signalized intersection. It is subject to (a) cycle length, (b) green split, (c) phase sequence, and (d) slack-time allowance of the optimal PASSER II-80 time-space diagram where the available slack-time allowance is the through green time, for a particular direction at each signal, that the offsets could be adjusted without losing the optimal progression bandwidth. This formulation is shown in Figure 1. At first, the input is given by geometric, traffic, and traffic signal control characteristics. The objective function is constructed according to constraints from the maximum bandwidth solution. Finally, the constrained optimum solution set is obtained by iteration using the modified system offset optimization-modified "sectioning" method (8).

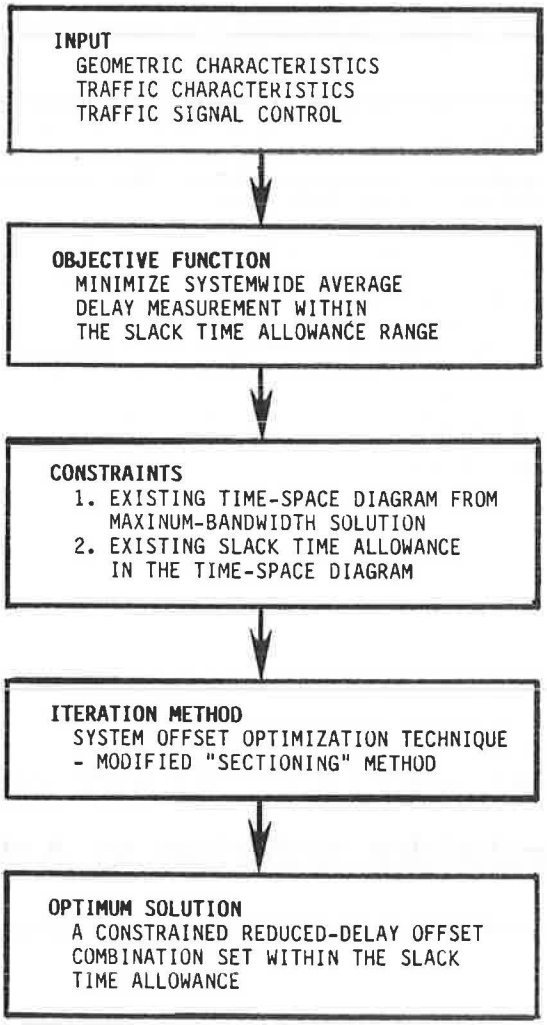


FIGURE 1 Formulation of constrained offset optimization problem.

PASSER II-84 OPERATIONS

Using the PASSER II-80 optimal time-space diagram as the starting solution and constraint, the PASSER II-84 optimization routine first identifies the offset slack-time allowance range to determine the possible optimum offsets for each intersection in the study arterial network. Then, the offset optimization algorithm starts a search within the slack-time allowance range for each intersection from the lowest possible optimum offset while keeping all the other intersection offsets constant. When the minimum arterial system delay is found within the slack-time allowance of a particular signal by simulating the system operations in PASSER II-84, the search will continue onto the next intersection until no further reduction of the total arterial system average delay value of this iteration can be found.

The major benefit of this systemwide offset optimization technique is that the objective function always remains the search for minimizing total arterial system average delay when performing the offset fine-tuning optimization within the slack-time allowance constraints. The optimization algorithm is constantly fixed at "system optimization," instead of at "local optimization" from link-to-link delay-offset analysis.

EVALUATION

Because actual traffic fluctuates, the NETSIM simulation program (9) was selected for its ability and complexity in changing traffic signal timing to achieve specific experimental conditions for field validation. As shown in Figure 2, the test procedure followed a straightforward analysis. First, the experimental plan was developed. Then, simulation techniques were used to enumerate the results of

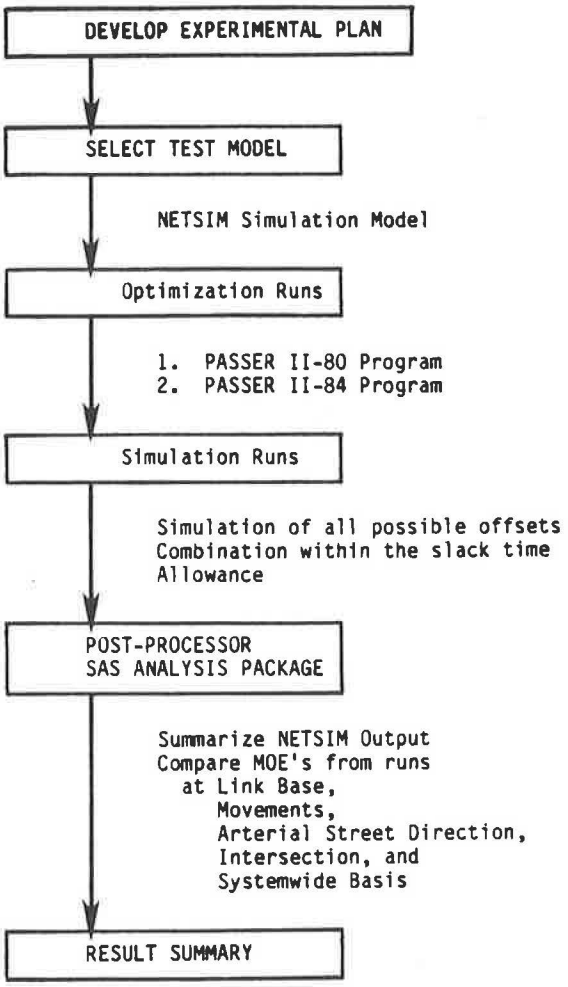


FIGURE 2 Evaluation of solution methodology.

arterial traffic signal systems. By measuring the systemwide measure of effectiveness (MOE), mainly delay and stops, at the end of the simulation study period, the operational performance of traffic on each segment in the analysis network can be evaluated. The simulation at first evaluated the enhanced offset fine-tuning optimization by analyzing one detailed test case as described by Rogness (5). A subsequent analysis was performed to evaluate PASSER II-80 and PASSER II-84 using 13 test scenarios under good, fair, and poor progression operations (10,11). This was done to answer the following questions:

1. Did the reduced-delay system optimization algorithm find the minimum delay offsets within the slack-time allowance?
2. On the basis of the PASSER II-84 evaluation, did the PASSER II-84 model improve on the systemwide delay and stops of the existing PASSER II-80 model?
3. On the basis of the NETSIM evaluation, did the PASSER II-84 model improve on the existing PASSER II-80 model?

Link Delay-Offset Analysis

An example of the NETSIM simulation analysis (summarized by SAS) for comparing the measures of effectiveness from different offset combinations either inside or outside the slack-time allowance ranges is shown in Figure 3. The horizontal axis is the offset value for intersection 2, in seconds. The vertical axis represents the average total systemwide delay per vehicle, in seconds per vehicle. The two vertical lines mark the range of slack-time allowance for intersection 2. The PASSER II-80 solution, as simulated by NETSIM, is labeled "OLD" and the enhanced PASSER II-84 solution is represented as "NEW." Figure 3 shows the average delay versus relative offsets at intersection 2. The figure also shows the sensitive change of average delay or stop performance with respect to any of the three offsets of intersections 2, 3, and 4 in the systemwide delay-offset response space. It could be considered as slicing the objective performance curve by "sectioning" at that particular intersection offset.

Each small dot in the diagram represents one individual NETSIM run for a given cycle length and progression phase sequence and their effects--either inside or outside the allowable slack-time range for each traffic signal. The feasible solution sets with all the intersection offsets inside the slack-time allowance ranges are represented by the solid boxes. These 310 NETSIM simulation runs evaluated the possibly feasible or infeasible solutions that may be generated through existing PASSER II-80, enhanced PASSER II-84, or any other traffic signal timing optimization program.

The solid curved line is the approximated regression line of the relationships between the total systemwide average delay and the offset value of intersection 2 in the Skillman Avenue network. The numerical value on the curve is calculated by averaging the amount of average delay per vehicle with respect to the different offset combinations having that particular offset value on intersection 2. The dashed curved lines delineate the 95 percent confidence limits for mean prediction of the average delay per vehicle as estimated by the SAS package.

The optimum offset obtained by this constrained reduced-delay offset optimization is represented by the offset combination with the offset at that intersection of that value. The optimum offset set, in this particular case, moved from the PASSER II-80 solution of 26, 85, and 49 sec to the PASSER II-84 solution of 31, 83, and 49 sec for offsets at intersections 2, 3, and 4, respectively. Applying a similar analogical relationship, the NETSIM simulation results were evaluated in terms of the average stops per vehicle versus the relative offsets for intersections 2, 3, and 4.

For this particular problem, the constrained optimum offset set was the same whether minimum delay or minimum stop was used as the selection criterion. Both the systemwide delay and the stop measurements of the optimum solution obtained from the maximum bandwidth procedure by PASSER II-80 were improved by the reduced-delay fine-tuning optimization technique in the PASSER II-84 model. Not only was the systemwide average delay per vehicle reduced to the minimum value within all the feasible offset solutions, but the average stops per vehicle were also reduced to the minimum value among all the possible offset solutions either inside or outside the slack-time allowance ranges.

As shown in Figure 3, both the PASSER II-80 and the PASSER II-84 solutions have a lower MOE than do arbitrary offset searches within the slack-time allowance range. This comparison indicated the superiority of offset optimization as calculated by PASSER II-80 and PASSER II-84 using the offset fine-tuning optimization algorithm in the enhanced PASSER II-84.

Even though absolute magnitude differences did exist between the measures of effectiveness of PASSER II-80, PASSER II-84, and NETSIM, the evaluation based on the NETSIM simulation provided a

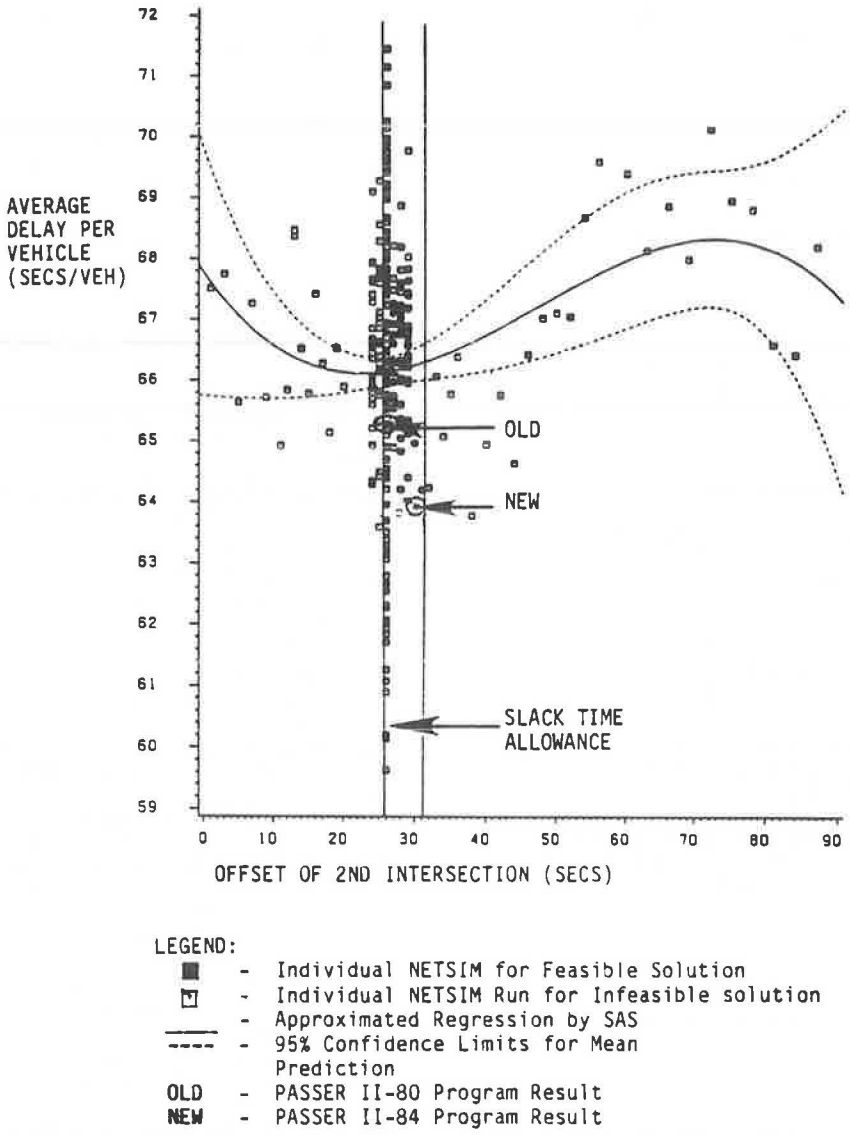


FIGURE 3 Summary of NETSIM simulation results—average delay per vehicle versus relative offset at intersection 2.

favorable examination of the system offset fine-tuning optimization algorithm developed for PASSER II-84. Several interesting results can be observed:

1. The reduction of average delay from PASSER II-80 (OLD OFFSET) to PASSER II-84 (NEW OFFSET) suggested that the offset fine-tuning optimization did reduce delay based on the original progression settings by the system offset optimization algorithm in PASSER II-84.
2. The minimum systemwide delay and stops may not always exist within the slack-time allowance range; they sometimes exist outside the allowable slack-time ranges based on the NETSIM evaluation.
3. The wide range of resultant total system MOEs from the NETSIM simulation analysis indicated sensitive interactions among the individual traffic signal timing parameters and the systemwide MOEs.
4. Difficulties exist in predicting the trend and magnitude of influence on the total systemwide MOEs if based only on the relative amount of change in the traffic signal offsets at any individual intersection.
5. Further reduction of average arterial system delay from the enhanced PASSER II-84 could possibly be obtained by moving to another offset combination outside the feasible slack-time allowance range.

Overall, the reduced-delay system offset optimization in PASSER II-84 found the signal offset solution with the minimum systemwide delay within the slack-time allowance of PASSER II-80 while not changing the optimum progression solution. In the next section the most significant results of the reduced-delay fine-tuning offset optimization in PASSER II-84, according to both PASSER II-84 and NETSIM evaluations, are summarized.

Performance Analysis

The NETSIM evaluation examined the robustness of the reduced-delay optimization under various test scenarios having different progression quality. To permit some range of progression efficiency while confining the number of alternatives, three spacings (full scale, half scale, and quarter scale) of the arterial were selected. The three cycle lengths

selected were 80, 90, and 100 sec, which appeared to be representative and still provided a nominal range of solutions. These NETSIM comparison runs were enumerated under 13 test scenarios, including the basic test case, with progression efficiency values ranging from 0.15 (poor progression) to 0.25 (fair progression), to 0.38 (good progression). These test arterial systems were selected so that the evaluation of the system offset optimization could be tested under various progression conditions. Figure 4 shows the systemwide average delay comparisons of PASSER II-80 on the horizontal axis versus PASSER II-84 on the vertical axis. Both PASSER II-84 and NETSIM evaluations are included. The consistent performance of PASSER II-84 over PASSER II-80 is identified by the locations of all the PASSER and NETSIM data on the lower side of the 45-degree line, which means that the enhanced PASSER II-84 calculated a consistently equal or lower delay than was calculated by the existing PASSER II-80. This result indicated that in some cases, even though the PASSER II-84 offset optimization could not find a lower system delay solution due to the limited slack time, it would still provide the original PASSER II-80 solution instead of a worse solution.

As shown in Figure 4, there are four possible results that both PASSER II-84 and NETSIM programs could possibly generate from PASSER II-84 and PASSER II-80 solutions. Quadrant I indicates that with both PASSER II-84 and NETSIM evaluations the average delay was reduced using PASSER II-84 compared with PASSER II-80; Quadrant III shows both PASSER II-84 and NETSIM evaluations and indicates that average delay was increased due to the use of PASSER II-84 for the PASSER II-80 original solution. Quadrants II and IV show the possible inconsistent results as evaluated by the PASSER II-84 and NETSIM programs. This figure indicates that the PASSER II-84 MOE provided a slightly higher estimation of the percentage of average delay reduction than did NETSIM but in an extremely consistent manner throughout all 13 test scenarios.

Summary

More than 310 NETSIM runs were made using the Skillman Avenue test bed. The detailed NETSIM analysis on this test case indicated that a reduced-delay solution could be found by PASSER II-84 among all the possible solutions including existing PASSER II-80, enhanced PASSER II-84, or any other signal timing program. The overall NETSIM simulation indicated that the system offset fine-tuning optimization can provide reduced delay and stops within the slack-time allowance range under all 13 NETSIM simulation test scenarios by trend analysis, binomial test, and paired t-tests. Figure 5 shows a summary of the percentage of systemwide delay reduction using the enhanced PASSER II-84 model with respect to the various progression efficiencies of the arterial progression systems used. This provided a basic indicator for estimating the possible systemwide delay reduction when the decision had to be made whether to implement the signal timing plan from the existing PASSER II-80 solutions or from the enhanced PASSER II-84 solutions. However, no specific trend of percentage of delay reductions versus progression efficiency could be observed.

The NETSIM analysis of average delay and stops on the link-to-link basis and total arterial travel direction indicated mixed results. That is, when the total arterial system delay is fine tuned based on the PASSER II-84 progression solution, the delay measurement may decrease on some links and increase on other links. The NETSIM analysis also indicated the difference that exists in total arterial system delay and delay incurred just in the arterial travel directions.

FINDINGS AND RECOMMENDATIONS

The maximum bandwidth procedures are based mainly on calculations of distance, travel speed, and continuity of available green time for progressive

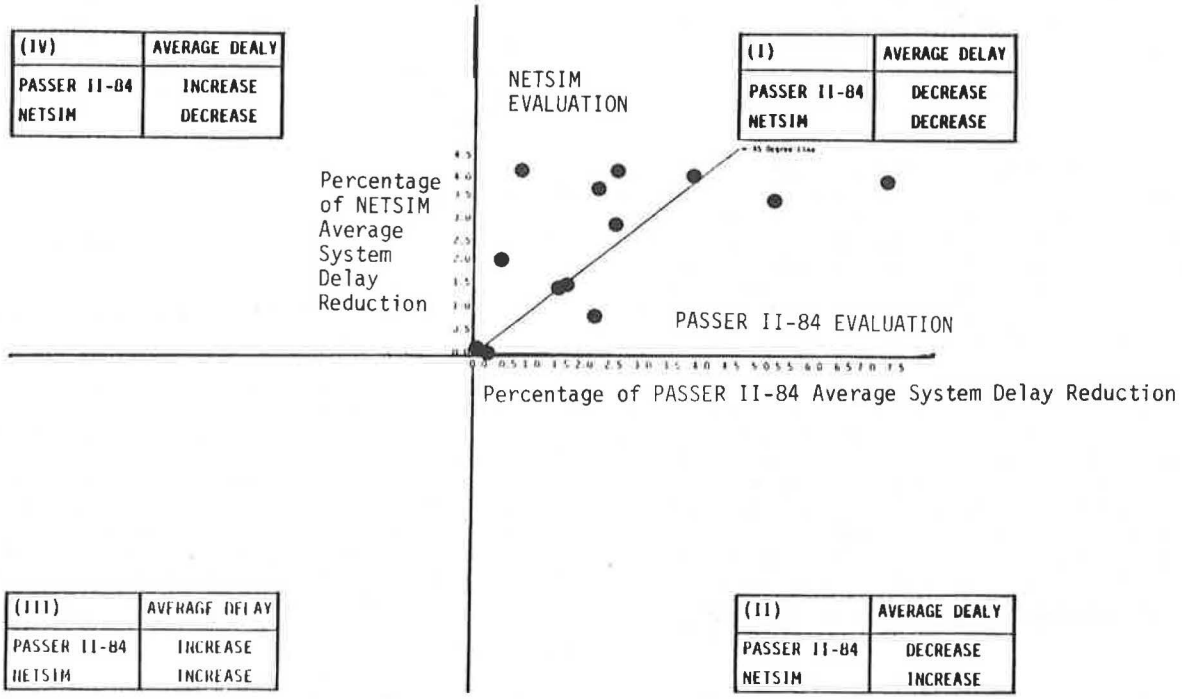


FIGURE 4 Comparisons of percentage change of average system delay in PASSER II-84 solutions and PASSER II-80 solutions as evaluated by PASSER II-84 and NETSIM programs.

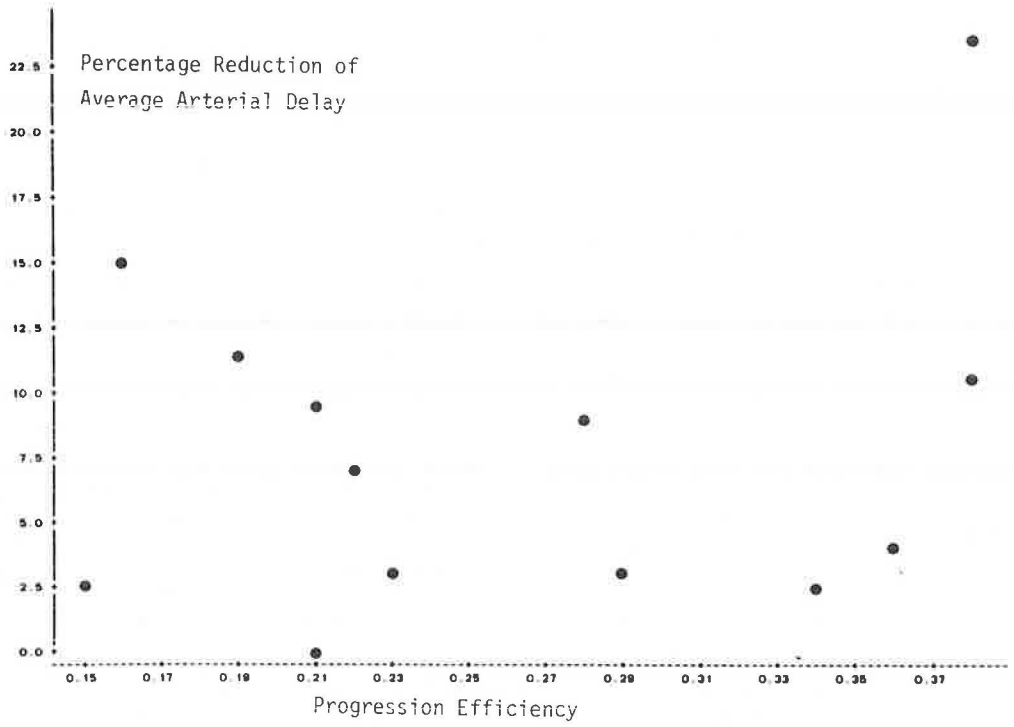


FIGURE 5 Comparisons of reductions in NETSIM average arterial delay per vehicle of PASSER II-84 solutions over PASSER II-80 solutions versus progression efficiency.

movements without direct relationship to delay. The minimum delay algorithm minimizes total system delay endured by all traffic in the analysis network whether there are progressive movements or not. The overall result offsets confirmed the feasibility of minimizing delay by the optimal offsets from the maximum bandwidth algorithm. When minimum delay and maximum progression are used, as calculated by the enhanced PASSER II-84, an improved level of service results thereby providing maximum progression and minimum total system delay within the offset slack-time range.

Findings

This study provided various insights into the operational characteristics of the enhanced PASSER II-84. It indicated the advantages and drawbacks of combining the state-of-the-art bandwidth maximization procedure and the delay minimization technique. However, the general improvement by PASSER II-84 relies primarily on the quality of the original PASSER II-80 solution. Several findings were obtained from this research:

1. PASSER II-84 consistently outperformed PASSER II-80 on the basis of two evaluations of delay as the measure of effectiveness. These delay evaluations were performed by the deterministic delay simulator found in PASSER II and by the unbiased microscopic delay simulator in NETSIM.
2. Consistent and satisfactory trends in the prediction of delay were found between PASSER II-84 and NETSIM. PASSER II-84 predictions of delay reduction, however, were somewhat higher than those derived from NETSIM.
3. PASSER II-84 solutions reduced total system delay from 0 to 4 percent for the 13 test scenarios. Arterial movement delays were reduced from 0 to 23 percent. These findings are based on NETSIM simulation using PASSER II-80 progression solutions as the base condition.
4. No apparent correlation was found between average arterial system delay reduction and progression efficiency. The amount of delay reduction possible depends heavily on the available slack-time allowances and on the efficient use of green time for progressive movements.
5. Instead of using the total directional traffic volume ratio and minimum green time constraint alone to provide the best directional bandwidth weighting, an optimization search outside the existing slack-time allowance ranges can further reduce the total arterial system delay by slightly varying the optimal progression solution.

Recommendations

As a result of this research, five recommendations are made:

1. The enhanced PASSER II-84 model should be implemented to further improve arterial traffic signal system operations.
2. Field validation of the reduced-delay offset-optimization algorithm to confirm the benefits yielded by the enhanced PASSER II-84 should be undertaken.
3. If minimum system delay is the most important objective for a particular arterial traffic signal system, the optimization search outside the existing slack-time allowance ranges is suggested. This is done by slightly varying the optimal progression solution to further reduce total system delay.
4. Alternative strategies should be developed for allocating the directional bandwidths to determine the best directional bandwidth weighting for minimizing total system delay.
5. This study did not explore the question of execution time and program efficiency. Benchmark

tests to examine the run time efficiency of the enhanced PASSER II-84 versus the existing PASSER II-80 on different computer systems are recommended. Developing possibly more efficient program architecture for PASSER II-84 to optimize calculation and eliminate duplicated FORTRAN coding is also recommended.

ACKNOWLEDGMENTS

This research was supported by the Texas State Department of Highways and Public Transportation; FHWA, U.S. Department of Transportation; and the Texas Transportation Institute, the Civil Engineering Department, and the Texas Engineering Experimental Station at Texas A&M University.

REFERENCES

1. E.C. Chang, C.J. Messer, and B.G. Marsden. Analysis of Reduced Delay and Other Enhancements to PASSER II-80-PASSER II-84--Final Report. Research Report 375-1F. Texas Transportation Institute, Texas A&M University, College Station, April 1984.
2. C.J. Messer, K.H. Whitson, C.L. Dudek, and E.J. Romano. A Variable Sequence Multiphase Progression Optimization Program. In Highway Research Record 445, HRB, National Research Council, Washington, D.C., 1973, pp. 24-33.
3. J.D.C. Little and M.D. Kelson. Optimum Signal Timing for Arterial Signal Systems: Vol. 1, Summary Report. Operations Research Center, Massachusetts Institute of Technology, Cambridge, April 1980.
4. C.E. Wallace. Development of a Forward Link Opportunities Model for Optimization for Traffic Signal Progression on Arterial Highways. Ph.D. dissertation. University of Florida, Gainesville, 1979.
5. R.O. Rogness. Evaluation of a Heuristic Programming Approach to Arterial Street Signal Timing Operation. Ph.D. dissertation. Texas A&M University, College Station, 1981.
6. S.L. Cohen. Concurrent Use of the MAXBAND and TRANSYT Signal Timing Programs for Arterial Signal Optimization. In Transportation Research Record 906, TRB, National Research Council, Washington, D.C., 1983, pp. 81-84.
7. PASSER II-80 User's Manual. Texas State Department of Highways and Public Transportation, Austin, Dec. 1982.
8. A. Glankwahmdee, J.S. Lieberman, and G.L. Hogg. Unconstrained Discrete Nonlinear Programming. Engineering Optimization, Vol. 4, 1979, pp. 95-107.
9. Traffic Network Analysis with NETSIM--A User Guide. FHWA Implementation Package FHWA-IP-80-3. FHWA, U.S. Department of Transportation, Jan. 1980.
10. K.G. Baass and B. Allard. Description of A Combined Approach for Arterial Signal Coordination Considering Bandwidth and Delays. In Transportation Research Record 957, TRB, National Research Council, Washington, D.C., 1984, pp. 32-46.
11. E.C. Chang. A Minimum Delay Approach To Maximum Bandwidth Optimization on Arterial Traffic Signal Timing. Ph.D. dissertation. Texas A&M University, College Station, 1984.

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

Analysis of Traffic Network Flow Relations and Two-Fluid Model Parameter Sensitivity

JAMES C. WILLIAMS, HANI S. MAHMASSANI, and ROBERT HERMAN

ABSTRACT

Presented in this paper is a systematic exploration, using microscopic simulation, of the sensitivity of network-level traffic flow descriptors and relationships, particularly those of the two-fluid theory of town traffic, to network features, traffic control, and traffic-interfering urban activity levels. Moving traffic interference, which is represented by stochastic short-term lane blockages of varying duration and frequency, is shown to be a key determinant of the traffic character of an urban street network and of the behavior described by the two-fluid theory and verified operationally. In addition, the sensitivity of the two-fluid model parameters to a change in traffic control strategy, in this case the coordination of signals to achieve progression, is demonstrated. Furthermore, keeping the same network configuration, the effect of network topology on traffic flow is examined by changing the identical length of the links.

The importance of characterizing traffic in a city network stems from the need to evaluate and compare control strategies in terms of their overall network impact, to assess the quality of operations and identify deficiencies in a given network, and to monitor the level of traffic service over time in a designated area. The complex interactions taking place in a traffic network effectively preclude the analytic derivation of network-level macroscopic relationships from the basic principles governing microscopic traffic behavior or from link-level macroscopic models. In addition, field validation of postulated theories requires considerable and often elaborate data-gathering efforts.

In a previous paper (1), Williams et al. established the feasibility of using microscopic simulation to study network-level traffic flow relationships. Computer simulation offers the opportunity to conduct systematic analyses of network traffic performance under controlled experimental conditions. In the previous effort, the NETSIM model was employed to examine the variation of mean speed with concentration in a closed test network, as well as to establish and verify relationships among certain fundamental network-level traffic flow descriptors. In addition, the sensitivity of network performance to various control and operational features was demonstrated. Furthermore, a preliminary investigation of certain aspects of Herman and Prigogine's two-fluid theory of town traffic (2) was conducted, particularly with regard to the dependence of the fraction of vehicles stopped in a network on the concentration in that network (1).

The two-fluid theory of town traffic provides a useful framework within which to attempt to characterize with a small number of parameters for traffic flow in a network. The two "fluids" consist of vehicles moving in the system and of stopped vehicles (still in the traffic stream, such as at traffic signals, as opposed to parked vehicles). The theory leads to a relatively simple relationship between the average running time and the average stopped time in a network (2). Validation studies, relying on chase-car techniques and, more recently, aerial photographs, have been ongoing for the past 5 years and have firmly established the basic underlying premises of the theory under operating conditions found in actual city networks (3-6).

In addition to the description of traffic flow in a given network under a particular traffic control scheme and usage pattern, it would be of considerable interest and practical importance to understand the mechanisms and factors affecting network-level performance characteristics such as those proposed by the two-fluid theory. Conducting this type of investigation with exclusive reliance on the observation of actual systems is clearly prohibitive and unpractical, given the need for extensive data from a large cross section of actual networks and operating conditions. As noted earlier, microscopic simulation has already yielded useful insights into these phenomena (1).

Presented here is a systematic exploration, using NETSIM, of the sensitivity of network traffic descriptors and relationships, particularly those of the two-fluid theory, to network topology, traffic control, and an urban area's "activity level" (events that interfere with the flow of traffic). The results presented contribute to two interrelated fundamental research objectives. The first is to determine how to capture, in a microscopic simulation model such as NETSIM, the fundamental character of an urban network, including the often intense levels of interfering activities. The second objective is the identification and ultimate representation of the effect of key traffic mechanisms and of a network's physical and operating features on network-level macroscopic descriptors and relations.

The relevant conceptual background is briefly presented in the next section, with particular emphasis on the key traffic flow descriptors used in the analysis as well as in the two-fluid model parameters. The basic structure of the simulation experiments is presented in the third section, along with a description of the features of the test network common to all the simulations performed in this study. The fourth section is a description of how interference with flow, which is an inherent characteristic of traffic in city networks, is captured using the "short-term events" capability in NETSIM. A systematic analysis of the effect of this mechanism for representing interfering urban activity levels on the two-fluid model parameters and their descriptors is presented in the fourth section. Further investigation of the sensitivity of network-level traffic flow relationships to network topology (block length) and traffic signal coordination is discussed in the fifth section. Concluding comments, including some implications for city traffic network operations and future research, are presented in the final section.

REVIEW OF RELATED CONCEPTS

The principal network-level traffic flow descriptors and their interrelation are described elsewhere (1). A brief review of concepts relevant to the present paper is presented in this section. These include the definition of average speed and concentration at the network level, the principal variables addressed in the two-fluid theory and the interpretation of its parameters, and the related variation of the average fraction of vehicles stopped in a network as a function of concentration.

Speed-Concentration Relation

The three fundamental traffic variables speed, concentration, and flow, extensively studied in the context of flow along arterials or through intersections, have been generalized to the network level elsewhere (1). Of interest here is the variation of average network speed (V) with average concentration (K). These averages are taken both over time and over all vehicles in the network. Average speed, in miles per hour, is thus taken as the ratio of total vehicle-miles traveled to total vehicle-hours on the network during a given observation period. The average concentration, for the same time period, consists of the time average of the number of vehicles per unit lane-length in the system. One element of control exercised in these experiments is to maintain concentration constant throughout any given observation period in order to avoid confounding effects due to varying concentrations with those under investigation.

Previous simulations (1) clearly indicated that V is a decreasing function of K at the network level also, as would be expected from the well-known speed-concentration behavior on arterials. Furthermore, the sensitivity of the K-V relationship to traffic control and usage patterns in the network was also observed in earlier simulations (1), suggesting that this relationship can provide a good indicator of a network's performance under varying network physical and operating characteristics. However, the way in which this relationship is affected by the various factors of interest remains to be established. The present computer experiments provide important insights into this relationship.

Two-Fluid Theory of Town Traffic

In their kinetic theory of multilane highway traffic, Prigogine and Herman (7) recognized the existence of two distinct traffic flow regimes: the individual and the collective flow regimes, which are a function of vehicle concentration. When concentration rises so that traffic is in the collective flow regime, the flow pattern becomes largely independent of the will of individual drivers. The two-fluid model (2) views traffic in this latter regime, in an urban network, as two "fluids": one consisting of moving vehicles and another of stopped vehicles, which remain in the traffic stream.

The basic postulate of the two-fluid theory states that V_r, the average speed of the moving vehicles, is related to f_s, the fraction of vehicles stopped, in the following manner:

$$V_r = V_m(1 - f_s)^n \quad (1)$$

where V_m is the average maximum running speed in the network and n is a parameter that has been found to be a useful indicator of the quality of traffic service in a network (4,5).

This postulate leads to a relationship between three principal variables, T, T_r, and T_s, respectively, average travel time, average running (or moving) time, and average stopped time, all per unit distance, of the following form (2):

$$T_s = T - T_m^{[1/(n+1)]}T^{[n/(n+1)]} \quad (2)$$

where T_m is a parameter of the model equal to V_m^{-1} and thus reflecting the minimum travel time per unit distance in the network under free-flow conditions.

The calibration of the model parameters T_m and n (both of which were found to be robust characteristics of a given network) to a particular city has been performed empirically using trip time and stop time information obtained by one or more circulating vehicles "sampling" the network according to the chase-car technique (4,5). From Equation 2, T_r is given by

$$T_r = T_m^{[1/(n+1)]}T^{[n/(n+1)]} \quad (3)$$

In simulation experiments, T_r and T can be calculated over any desired observation period and obtained from the output of the simulation package (1). The two parameters T_m and n can then be found by simple linear regression using a log transformation of Equation 3:

$$\ln T_r = [1/(n+1)] \ln T_m + [n/(n+1)] \ln T \quad (4)$$

The network parameter n can be viewed as an approximate measure of the slope of the T-versus-T_s relation (2). If $n = 0$, $T_r = T_m$ (see Equation 3), and trip time would increase at the same rate as stop time. If $n > 0$, trip time increases at a faster rate than stop time, meaning that running time (T_r) is also increasing. Intuitively, n would be expected to be greater than zero because the usual cause for increased stop time is heavier congestion, which generally results in vehicles moving at lower speeds, which implies higher average running time per unit distance.

The two parameters T_m and n, as well as the interrelation between T_r, T_s, and T and their dependence on network concentration, are important performance indicators of the effect of the various network and traffic control features of interest.

Another important quantity in the two-fluid conceptualization of traffic is f_s, the average fraction of vehicles stopped in the network, taken over a given observation period. This quantity was postulated to be an increasing function of network concentration (2), a result that was later verified (1). The variation of f_s with K and the factors that affect this variation are given particular consideration in these simulation experiments. Note in this regard that the mean fraction of time stopped (T_s/T) is used here in lieu of f_s because it is more readily calculated from the simulation output. The equality of f_s and T_s/T in a closed network has been shown theoretically (3) and verified by simulation (1).

EXPERIMENTAL DESIGN AND DESCRIPTION OF BASE CASE

In this section, the structure of the simulation experiments, as well as the characteristics of the network and traffic control scheme specified for the base condition used in this paper, is described. Although the same basic network configuration was used in all simulations, three different factors were varied in accordance with the experimental design. These factors are (a) level of traffic-interfering urban activity, (b) network topology, and (c) traffic control scheme. For each combination of factor levels, a set of five runs was performed, each run corresponding to a different network concentration level. The approximate concentration levels used within each set were 8, 15, 30, 45, and 60 vehicles per lane-mile. Note that, in a given simulation run, the interest is in the network-level properties of a fixed number of vehicles circulating in the specified closed system, thereby maintaining a constant concentration in the network throughout the simulation period.

A series of sets of simulation runs was first conducted to analyze the effect of the first factor, namely the level of interfering activity. Next, a representative activity level was selected for use in the reference base case and a second series of sets of runs was performed with each set corresponding to a change in one of the other two experimental factors of interest.

The NETSIM model, which was used to perform these experiments, is a fixed-step, microscopic, network traffic simulation model. Each vehicle in the system is treated separately during the simulation; its behavior is governed by a set of microscopic car-following, queue-discharge, and lane-switching rules (8,9). The feasibility of using the NETSIM model in this type of study was established previously (1).

Further details on the simulation experiments are presented hereafter; particulars concerning changes in activity levels and other network and traffic factors are presented in the next two sections, along with the results of these changes.

Network Configuration and Geometric Features

As noted in the previous section, a degree of regularity and uniformity was sought in the test network. This network consists of 25 nodes, arranged in a 5-node by 5-node square, connected by two-way, four-lane streets forming a regular, central business district (CBD)-like grid. Because only directed links can be used in representing the network in NETSIM (i.e., all links are one way), there are 80 one-way, two-lane links, as shown in Figure 1. Each link (block) is 1,000 ft long, with no right- or left-turn bays, and all grades are zero.

Vehicles are injected onto the network via 12 entry links placed around the perimeter, three to a side, with each entry link connecting a source node (source nodes are labeled 801 to 812 in Figure 1) to

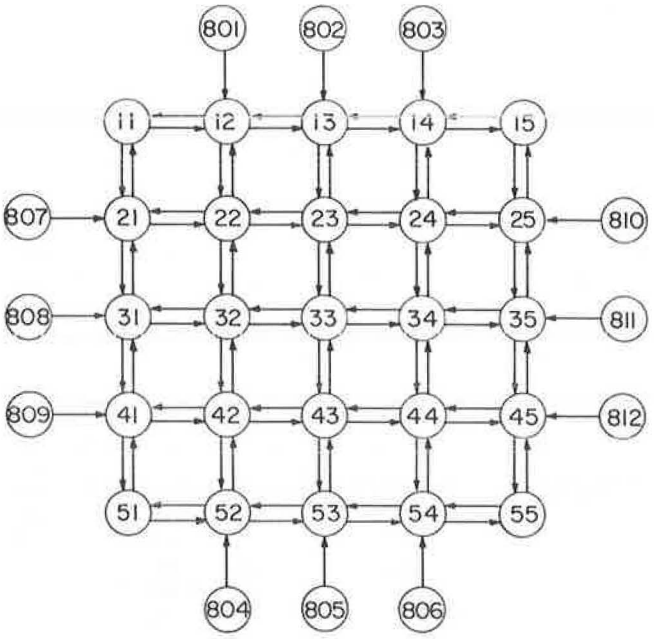


FIGURE 1 Network configuration.

a noncorner boundary node. No sink nodes have been designated because a closed system, where vehicles remain in the network once they have entered it, is under consideration.

Traffic Characteristics and Control Strategy

The mean desired speed for all the simulations was 35 mph, the desired speed of a given driver being the speed at which he would travel in the absence of other vehicles and of traffic controls. The model assigns desired speeds to vehicles as they enter the system according to a distribution about the specified desired speed. Likewise, the same vehicle-turning movements were used throughout the simulation runs: 10 percent of the vehicles turned left, 15 percent turned right, and 75 percent continued straight through at each interior intersection, whereas at the boundary nodes traffic split equally between the two available options. Fixed-time traffic signals were placed at all but the four corner nodes. At the interior nodes, two-phase timing with a 50-50 split and no protected turning movements was used. Three-phase signals were used at the boundary intersections, providing a protected left turn for vehicles reentering the interior of the system. The time was nearly equally allocated between the vehicles leaving the interior of the network and those reentering it from the boundary. Two-way progression at the mean desired speed was provided along the interior arterials by using a 40-sec cycle length with single alternate operation (offsets between adjacent signals were all 50 percent of the cycle length). Application of Webster's equation (10,p.57) indicated that this cycle length was adequate to handle the traffic volumes encountered in the various simulation runs.

There were no pedestrians and right turn on red was allowed at all intersections in all simulations. Additional details of the conduct of individual runs are presented next.

Individual Runs

A start-up period ranging from 5 to 15 min was used in all runs, during which vehicles were generated uniformly on the 12 entry links. The length of the start-up period depended on the number of vehicles necessary to achieve the desired concentration level in the network. Vehicles were injected directly into the network interior by not allowing turns onto the boundary from the entry links. The vehicles were then allowed to circulate in the network for the desired observation period of 15 min. Intermediate output was printed every minute, providing a "snapshot" of each link's condition (at that time) along with some cumulative information about each link. Network-level information, as well as additional cumulative link data, was printed every 3 min during the simulation period.

As mentioned earlier in this section, the vehicle concentration ranged from 8 to 60 vehicles per lane-mile within each set of runs. Specific changes from the base case and the results of these changes are analyzed in the next two sections, starting with the traffic-interfering activity factor.

ACTIVITY LEVEL EFFECTS

Considered here is the effect of differing levels of traffic-interfering activity in the test network on the quality of traffic service, as described by the network-level quantities and relations presented in the second section. "Activity level" refers to the degree of interference experienced by moving vehicles. Such interference may be due to intralink perturbations such as vehicles stopping to pick up or drop off passengers or goods, illegal parking and related maneuvers, pedestrian activity (other than that at intersections), and many other similar activities that are an inherent feature of a city street network. The NETSIM model provides for activities of this type with a "short-term events" option, whereby the user may specify, by link, the mean duration and the frequency of the events. In the model these are introduced stochastically and independently as blockages in the right lane in the center of the link. In the experiments described herein, short-term events were identically specified for all 80 links in the test network at the six levels given in Table 1. The first column gives the level designation; the second, the mean duration of each event; the third, the mean time interval between the beginning of successive events on a link; and the fourth, the fraction of time the right lane of a link is blocked by an event, which is the duration (in column 2) divided by the time between events (in column 3). Note that Level B is the one specified in the base case mentioned in the previous section. For each of these six activity levels, a set of five runs at varying concentrations was performed with all other factors remaining the same.

TABLE 1 Activity Levels

Activity Level	Mean Event Duration (sec)	Time Between Events (sec)	Fraction of Time Lane is Blocked
0	0		0.000
A	30	240	0.125
B	45	120	0.375
C	14	36	0.389
D	30	36	0.833
E	30	120	0.250

Single values of T and T_r were calculated from the final 12 min of each simulation run, resulting in five pairs of data for each set of runs. The

values of the two-fluid parameters T_m and n were then obtained as described in the second section. The results are given in Table 2 along with the corresponding r^2, using only the lowest four of the five concentration levels. The omission of the highest concentration level in the estimation of T_m and n is based on a comparison of estimates with and without this highest point, which revealed substantial discrepancies in the magnitudes of the parameter estimates. Further examination of the variation of T_r with concentration, shown in Figure 2, revealed clearly that, at extremely high concentrations, T_r exhibits as substantially lower rate of increase than at lesser, and more practically meaningful, concentration levels. This may mean that these are limiting cases of the two-fluid theory in that they go to concentration levels that are clearly beyond the range encountered in previous supporting empirical work.

The effects of these activity levels on the network performance parameters can be analyzed in three ways: (a) as a function of the mean event duration for a given mean interevent time, (b) as a function of the interevent time for a given mean duration, and (c) as a function of the fraction of time that the right lane of each link is blocked, thus allowing both the mean event duration and interevent time to vary.

TABLE 2 Estimates of the Two-Fluid Parameters T_m and n

Activity Level	n	T_m	r^2
0	0.076	2.238	0.947
A	0.338	2.196	0.964
B	0.845	2.135	0.973
C	0.784	2.123	0.982
D	1.738	1.997	0.993
E	0.573	2.173	0.966

Effect of Mean Event Duration

Two groups of runs can be distinguished: Activity Levels 0, C, and D (with a mean interevent time of 36 sec) and 0, E, and B (with a mean interevent time of 120 sec). Average speed is plotted against concentration in Figure 3(a) and fraction of vehicles stopped (f_s or T_s/T) is plotted against concentration in Figure 3(b) for Activity Levels 0, C, and D. As the mean event duration increases from 0 to 14 to 30 sec, both the average speed and the fraction of vehicles stopped decrease. Figure 3(a) also reveals that, at the lowest concentration considered, the average speeds are nearly identical for the three duration levels because of the ability of vehicles to get around the right lane blockages in the absence of significant competing traffic. As concentration increases, this ability is substantially hampered, resulting in a decrease of the average speed with increasing mean event duration. It can also be noted that, at higher concentrations, the speed differentials across mean event durations tend to decrease, as typified by the speed values for $K \approx 60$ vehicles per lane-mile in Figure 3(a). This is because, at higher concentrations, the fraction of time vehicles are stopped or slowed due to short-term events becomes smaller relative to the fraction of time they are stopped or slowed for other reasons, namely queueing at traffic signals and overall congestion.

Figure 3(b) reveals that, as expected, f_s increases with concentration, though at a slower rate with increasing mean event duration. At the lowest concentration, f_s is nearly identical for all

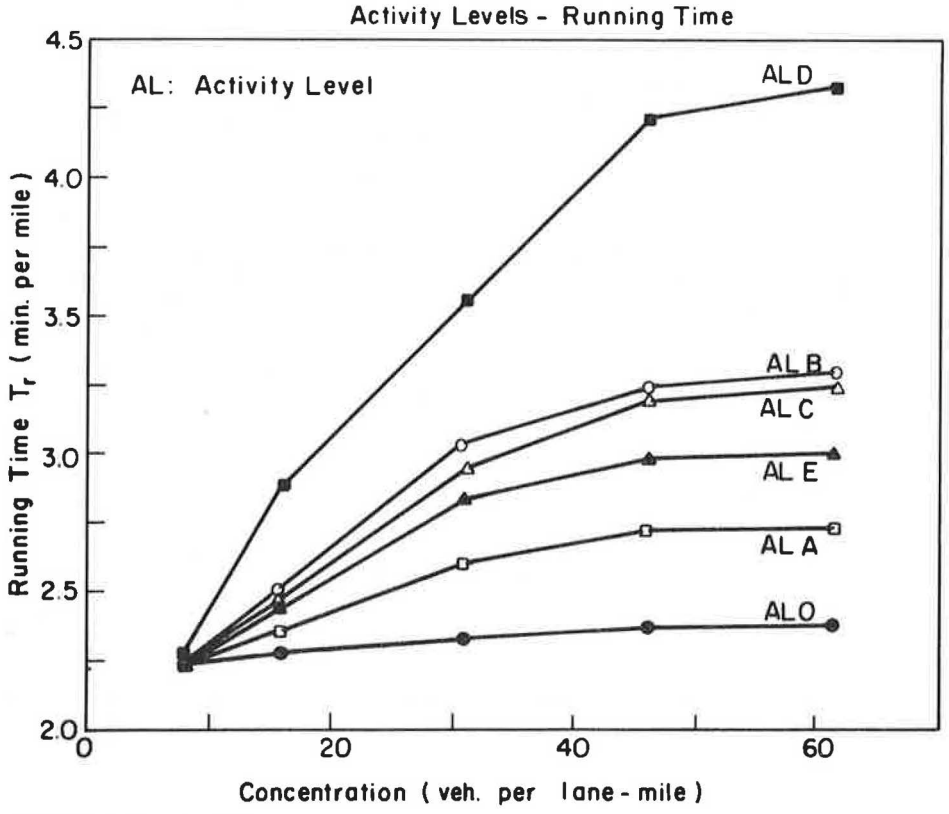


FIGURE 2 Running time versus concentration: activity levels.

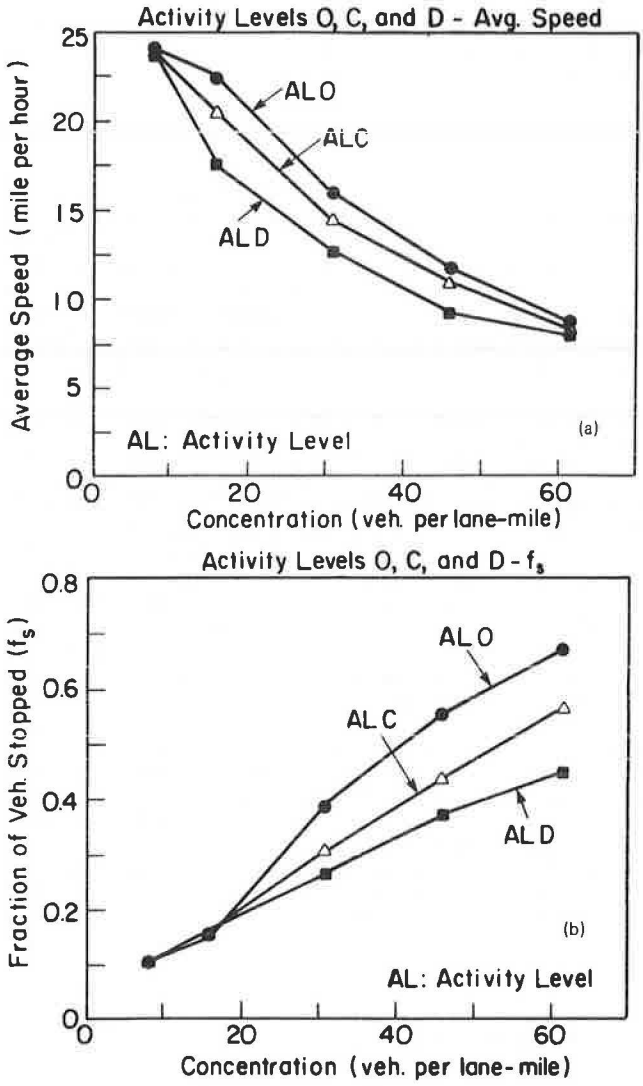


FIGURE 3 Effect of mean event duration for a mean interevent time of 36 sec (Activity Levels 0, C, and D): (a) average speed versus concentration and (b) fraction of vehicles stopped versus concentration.

three durations considered for the same reason that speeds are almost equal: vehicles in this nearly empty network encounter little impedance in getting around temporary obstructions. However, at each higher concentration level, and somewhat counter-intuitively, f_s is consistently smaller for longer event durations. This phenomenon can best be explained be realizing the identity, mentioned in the second section, of f_s and T_s/T. Essentially, as seen in Figure 2, increasing event duration results in a faster rate of increase of the average running time (T_r) with respect to concentration. Therefore, because $T = T_s + T_r$, the relative contribution of T_r to total trip time increases with increasing event duration, resulting in the observed decrease in f_s, particularly because T_s decreases with increasing mean event duration for a given concentration, as discussed hereafter.

Table 3 gives the average speed (V), trip time (T) and its two components T_r and T_s, along with the fractions T_r/T and T_s/T, corresponding to the three activity levels (0, C, and D) under three levels of concentration (approximately 8, 30, and 60 vehicles per lane-mile). The data in this table reveal that, at low concentrations ($K \approx 8$), there is essentially no difference in the descriptors across activity levels because of the ability of vehicles to get around the obstructions with minimal running or stopped delay. However, at higher concentrations, average (total) trip time (T) increases with activity level, as shown in Figure 4(a). This increase is due exclusively to the increase in running time (T_r), which even compensates for a decrease in average stop time (T_s). This somewhat counter-intuitive decrease in T_s is more evident at the highest concentration considered ($K \approx 60$). This probably arises because (stopped and moving) delays at the obstructions reduce the otherwise dominant stopped delays at the intersections.

Similar results were found for Activity Levels 0, B, and E, where the mean duration time varied from 0 to 30 to 45 sec for a mean interevent time of 120 sec. However, the effects were present to a lesser extent because the essentially similar durations took place within a much longer interevent time and therefore had a lesser impact on traffic. The effect of interevent time is further discussed hereafter. Finally, the two-fluid parameter n is plotted in Figure 4(b) as a function of mean event duration for the two levels of mean interevent time considered, indicating that n appears to increase linearly with duration time (for a given event frequency).

Effect of Mean Interevent Time

Activity levels at a 30-sec mean duration can be arranged in the order of least frequent (i.e., never occurring) events to most frequent events as follows: Level 0 (no events), Level A (240 sec between events), Level E (120 sec between events), and Level D (36 sec between events). Speed and fraction of stopped vehicles are plotted against concentration in Figures 5(a) and 5(b), respectively, for the four levels. The effect of decreasing interevent time (or increasing frequency) is essentially similar to that

TABLE 3 Effect of Activity Level on Average Speed and Trip Time Components Under Varying Concentrations

Concentration K (veh/ln-mile)	Activity Level	Average Speed V (mph)	Trip Time T (min/mile)	Stop Time T_s (min/mile)	Running Time T_r (min/mile)	T_r/T	T_s/T
7.92	0	24.03	2.50	0.26	2.24	0.897	0.103
	C	23.84	2.52	0.27	2.25	0.895	0.105
	D	23.64	2.54	0.26	2.28	0.896	0.104
30.89	0	15.84	3.79	1.45	2.33	0.616	0.384
	C	14.27	4.20	1.26	2.94	0.700	0.300
	D	12.52	4.79	1.25	3.54	0.738	0.262
61.38	0	8.53	7.03	4.65	2.38	0.339	0.661
	C	8.18	7.33	4.10	3.24	0.441	0.559
	D	7.73	7.76	3.44	4.32	0.557	0.443

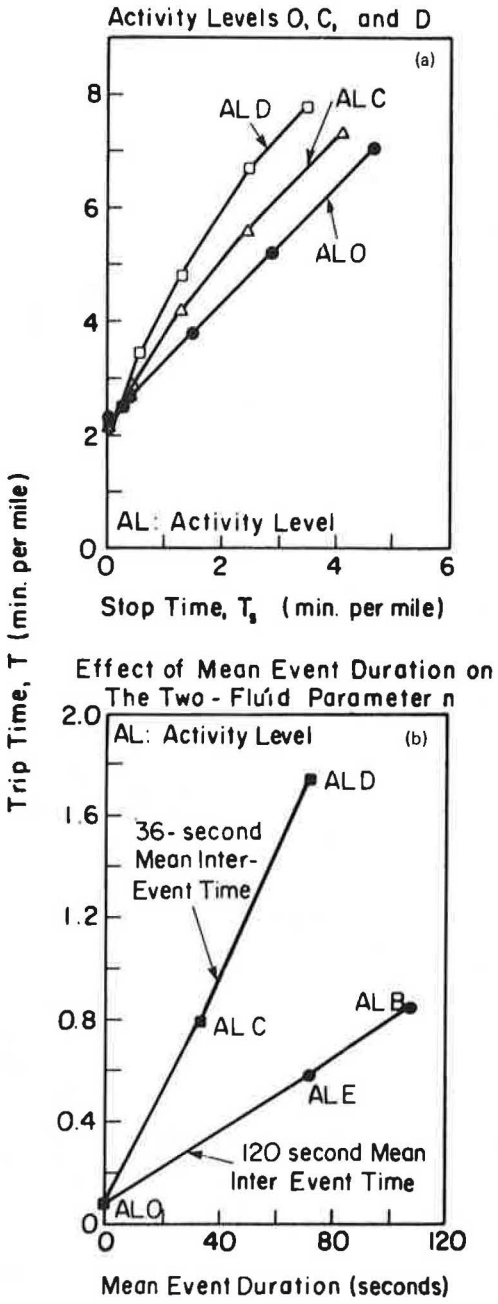


FIGURE 4 Effect of mean event duration for mean interevent times of 36 and 120 sec: (a) trip time versus stop time, 36-sec mean interevent time only (Activity Levels 0, C, and D) and (b) n versus mean event duration for two levels of mean interevent time (36 and 120 sec).

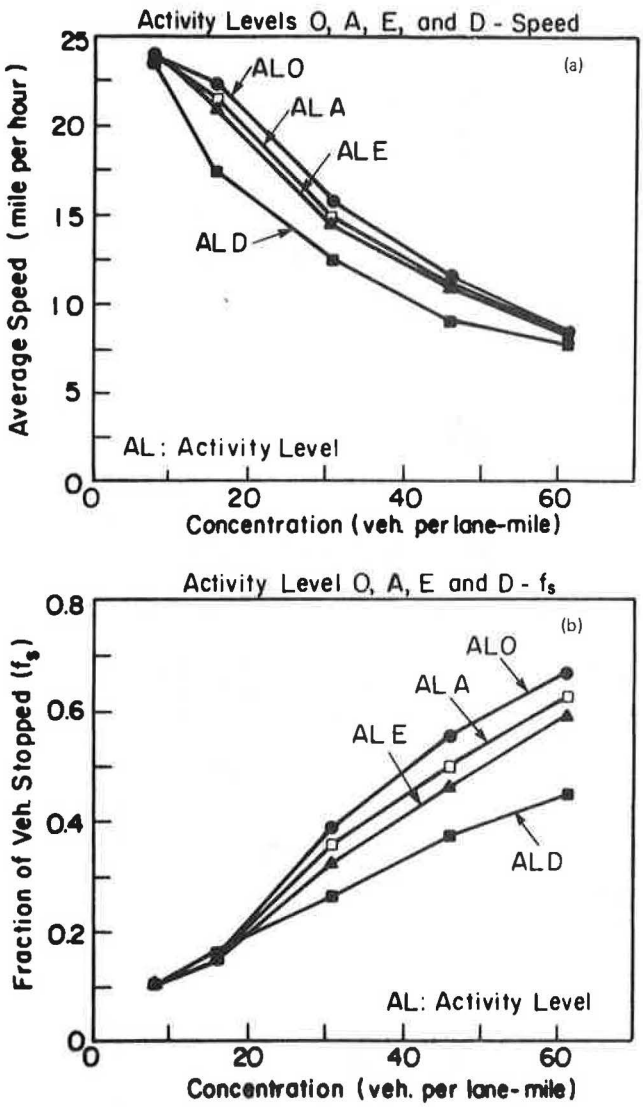


FIGURE 5 Effect of mean interevent time for a mean duration of 30 sec (Activity Levels 0, A, E, and D): (a) average speed versus concentration and (b) fraction of vehicles stopped versus concentration.

of increasing event duration, discussed earlier, and for the same reasons: as either mean duration increases or the interevent time decreases, the fraction of time a lane is blocked increases. The change between Levels E and D in Figures 5(a) and 5(b) is appreciably larger than the change between Levels E and A because events occur twice as frequently in Level E as they do in Level A but more than three times more frequently in Level D than in Level E.

Figure 2 shows that T_r increases with concentration at a faster rate for decreasing interevent times (Levels 0, A, E, and D, respectively), indicating that the parameter n increases correspondingly. The T-versus-T_s plot for these four levels is shown in Figure 6(a), revealing an increasing slope with decreasing interevent time. Figure 6(b) shows n as a function of the frequency of events, holding mean duration constant at 30 sec. As in Figure 4(b), this relationship appears to be linear.

Effect of Fraction of Time Lane Is Blocked

Next the effect of the fraction of time that the right lane is obstructed, without regard to the actual durations and interevent times, is considered. Activity Levels B and C were selected to have nearly equal values of this fraction (the slight difference is due to the necessity of specifying integers in the NETSIM model). The data in Table 2 indicate that n and T_m are extremely close (though not identical) for both levels, and only Level B will be used in the following comparisons. The remaining levels can be arranged in order of increasing fraction of time of lane blockage: Level 0 (no events), Level A (0.125), Level E (0.250), Level B (0.375), and Level D (0.833).

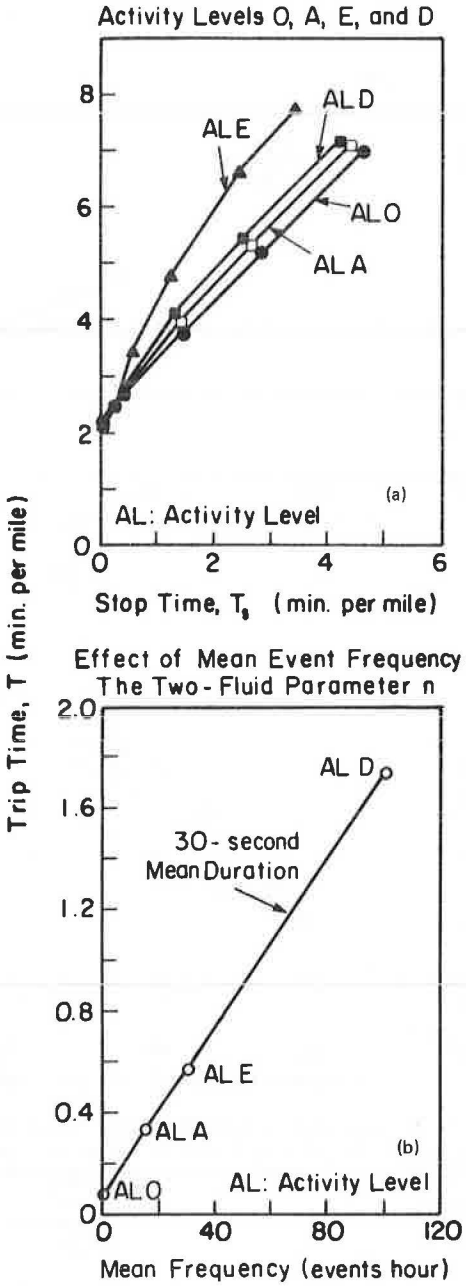


FIGURE 6 Effect of mean interevent time for a mean duration of 30 sec (Activity Levels 0, A, E, and D): (a) trip time versus stop time and (b) n versus event frequency.

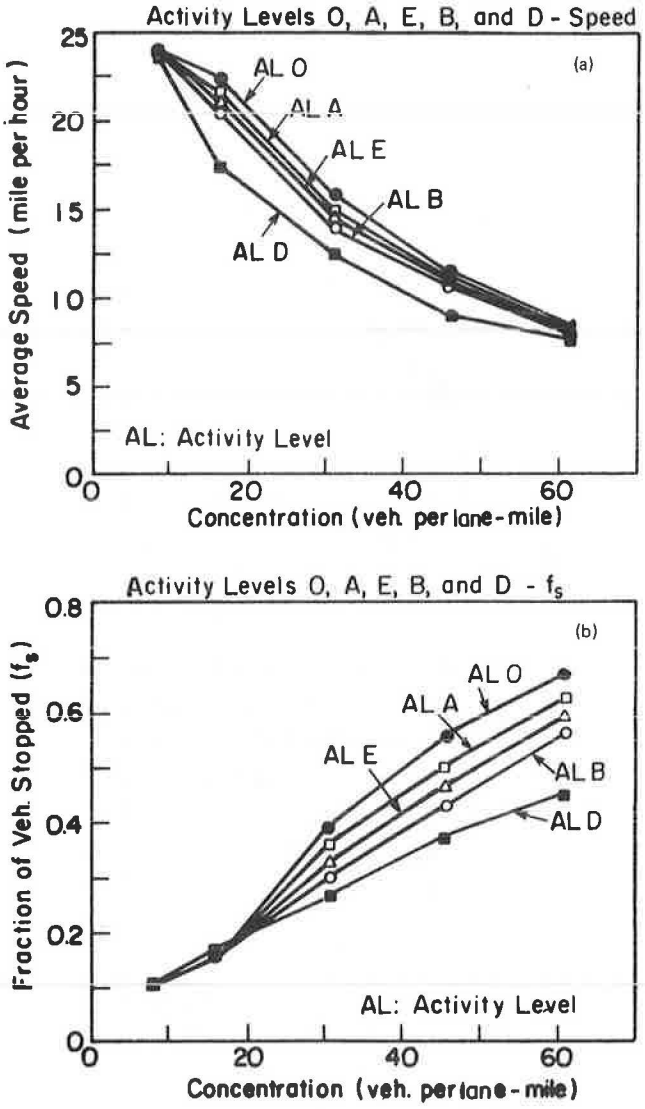


FIGURE 7 Effect of fraction of time lane is blocked (Activity Levels 0, A, E, B, and D): (a) average speed versus concentration and (b) fraction of vehicles stopped versus concentration.

Speed and fraction of stopped vehicles are shown plotted against concentration in Figures 7(a) and 7(b), respectively, and, as expected, bear a marked resemblance to the previous figures corresponding to duration and interevent time individually. Similarly, n increases with the fraction of time that the lane is blocked, as seen in Table 2 and reflected in the T-versus-T_s plot in Figure 8(a). In addition, n appears to be a linearly increasing function of this fraction, as shown in Figure 8(b).

Remarks

As was pointed out in the second section, an increasing T_r with respect to concentration is a result of congestion and perturbations in the moving traffic. When no short-term events are specified, n is nearly zero, intersection effects dominate intralink effects, and traffic-interfering activities normally observed in city streets at all but the lowest concentration levels are not represented in the simulations. As short-term events are introduced, T_r is allowed to increase with increasing concentration, resulting in non-zero values of n. When n is related to the specific components of the short-term events, as well as to a single measure of the event intensity, a linear relationship develops. Of course, other factors also affect the value of n, some of which will be discussed in the next section and others of which will be mentioned in the final section. In investigating these other factors, a moderate level of short-term events (Activity Level B) was used because the resulting higher values of n (as opposed to a case with a zero activity level) are more representative of traffic conditions in cities, thus allowing this parameter to better respond to changes in the network and traffic factors under consideration, as described in the next section.

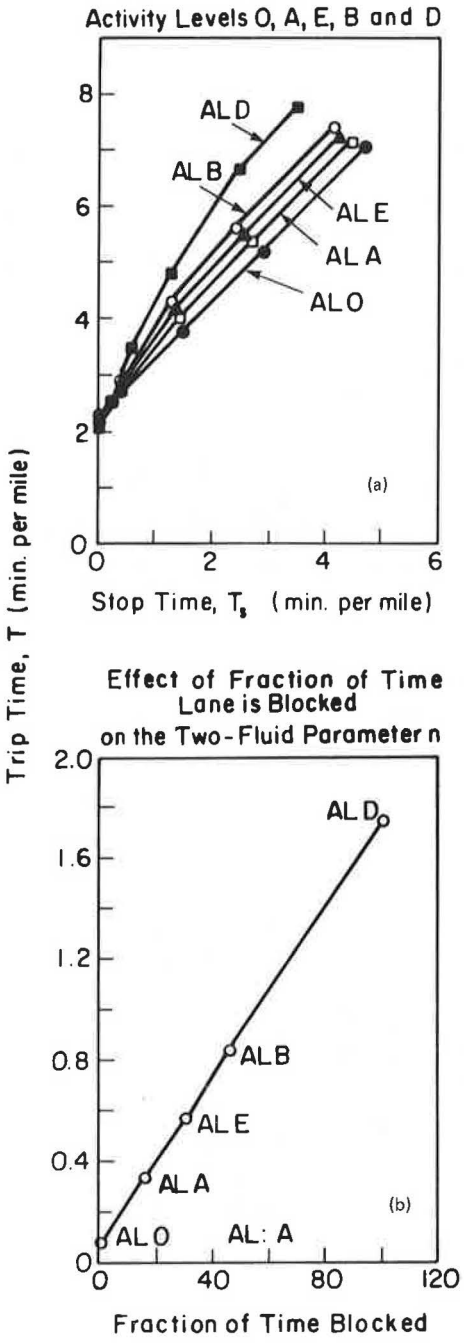


FIGURE 8 Effect of fraction of time lane is blocked (Activity Levels 0, A, E, B, and D): (a) trip time versus stop time and (b) n versus fraction of time lane is blocked.

NETWORK FEATURES AND TRAFFIC CONTROL

The sensitivity of the selected network parameters and traffic flow relationships to changes in traffic control strategy and network topology (block length) is addressed in this section. The base conditions described in the third section and used in the fourth section are also employed here, except to examine the effect of block length as indicated in the appropriate subsection. As mentioned in the previous section, Activity Level B was specified in order to provide a suitable degree of intralink friction in all runs.

Values of T_m and n used in this section were estimated in the manner described earlier and are given in Table 4, where the base case consists of Activity Level B applied to the base conditions described previously. Again, values obtained by regressing T and T_r from the four lower concentration levels only will be used, as discussed in the previous section. The results are presented hereafter for each of the two factors.

TABLE 4 Estimates of the Two-Fluid Parameters T_m and n

Case	n	T_m	r^2
Base case[a]	0.845	2.135	0.973
Traffic control strategy, no progression	0.490	2.196	0.921
Network topology, 400-ft links[b]	0.394	2.580	0.994

[a] 1,000-ft links, progression.

[b] This case should be compared with the no-progression case not the base case.

Traffic Control Strategy

A question of considerable practical importance is that of the effect of a major change in the traffic control scheme of a city network on the two-fluid parameters and on n in particular. To address this question, the base case, where signals were timed for progression (see the third section), was compared with a situation in which signal timing did not explicitly provide for such progression. Under the latter scheme, the cycle length of all the traffic signals in the test network was 50 sec (versus 40 sec in the base case). The percentage of signal splits and offsets remained unchanged.

The average speed and the fraction of vehicles stopped are plotted against concentration in Figures 9(a) and 9(b), respectively. Over all concentration levels, average speed dropped and the fraction of vehicles stopped rose. Because signal progression is implemented to reduce stopping, these results are not surprising. The drop in average speed results from vehicles being stopped a greater fraction of time while holding the desired speed constant.

After beginning at nearly the same value, average running time (T_r) increases with concentration at a faster rate under progression (see Figure 10), meaning that while they are moving vehicles are traveling at a slower speed than when there is no progression. But the higher average overall speed under progression along with the reduced fraction of vehicles stopped (and therefore fraction of time stopped for any one vehicle, averaged over all vehicles) indicate a system that is operating more smoothly with progression.

The slower increase of stopped time (T_s) with an increasing T when the signals are timed for progression is shown in Figure 11(a). This difference in slopes is also reflected in the values of n: 0.490 for no progression, 0.845 for progression. The values of T_m are nearly identical (see Table 4) with that for progression being slightly lower indicating a slightly higher V_m, the average minimum running speed in the network. The change in cycle length may be responsible for part of the difference between progression and no progression; however, such cycle changes are often necessary when implementing progression schemes in urban areas. The main result from the perspective of the present paper is that n can be quite sensitive to major changes in a network traffic control scheme. Such sensitivity had not been established previously and is quite important for practical engineering use of the two-fluid model.

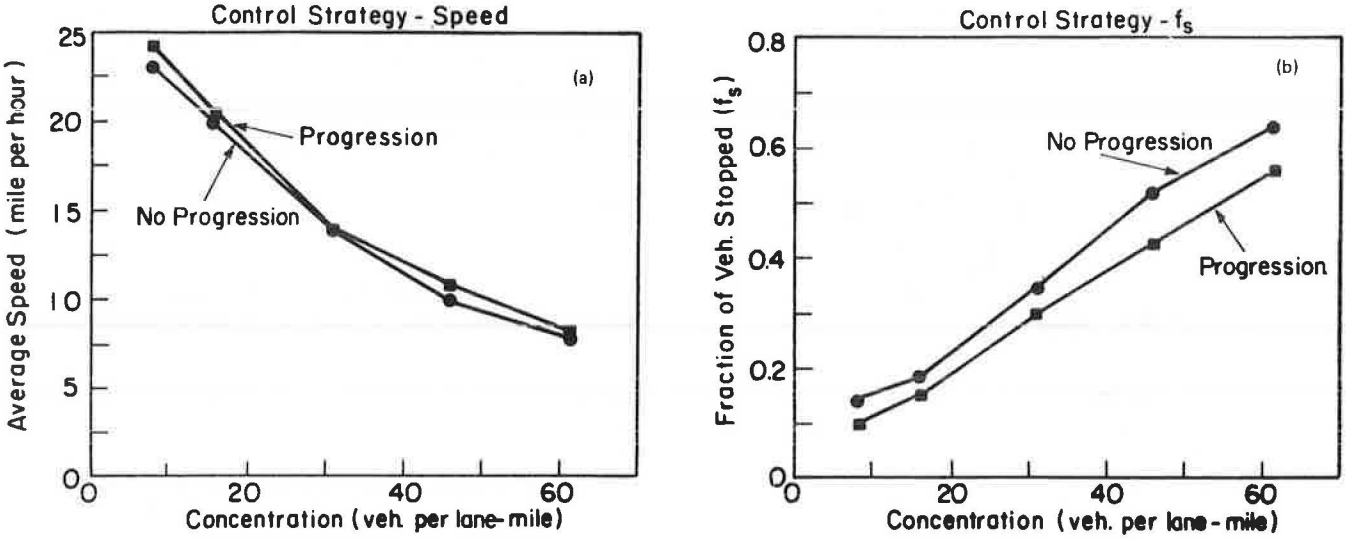


FIGURE 9 Traffic control strategy—progression versus no progression: (a) average speed versus concentration and (b) fraction of vehicles stopped versus concentration.

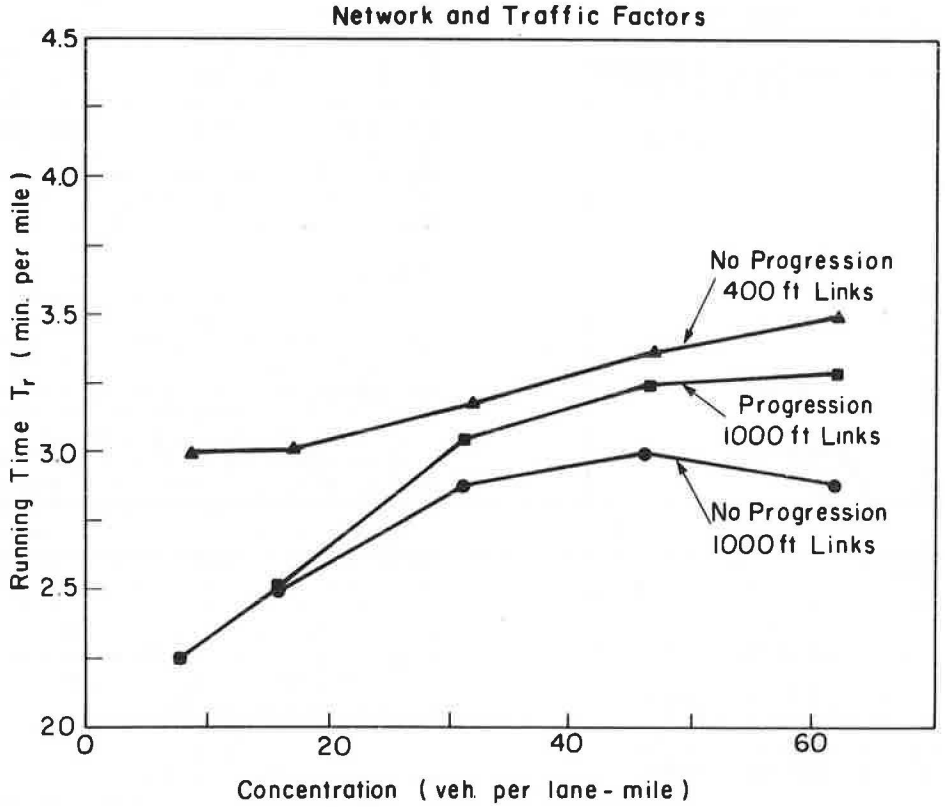


FIGURE 10 Running time versus concentration: network and traffic factors.

Network Topology--Block Length

The block length used to study network topology was reduced from 1,000 to 400 ft. A 50-sec cycle length was used so that these runs could be compared with the no-progression case, because a 40-sec cycle length no longer provided progression with the reduced block length. Average speed and f_s are plotted against concentration in Figures 12(a) and 12(b), respectively, and both indicate that the effect of block length is substantial. At low concentrations average speed for 400-ft links is reduced by almost half of that for the longer links. The more frequent intersections (associated with the shorter links) require vehicles to stop much more often (particularly because there is no progression in either case), and many are unable to ever reach their desired speed. However, at higher concentration, intense congestion does not allow vehicles to build up much speed in either case.

The fraction of vehicles stopped at the low concentrations is much higher for the shorter blocks,

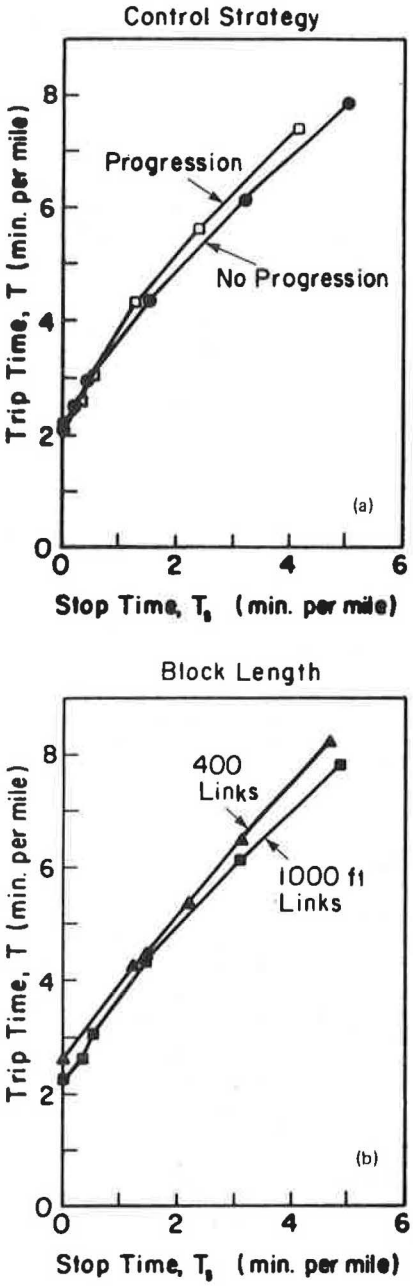


FIGURE 11 Trip time versus stop time for network and traffic factors: (a) progression versus no progression and (b) 1,000-ft versus 400-ft block lengths.

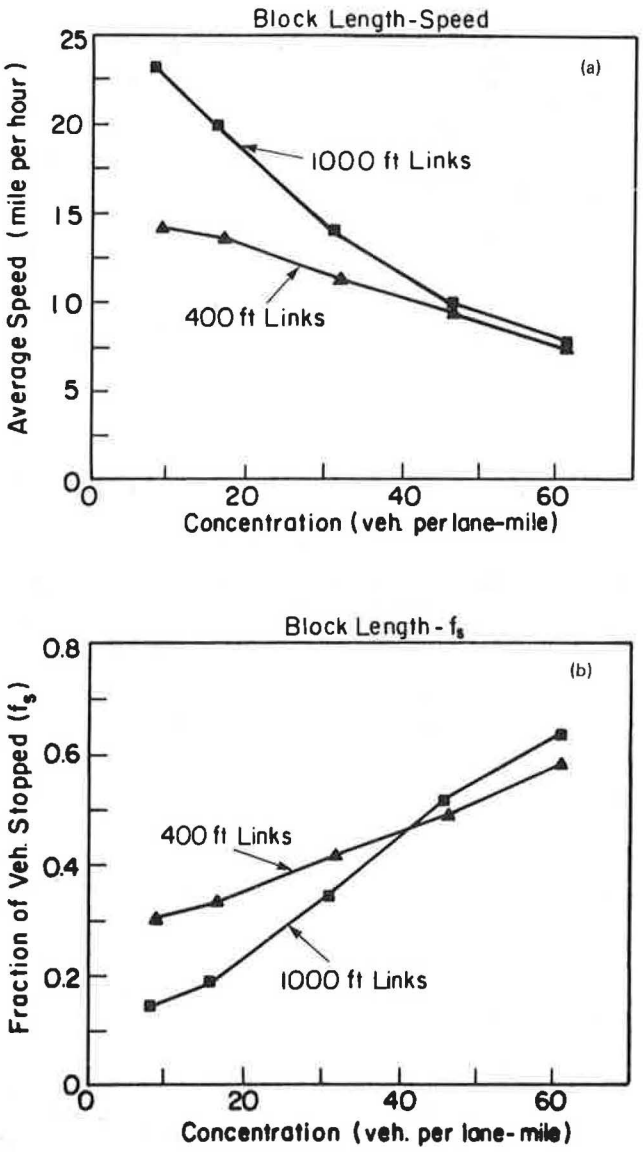


FIGURE 12 Network topology—1,000-ft versus 400-ft block lengths: (a) average speed versus concentration and (b) fraction of vehicles stopped versus concentration.

as expected. However, as concentration increases, Figure 12(b) reveals that the f_s-versus-K lines corresponding to the two lengths cross, and f_s for short block lengths becomes less than that for the longer block length.

The running time (T_r) for the short blocks is consistently higher than that for the longer blocks (see Figure 10), but the slope for the 1,000-ft blocks is somewhat steeper. The effect of this difference in slope on the value of n is somewhat offset by the fact that the slope of T (travel time, the inverse of the average speed) as a function of concentration K is steeper for the long links. The resulting values of n are 0.394 and 0.490 for the 400- and 1,000-ft links, respectively (Table 4). The difference in the values of T_m is substantial as well (see Table 4), with T_m for short links (2.580 min per mile) being larger than that for the long links (2.196 min per mile). The difference between the values of n and T_m are reflected in the T-versus-T_s plot in Figure 11(b).

CONCLUDING REMARKS

The analysis of network-level traffic flow relationships presented in this paper has addressed two aspects of this problem. First, although the feasibility of using the NETSIM microscopic traffic simulation model was established in previous work (1), serious limitations were present due to the highly idealized nature of traffic operations in the test network. In particular, the values of the two-fluid model parameters indicated that insufficient moving traffic interference was encountered by the vehicles circulating in the test network under the specified conditions. Therefore, as trip time increased with concentration, running time remained virtually un-

changed (1). By use of the short-term events feature of NETSIM, this interference, or intralink friction, could be introduced into the traffic stream at specified levels. The model proved quite sensitive to these traffic-interfering urban activity levels, as was shown by the two-fluid parameter values. Essentially, it is such perturbations that give a city street network its inherent traffic character and the kind of behavior described by the two-fluid theory and verified by observation.

Second, although the two-fluid theory has been extremely successful at capturing relationships between macroscopic traffic flow descriptors, it does not explicitly address the key underlying mechanisms or how the model parameters may be affected by the physical and operational features of a given network. Analyses such as the present one can contribute greatly to those aspects that are outside the scope of the theory. For instance, this work has helped confirm the previously mentioned role of traffic-interfering activities in determining the behavior of a system. In addition, the effect of two major factors, namely block length and the traffic control scheme, were ascertained. Changes in both of these factors resulted in significant changes in the two-fluid model parameters as well as in other network traffic flow descriptors and relationships. It was thus seen that a major change in traffic control strategy, in this case retiming of all signals to achieve flow progression, could be detected by the two-fluid model parameters.

The two-fluid parameter n can be viewed as an indicator of overall network intralink friction, the effect of which becomes more apparent with increasing concentration. Specifically, n is a measure of how quickly average running time increases with average (total) trip time: the higher n is, the greater is the rate of increase of running time with respect to trip time.

Although one of the modeling objectives of this study was to achieve relatively high values of n that were comparable to values that have been observed in related field work in order to provide a realistic system representation, high values of n are not desirable from a traffic engineering design perspective. Indeed, all else being equal, a small n value would indicate less intralink friction (i.e., smoother flow) and a shorter running time for a given stop time. This clearly indicates that this parameter must not be used in isolation as an indicator of traffic service quality. A case in point is the introduction of progression in the timing of signals, whereby the value of n actually increased, though other indicators, such as average speed (or trip time) and fraction of vehicles stopped, clearly supported the desirability of the new timing scheme. Thus, although T_r for a given concentration level was lower under progressive timing, the corresponding increase in average speed (or, conversely, decrease in T) offset the effect of the variation of T_r, resulting in the observed higher n value.

Field studies indicate that running time continues to increase with trip time and rising concentration. However, field studies can only observe existing conditions, and the highest observed concentrations have been around 40 vehicles per lane-mile. The simulation experiments discussed here have gone as high as 60 vehicles per lane-mile and appear to indicate that the rate of increase of T_r may decrease at exceptionally high concentrations, with T_r perhaps approaching an upper limit. After all, there is a limit below which drivers will just stop instead of creeping along.

Finally, it has already been noted that the activity level was the one factor that allowed the greatest control over the selected network parameters. However, it is likely that there may be an upper limit to the extent to which the activity level (at least as defined here) can affect these parameters. The other network and traffic factors, discussed in the previous section, were also shown to have an effect on these parameters. The test network used herein, however, is still rather idealized, with uniform streets and intersections. Field studies conducted in reference to the two-fluid model have shown that relatively regular street systems (e.g., the Houston CBD) result in lower values of n than do the more irregular street networks of other cities (5). Thus, one of the next steps in these simulation experiments is to introduce network topological irregularities, such as boulevards, missing links, one-way streets, and so on. Naturally, numerous other avenues for additional research exist in the development of a macroscopic network-level theory of traffic flow.

ACKNOWLEDGMENTS

Partial funding for this research has been provided by a grant from the Bureau of Engineering Research at the University of Texas at Austin to H.S. Mahmassani. Computer funds were provided in part by the Department of Civil Engineering and the Center for Studies in Statistical Mechanics at the university. The assistance of Alfredo Hinojos in the preparation of the graphic material is appreciated. The authors of course remain solely responsible for the contents of this paper.

REFERENCES

1. H.S. Mahmassani, J.C. Williams, and R. Herman. Investigation of Network-Level Traffic Flow Relationships: Some Simulation Results. In Transportation Research Record 971, TRB, National Research Council, Washington, D.C., 1984, pp. 121-130.
2. R. Herman and I. Prigogine. A Two-Fluid Approach to Town Traffic. Science, Vol. 204, 1979, pp. 148-151.
3. M.F. Chang and R. Herman. Trip Time Versus Stop Time and Fuel Consumption Characteristics in Cities. Transportation Science, Vol. 15, 1981, pp. 183-209.
4. S. Ardekani and R. Herman. Quality of Traffic Service. Research Report 304-1. Center for Transportation Research, University of Texas at Austin, 1982.
5. R. Herman and S. Ardekani. Characterizing Traffic Conditions in Urban Areas. Transportation Science, Vol. 18, 1984, pp. 101-139.
6. S. Ardekani, V. Torres-Verdin, and R. Herman. The Two-Fluid Model and the Quality of Traffic in Mexico City. University of Texas at Austin, 1984.
7. I. Prigogine and R. Herman. Kinetic Theory of Vehicular Traffic. American Elsevier, New York, 1971.
8. Peat, Marwick, Mitchell and Company. Network Flow Simulation for Urban Traffic Control System--Phase II, Vol. 1-5. FHWA, U.S. Department of Transportation, 1973.
9. Traffic Network Analysis with NETSIM--A User Guide. Implementation Package FHWA-IP-80-3. FHWA, U.S. Department of Transportation, 1980.
10. F.V. Webster and B.M. Cobbe. Traffic Signals. Road Research Technical Paper 56. Her Majesty's Stationery Office, London, England, 1976.

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

Macroparticle Traffic Simulation Model To Investigate Peak-Period Commuter Decision Dynamics

GANG-LEN CHANG, HANI S. MAHMASSANI, and ROBERT HERMAN

ABSTRACT

A special-purpose, macroscopic highway corridor traffic simulation model is presented. The model views traffic as discrete vehicle bunches or macroparticles that are moved according to local speeds defined by local concentrations, resulting in high computational efficiency. The model allows the investigation of commuter decision dynamics and their interrelation with time-dependent congestion patterns. Two applications of the model are described illustrating its use (a) in conjunction with behavioral rules by which commuters respond to experienced congestion and (b) as part of an interactive experiment involving decision making by real commuters.

Peak-period traffic congestion is an everyday occurrence in most metropolitan areas, particularly in commuting corridors. The excessive delay, instability of travel time, and increased fuel consumption that accompany congestion translate into significant economic and social cost. With the well-known difficulties that preclude the expansion of roadway capacity in urban areas, most efforts to relieve congestion during the past decade have centered on improving system usage through traffic control as well as demand-side strategies. An understanding of the day-to-day interaction between traffic patterns and commuters' departure time and route choice behavior is essential to the analysis of peak-period congestion and the design and proper evaluation of control measures.

Some recent analytical studies have addressed commuter departure time decisions and the resulting temporal traffic pattern in a dynamic user equilibrium framework (1-5). Most of these studies have considered an idealized system with a single origin and a single destination, with congestion occurring at a unique bottleneck along the only available route. However, a more realistic though more complex situation is that of multiple origins and multiple destinations in a particular commuting corridor, where congestion may develop at more than one location and vary from day to day.

The analysis of the interrelationship between the time-dependent congestion patterns and commuters' responses in such a system is actually too difficult to be tackled analytically under realistic assumptions. A more convenient and effective approach is to use a special-purpose simulation tool that can capture the complex dynamics of the system, particularly the fluctuation of travel time with departure time and the time-dependent congestion patterns. This tool should also possess the capability to explore the impact of control strategies of alternative user decision rules on the performance of the system and its convergence properties. It was in response to these needs that the Macroparticle Simulation Model (MPSM) was developed.

The purpose of this paper is to introduce the logic employed in this special-purpose simulation model and to describe its application to ongoing research on the dynamics of user decisions and peak-period congestion. Further, because appropriate data for the study of commuters' departure time choice dynamics are quite difficult to obtain in a real-world context, interactive use of this program, in conjunction with survey techniques to study these phenomena, is discussed.

The general structure of this model and some of its key features are outlined in the next section, including a discussion of the inputs and major options available to the model user. In the second section computational experience and a brief sensitivity analysis with respect to key control variables are presented. The application of this model is the principal concern in the third section, in which is presented a summary of a number of computer simulation experiments that were conducted to investigate a commuting corridor system's convergence status under prespecified user decision rules. Further illustration of the model's application is given in the fourth section, in which a description is given of the use of the MPSM in conjunction with an interactive experimental procedure in which participants make daily departure time decisions in a hypothetical commuting situation. The variation of congestion patterns as well as the evolution of system concentration predicted by the MPSM, given the participants' choices, are examined therein. Finally, concluding comments and possible further extensions of this work are presented in the last section.

DESCRIPTION OF THE MPSM

Overview

Because the major concern is with the dynamics and convergence properties of the corridor system, the simulation program must possess the capability of providing the following information: (a) the concentration fluctuation in each section along the highway corridor for a given time-dependent demand, (b) the variation of travel time versus departure time for a given origin-destination pattern, and (c) the day-to-day evolution of the system's performance under various postulated rules that might govern users' departure time choice behavior.

Most of the existing macroscopic corridor simulation packages, such as MACK and SCOT, can meet these requirements (6,7). However, each simulation package was developed for applications that are not necessarily consistent with the present purpose, result-

ing in excess requirements and processing information and additional computational cost. These are important considerations in this research because the search for system covergence typically requires a considerable number of repeated simulations, often in excess of 50 simulation days. These considerations have led to the development of the special-purpose MPSM.

The MPSM is a fixed-time macroscopic traffic simulation model, which uses established traffic flow relationships to simulate the movement and the interaction of vehicles in a commuting highway corridor. Unlike most macroscopic simulation programs, for instance the MACK or FREFLO family (6,7), the traffic flow in the highway is not modeled as a compressible fluid but is viewed as a collection of vehicle groups or bunches, termed macroparticles [somewhat similar to the platoons used in the DAFT model (8,9)]. The model keeps track of the physical position of those particles using a prespecified speed-density relationship. The macroparticle approach avoids the significant computational cost of tracing individual vehicles, which is not essential to the present research. In addition, it avoids approximating traffic as a continuous fluid and the resulting inaccuracies such as the occurrence of nonphysical speeds. Further details regarding the tracing of vehicle movement are presented later in this section.

The MPSM consists of two principal components. The first component, or vehicle generation component, processes the daily decisions of users into sector-specific discretized time-dependent vehicle demand patterns. A user-decisions subroutine is also included to allow the researcher to specify departure time readjustment rules governing the behavior of commuters on a daily basis. The second component, or simulator, actually simulates the flow of traffic, including queueing at entrance ramps and vehicle movement along the highway. The overall framework and the interrelationship between each principal component are shown in Figure 1.

The input to the MPSM consists of two principal categories: supply-side data and demand-side information. The first category contains information on the key physical and operational features of the highway facility, such as the total corridor length, the number of analysis sections that the corridor is subdivided into, and the number of lanes as well as the allowable free flow speed in each section. The parameters of the speed-density model, as well as various simulation control parameters (described later in this section), are also included in this category.

The second category consists of information about usage level and commuter behavior. These are discussed hereafter in conjunction with the vehicle generation component of the model.

Vehicle Generation Component

The key demand-side input to the traffic simulation is the time-dependent vehicle departure patterns from each of the residential or origin sectors considered in the analysis. There are two principal approaches available to the model user; the approach used depends on the particular research application. The first approach consists of specifying the generation functions directly for a particular simulation day.

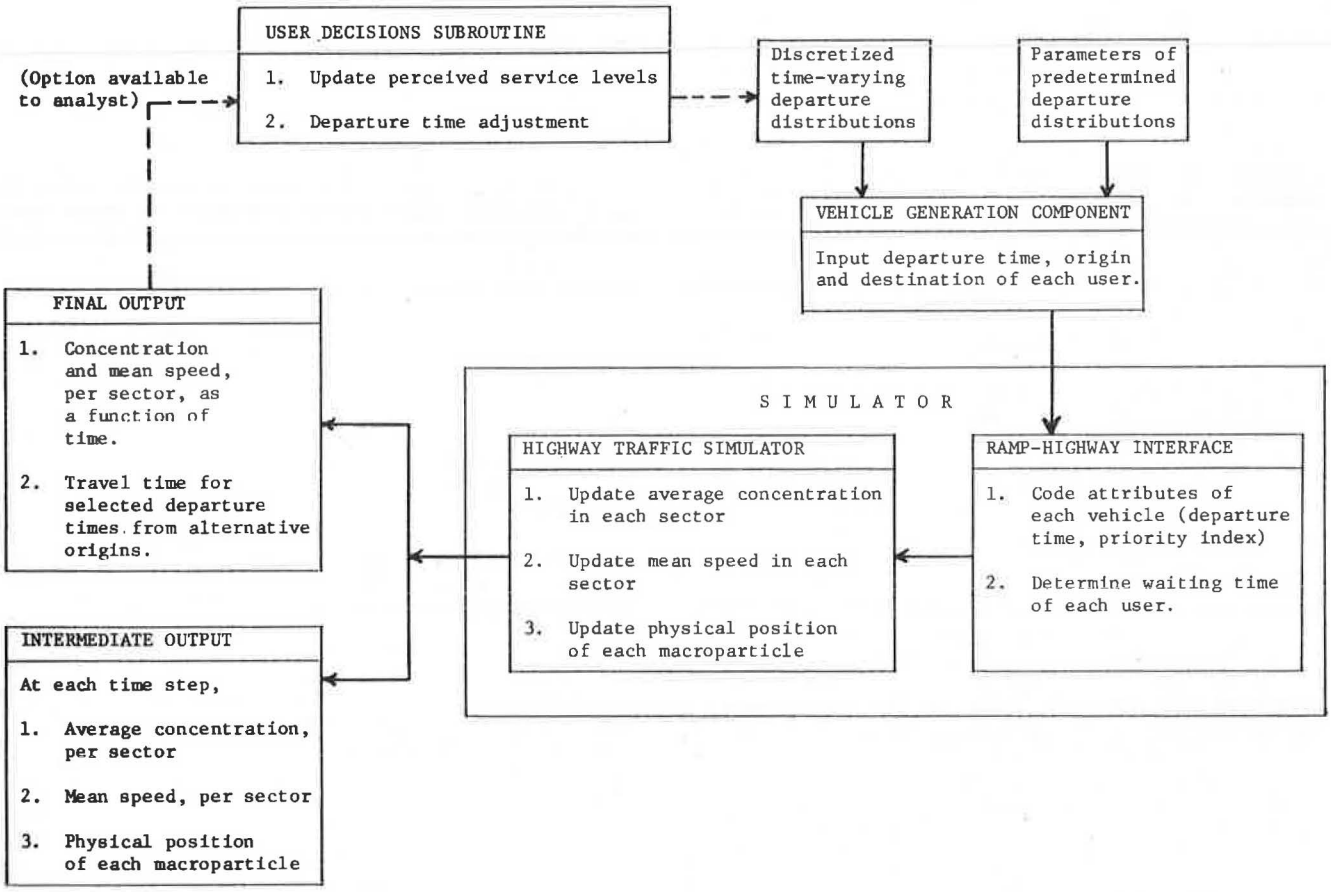


FIGURE 1 Overall framework of the MPSM.

The second approach treats the formation of the time-varying departure pattern endogenously, using individual decision rules to determine users' departure time choices on a given simulation day in response to the service levels experienced on previous days, as determined by the traffic simulation. This second approach would be appropriate when the objective is to investigate the effect of alternative models of individual behavior, or when such models have been calibrated on actual observations and are used to evaluate the effect of alternative control strategies on peak-period congestion.

When following the first approach, the model user can specify the departure time distributions in one of two forms: (a) as a discretized function giving the number of departures, from a particular origin, in each time interval or (b) as one of a number of commonly used distributions for which the model user has to supply the appropriate parameter values. An example application using this first approach will be given in the fourth section, where departure time decisions are made by real commuters participating in an interactive experiment. The vehicle generation component then processes the individual decisions into the desired time-dependent input functions.

Under the second approach, a user-decisions subroutine needs to be specified. It essentially takes the output of the traffic simulation on a given day, particularly travel time and schedule delay (defined as the difference between a commuter's desired arrival time and actual arrival time at the work destination), and determines each user's response in the form of a departure time selection for the following day. The rules used for this readjustment process constitute one important area of the overall research effort. Possible behavioral rules include those based on the widely used utility maximization principle (1,5), as well as other less restrictive heuristic rules reflecting so-called boundedly rational or "satisficing" behavior (10). A brief example application of this approach is given in the third section; further details of this particular aspect are outside the scope of this paper and can be found elsewhere (10).

Essentially, under the second approach, the simulations automatically proceed from one day to the next, as the individual decisions determined in the user-decision subroutine are internally transformed into time-dependent departure patterns. The process takes place over a prespecified maximum number of days or until steady state is reached, whichever occurs first. It should also be noted that, under either approach, additional demand-side input information includes users' respective desired arrival times (used in the schedule delay calculations), which need not necessarily be identical to their work start times, as well as the desired destination.

SIMULATOR

Ramp-Highway Interface

Every departing vehicle is assumed to start directly at the corresponding entry ramp onto the highway facility. Therefore access time from a specific residence to the ramp is not explicitly considered in these analyses, which is equivalent to implicitly assuming that it is a constant term and thus would merely shift a user's departure time by a fixed amount. A departing vehicle is identified by its departure time and the required travel distance (based on the desired destination). Naturally, vehicles cannot all enter the facility simultaneously due to physical and operational capacity constraints (including possible traffic control devices). Because the microscopic details of merging maneuvers are beyond the level of detail of the model, a simple deterministic queueing approximation is employed to handle this phenomenon.

Denoting the service rate by s, the queue length at time t by D(t), and the fixed simulation time step by Δt, a user wishing to depart in the interval $[t, t + \Delta t]$ is considered to incur a wait time only if $D(t) > s\Delta t$.

Note that the service discipline at the ramp queue is first come, first served. However, in the event that some users wish to depart simultaneously, the departures need to be ordered. A priority index is defined for this purpose and can be either assigned exogenously by the model user or randomly generated by the model using the Monte Carlo technique.

Vehicles leaving the queue are subsequently grouped in macroparticles for moving on the highway proper. When using a fixed macroparticle size, a minor problem arises when the number of entering vehicles does not form an integer number of complete macroparticles, thus delaying some vehicles until a sufficient number are present to complete that last group. Because a typical macroparticle size is between 5 and 10 vehicles, the resulting delay is often negligible and not in excess of Δt (Δt = 1 min was used in most of the simulations), except possibly at extremely low usage levels (when congestion is not of concern anyway). One way to get around this issue altogether is to use variable size macroparticles; in this case it is, however, recommended not to go below 5 vehicles per particle in order to maintain the computational cost advantages. Issues relating to macroparticle size are discussed hereafter in conjunction with traffic movement on the highway proper.

Highway Traffic Simulator

The highway traffic movement simulator is the core of the MPSM. This part of the program executes a set of procedures at every simulation interval, the length of which is user controlled and possibly different from that of the vehicle generation component. These procedures are described here and contrasted with other available models.

Most of the commonly used macroscopic simulation models (6,7,11,12), though developed for their own particular purposes, share the following set of assumptions: (a) time is discretized into small, equal intervals; (b) the highway facility is divided into sections; (c) traffic demand and system performance are effectively constant over a given time interval; and (d) traffic flow is viewed as a compressible fluid where the details of individual vehicle movement are inconsequential. Three basic equations are then used to govern the flow of traffic in the facility (12): a conservation equation, a speed-concentration relation, and the identity of flow to the product of speed and concentration. The conservation of vehicles can be stated as

$$k_i^{t+1} = k_i^t + [\Delta t/(\ell_i\ \Delta X_i)]\left[\ell_{i-1}\ q_i^{t+1} - \ell_i\ q_{i+1}^{t+1} + (N_i^{t+1}/\Delta t)\right] \quad (1)$$

where

k_i^t = concentration in section i during the t-th time step, in vehicles per lane-mile;

q_i^t = flow into section i during the t-th time step;
N_i^t = number of vehicles generated in section i minus those exiting in section i during the t-th time step;
ℓ_i = number of lanes of section i;
ΔX_i = length of section i; and
Δt = simulation time step.

The second equation is the speed-concentration relationship, which varies between the various models, and the third equation simply states that

$$q_i^{t+1} = k_i^t \, v_i^t \tag{2}$$

In the MPSM, both conservation and speed-concentration equations are used. However, as noted earlier, the flow relation (Equation 2) is not used. Instead, vehicles in the flow are viewed as groups of physical entities, termed macroparticles, and move in accordance with the local speed field, specified by a speed-concentration relation. Thus, the concentration of each section can be updated at every time step by tracing the actual physical positions of the particles. The general approach of tracing the position of groups of vehicles seems to have been adopted by the DAFT/SCOT family of simulation packages (8,9), in which vehicles are moved in platoons whose lengths are followed. In the MPSM, however, the length of the macroparticle need not be explicitly considered, as long as its size is not excessively large. The logic of the macroparticle approach adapted in this work follows that of the "magnetohydrodynamic particle code" developed for the fluid simulation of plasmas (13).

The conservation equation used in the MPSM then has the following form:

$$k_i^{t+1} = k_i^t + (1/\ell_i \, \Delta X_i) \left(M_i^{e,t+1} - M_i^{o,t+1} + N_i^{t+1} \right) \tag{3}$$

where $M_i^{e,t+1}$ and $M_i^{o,t+1}$ denote the vehicles that enter section i from the preceding section and those that move onto the next section, respectively, in a given time step Δt.

In the MPSM, the concentration in each section is updated, using Equation 3, at the beginning of every time step, and it is assumed to remain constant over the interval [t, t + Δt]. The corresponding mean speed prevailing during this interval can then be obtained from the speed-concentration relation. The functional form adopted in most of the simulations is

$$v_i^t = (v_f - v_0) \, (1 - k_i^t/k_0)^{\alpha} + v_0 \tag{4}$$

where

v_i^t = mean speed in section i during the t-th time step;
v_f, v_0 = mean free speed and minimum speeds on the facility, respectively;
k_0 = maximum or jam concentration; and
α = a parameter.

Note that the speed-concentration relation could be modeled using different or more elaborate formulations. However, this particular aspect has not been the focus of the present research, and Equation 4 has been found perfectly adequate to capture the character of the system's performance in the context of investigating the dynamic interaction between commuters' decisions and peak-period congestion.

Macroparticles are moved at the prevailing section mean speed, yielding the respective distances traveled during a particular time step and the resulting positions at the end of the interval. Section concentrations are subsequently updated, as described earlier, for the next time interval. In addition to its computational efficiency, tracing the macroparticles obviates the need for monitoring the traffic flow with a macroscopic flow equation (Equation 2). As mentioned earlier, the use of such a flow equation to control the flow from one finite section to another can transport material in unrealistically short times over long distances, thereby resulting in nonphysical high transport speeds. For example, if there are n finite sections in a linear highway, material will be transported every time interval Δt, no matter how small, so that some material will traverse n sections of finite length n $\cdot$ ΔX in the time nΔt, yielding the possibly unrealistic speed $\Delta X/\Delta t$, regardless of prevailing conditions. Further details on the particle-moving process are given hereafter.

Vehicle-Moving Process

The physical position of each macroparticle is updated at the end of every simulation interval and stored as one of the attributes of that macroparticle. Let X(i,m,t) denote the position of macroparticle m in section i (measured, in the direction of flow, relative to the beginning of the section) at the end of interval t, and R(i,m,t) the distance from its current position to the beginning of the next (downstream) section, as shown in Figure 2. Of course, R(i,m,t) = ΔX_i - X(i,m,t). The new position of particle m at the end of interval t+1 will be obtained by advancing it by a distance d(m,t+1) = $\Delta t \cdot v_i^{t+1}$ as long as d(m,t+1) $\leq$ R(i,m,t), meaning that its new position remains in section i.

In the event that d(m,t+1) > R(i,m,t), the particle will have to travel in the next section, i+1, during a fraction of the interval Δt. However, the mean speed in section i+1 may be different from that in the preceding section. The travel distance of macroparticle m during this Δt can thus be broken into two parts. The first part, which is equal to R(i,m,t), proceeds at the mean speed, v_i^{t+1}, for a fraction of the Δt equal to $R(i,m,t)/v_i^{t+1}$. In the remaining part of the time interval, namely $\Delta t' = \Delta t - [R(i,m,t)/v_i^{t+1}]$, travel takes place in section i + 1. Denote this second part of the distance traveled by R'(i+1, m,t+1). However, it does not seem reasonable to maintain a speed of v_i^{t+1}, nor to drastically change, midway, to v_{i+1}^{t+1}. A plausible way to handle this aspect is to assume that drivers will readjust their travel speed while entering i + 1 in a manner that is consistent with prevailing traffic conditions; the following averaging mechanism has been adopted in the model:

$$R'(i+1,m,t+1) = \Delta t' \left[(1/2) \left(v_i^{t+1} + v_{i+1}^{t+1} \right) \right] \tag{5}$$

Of course, the significance of this problem depends greatly on the size of the time step Δt, as well as on the section lengths. For sufficiently small Δt, the correction should not be of concern,

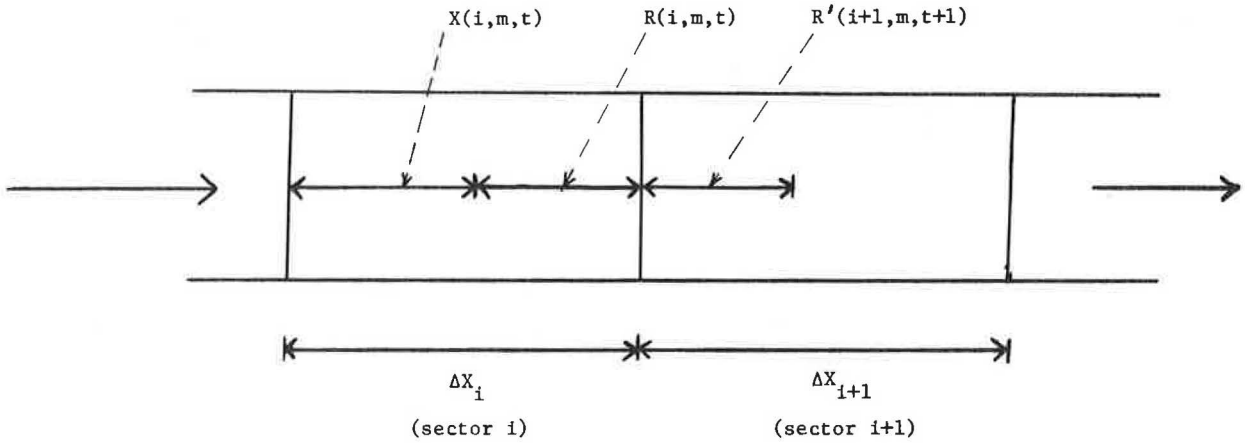


FIGURE 2 Graphic illustration of vehicle-moving process.

particularly for the present research applications. Selection of the number and lengths of the sections, of the time step Δt, and of the macroparticle size should be made judiciously. The frequently encountered trade-off between accuracy and computational efficiency is naturally present here. Sensitivity analyses have been conducted to provide useful insights into this aspect, as described in the next section.

MODEL USE CONSIDERATIONS AND SENSITIVITY ANALYSIS

In this section, some general considerations for selecting the key simulation parameters of the MPSM are discussed. In addition, the results of a number of numerical experiments conducted on the CDC system (Dual cyber 170/750) of the University of Texas at Austin are presented in order to better understand the effects of each key parameter. The numerical tests reported hereafter were carried out using a hypothetical 7-mile commuting corridor, with 2,400 trip makers uniformly distributed along its entire length, except for the last 1-mile section that immediately precedes the presumed common destination [central business district (CBD)]. In all cases, unless otherwise indicated, the highway facility was divided into seven 1-mile sections, with two lanes in each direction, though only the inbound (home-to-work) direction was considered here. The parameter values used in Equation 4 were $v = 45$ mph, $v_0 = 6$ mph, $k_j = 180$ vehicles per lane-mile, and $\alpha = 3.1416$ (or π).

There are three key simulation parameters in the MPSM: the simulation time step Δt, the macroparticle size Δm, and the section length ΔX. These are discussed in turn hereafter.

Selection of Section Length ΔX

As noted earlier, the continuous highway facility is usually divided into several discrete sectors in most macroscopic simulation models. Because system performance characteristics, such as concentration and speed, along each section are assumed to remain constant within each simulation time interval, a shorter section length can better reflect the location-dependent performance variation. Although the computational cost of course increases with the number of sectors specified (for a given total highway length), the relatively high efficiency of the MPSM is such that the incremental cost of increasing the number of sectors, or conversely of reducing ΔX, is rather small. This can be seen in Table 1 that gives the relationship between computational cost and number of sectors based on a set of numerical experiments under these conditions. In general, it is suggested that sector length not exceed 1 mile. Naturally, geometric features should be homogeneous within each sector. In addition, the sector length should provide adequate space for the vehicles discharged from the on-ramp during any simulation interval.

TABLE 1 Effect of Section Length (ΔX) on Computational Effort

No. Sectors[a]	ΔX(mile)	CPU[b]	CPU[c]
21	0.3	3.56	9.304
13	0.5	3.25	8.468
7	1.0	2.98	8.042
5	1.5	2.54	6.89
4	2.0	2.01	5.77

Note: In all experiments, $\Delta t = 1$ min, $\Delta m = 10$ vehicles, $V_f = 45$ mph, $V_0 = 6$ mph, $\alpha = \pi$, and simulation period = 60 min.

[a] The last 1 mile of this 7-mile facility was treated as one sector in all the cases reported here.

[b] No intermediate report is included in the program output.

[c] Intermediate report of system performance and all vehicle positions during each simulation interval is included in the output.

Selection of Simulation Time Step Δt

It can generally be expected that a smaller simulation time step Δt will result in greater simulation accuracy, particularly when there are large fluctuations in the vehicle generation pattern or significant variations in traffic conditions from sector to sector. Because traffic performance characteristics are assumed to remain constant during Δt, excessively long simulation intervals will, in general, lead to highly artificial discretized traffic flow. The resulting loss of accuracy is evident in Table 2, which gives a summary of the results of numerical experiments for values of Δt equal to 0.5, 1.0, 1.5, 2.0, and 3.0 min, respectively. For each case, the travel times associated with selected departure times from the same sector are given in this table, indicating that the magnitude of the discrepancies could be substantial, particularly for the largest value of Δt considered here.

TABLE 2 Travel Time (min) Versus Departure Time for Alternative Simulation Time Intervals (Δt), Sector 1

Departure Time (a.m.)	Δt (min)				
	Δt = 0.5	Δt = 1.0	Δt = 1.5	Δt = 2.0	Δt = 3.0
7:15	9.0	9.0	9.0	12.0	18.0
7:20	9.0	9.0	10.0	12.0	18.0
7:25	9.5	10.0	10.5	12.0	18.0
7:30	10.5	11.0	12.0	14.0	30.0
7:35	16.0	17.0	19.5	22.0	36.0
7:40	23.0	22.0	22.5	24.0	39.0
7:45	26.0	25.0	29.0	32.5	36.0

Note: In all cases Δm = 10 vehicles, 400 vehicles/sector, two lanes in each direction, v_f = 45 mph, v_0 = 6 mph, $\alpha = \pi$, and simulation duration = 60 min.

Fortunately, the computational efficiency of this model relieves to a great extent the concern for computation cost in selecting Δt. The computational effort is nonetheless quite sensitive to the size of Δt, as illustrated in the results, given in Table 3, of several numerical experiments.

TABLE 3 Effect of Simulation Time Step (Δt) on Computational Effort

Δt (min)	CPU[a]	CPU[b]
0.5	5.32	14.105
1.0	2.98	8.042
1.5	2.16	5.975
2.0	1.82	4.684
3.0	1.43	3.873

Note: Δm = 10 vehicles/particle, 400 vehicles/sector, two lanes in each sector, v_f = 45 mph, v_0 = 6 mph, $\alpha = \pi$, and simulation duration = 60 min.

[a]No intermediate report is included in the program output.

[b]Intermediate report of system performance and all vehicle positions during each simulation interval is included in the output.

Selection of Macroparticle Size Δm

As described in the previous section, vehicles bound for the same destination are grouped into macroparticles and thus enter the facility simultaneously. Conceptually, no problem would arise if each particle consisted of a variable number of vehicles, which would be specified as an additional attribute identifying each particle. However, for operational convenience, the use of identically sized macroparticles is suggested unless the number of vehicles arriving during Δg is insufficient to meet the prespecified size. In general, heavy traffic generation during the peak period can accommodate any reasonable predetermined particle size.

The specification of the particle size does not appear to significantly affect the accuracy of the simulation, as revealed by several numerical experiments. A summary of the results is given in Table 4, which depicts the travel time associated with various departure times under alternative macroparticle sizes (Δm = 5, 10, 15, and 20 vehicles per particle, respectively). The size of Δm does, however, contribute to the model's computational efficiency, although the marginal contribution of an increase in Δm beyond about 10 vehicles per particle decreases sharply, as seen in Table 5.

TABLE 4 Travel Time (min) Versus Departure Time for Alternative Macroparticle Sizes, Sector 1

Departure Time (a.m.)	Particle Size			
	Δm = 5	Δm = 10	Δm = 15	Δm = 20
7:15	9	9	9	9
7:20	9	9	9	9
7:25	10	10	10	10
7:30	11	11	11	11
7:35	16	17	16	15
7:40	28	26	26	25
7:45	28	29	28	28

Note: In all cases Δt = 1 min, 400 vehicles per sector, two lanes in each direction, v_f = 45 mph, v_0 = 6 mph, $\alpha = \pi$, and simulation duration = 60 min.

TABLE 5 Effect of Macroparticle Size (Δm) on Computational Effort

Δm	CPU[a]	CPU[b]
5	3.541	10.025
10	2.980	8.042
15	2.811	7.854
20	2.703	6.945

Note: In all cases, the following conditions hold: Δt = 1 min, 400 vehicles/sector, two lanes in each direction, v_f = 45 mph, v_0 = 6 mph, $\alpha = \pi$, and simulation duration = 60 min.

[a]No intermediate report is included in the program output.

[b]Intermediate report of system performance and all vehicle positions during each simulation interval is included in the output.

Note that in the selection of Δm, care must be taken to ensure that it is less than the maximum ramp entry rate (per unit simulation time step). A size of 5 vehicles per particle per lane (i.e., 10 vehicles for a facility with two lanes in each direction) has been found particularly convenient and appropriate in these experiments.

In the remainder of this paper an illustration is presented of the application of this model to the type of investigation for which it was developed. In particular, its application to the study of the dynamic behavior of traffic systems during peak-period congestion, under both mathematical decision rules and in conjunction with real commuters supplying departure time information, is described.

APPLICATION 1: SIMULATION EXPERIMENTS TO INVESTIGATE SYSTEM CONVERGENCE

This section contains a description of the use of the MPSM to explore the dynamics of a commuting corridor through a set of simulation experiments aimed at investigating the system's convergence status under alternative behavioral mechanisms. The latter are specified in the user-decisions subroutine introduced in the first section. Note that the presentation herein is intended primarily for illustrative purposes; further discussion of the details of the experiments, of the underlying assumptions, particularly those regarding user behavior, as well as of the results can be found in Mahmassani and Chang (10). In this section, the relevant features of the system and the type of behavioral rules that were considered are presented. Next, the key results per-

taining to system convergence that emerged from this particular set of experiments are highlighted.

System Features and Experimental Conditions

Supply-Side Input

The commuting corridor considered here is comprised of seven 1-mile sectors, with the common single work destination at the end of the seventh (and last) sector, as would be the case for CBD-bound work trips. Sectors are numbered from 1 to 7 in decreasing order of distance from the CBD. The highway facility consists of two lanes in each direction, although only the inbound direction is of concern here, with a mean free speed (v_f) of 40 mph and "jam" concentration (k_0) of 200 vehicles per lane-mile. The maximum entry rate, reflecting physical capacities as well as operational controls, is specified here as a constant of 80 vehicles per minute per sector. Note that a minimum speed of v_0 = 6 mph was specified in all cases to prevent a complete blockage of the system at high concentrations.

Demand-Side Input and User-Decisions Subroutine

The same number of users (in vehicle trips) was assumed for each of Sectors 1 through 6, with no trips generated in Sector 7, which can be viewed as a non-residential fringe sector. For simplicity, and without loss of generality, all commuters in these experiments were assumed to have the same work starting time of 8:00 a.m., although the model allows for a more general distribution of work start times.

Commuter behavior mechanisms of particular interest to this research application were specified in the user-decisions subroutine in order to provide daily departure time distributions. Two basic mechanisms can be distinguished in this subroutine: the first determines the acceptability, to a given user, of his most recent departure time, and the second determines the amount by which an unsatisfied commuter will adjust his departure time on the following day.

The first mechanism reflects a "satisficing," or boundedly rational view of commuter decisions in everyday situations (10,13). Essentially, it states that a user evaluates his departure time choice based on its outcome. A plausible rule is that users will accept a particular departure time, and maintain it on the next day, if the schedule delay, or the discrepancy between the resulting arrival time and the desired arrival time, is less than a certain tolerable level. This "indifference band" of tolerable schedule delay is viewed as a characteristic of each user (j) and is denoted by IB_j. It is natural to expect a distribution of IB_j across the population of commuters; the mean of this distribution, denoted by E[IB], is one of the experimental factors of concern in this research application. Three non-zero values of E[IB] were considered, namely, 5, 10, and 15 min, in addition to the extreme case of no indifference band. The distribution of IB_j across users was taken as a truncated normal distribution, coarsely discretized for the present purpose with a constant variance-to-mean ratio equal to 0.2.

The second mechanism, which specifies the departure time readjustment in response to previous days' performance, has been examined under two possible rules:

- A "myopic" rule, which says that the readjustment depends only on the latest day's schedule delay and
- A "learning" rule, whereby the readjustment explicitly takes into consideration the travel outcomes on more than just the latest day; typically, the relative weight accorded to recent experience is greater than that accorded to earlier experience.

The specific mathematical expressions for these mechanisms are presented in the Appendix. Simulation experiments were conducted to explore the dynamics of the system's behavior under the alternative myopic or learning readjustment rules. In addition, these experiments addressed the effect of two key factors: (a) the mean indifference band (E[IB]) for which the four different cases mentioned earlier were considered and (b) the usage level, whereby three different values were considered: a "reference" value V = 420 vehicles per sector, and V_1 = 0.6V and V_2 = 1.4V. The interaction between the usage level and the indifference band in determining the system's convergence properties is of particular interest and is described hereafter. Note that other factors, such as the parameters of the previously mentioned mechanisms (see Appendix) and the system's initial conditions (specified in terms of initial departure time distribution for each sector), were also explored, as described elsewhere (10).

Summary of Results

The corridor system was simulated for a period ranging from 50 to 70 consecutive days under each combination of values for the mean indifference band and the usage level, for both myopic and learning departure time adjustment rules. Convergence is examined in terms of the departure time distribution in each sector. When all users in a sector settle on their respective departure times, steady state is attained in that sector. For each simulation, the state of each sector after 70 days was recorded as one of the following: C for convergence to a steady state, O for regular oscillatory behavior, or NC for no convergence. A summary of this information is given in Table 6 for all the experiments in which the myopic adjustment rule was used and in Table 7 for those in which the learning-based rule was used. In both tables the number of days after which a steady state or regular oscillations were attained is indicated (in parentheses following the symbols C or O, respectively), where applicable. Note that in those situations where steady state or regular oscillations were not attained in all six sectors, partial steady state or oscillations, or both in only some of the sectors, are not necessarily guaranteed to be maintained indefinitely (10).

Effect of Mean Tolerable Schedule Delay

The data in both Tables 6 and 7 reveal that, all else being equal, the system is more likely to stabilize for larger values of E[IB]. This seems to be the case under both myopic and learning-based adjustment rules. For instance, consider the state of the system under experiments 2, 5, 8, and 11 in Table 7, corresponding to E[IB] values of 0, 5, 10, and 15 min, respectively. For the largest E[IB] value of 15 min, convergence to a steady state is reached in all sectors by Day 13. For E[IB] = 10 min, only partial convergence is attained, with the three sectors closest to the destination attaining a steady state, and under E[IB] = 5 min, convergence occurs in Sectors 5 and 6 only. Finally, under the assumption of no tolerable schedule delay, only sector 6 attains a steady state. This suggests the existence of a critical, or threshold, level of tolerable schedule delay above which a given system will be able to reach a steady state. This critical

TABLE 6 State of System After 70 Days, Myopic Behavior

Experiment No.	Usage Level[a]	Sector 1	2	3	4	5	6
No Indifference Band; E[IB] = 0							
1	0.6V	NC	NC	NC	NC	NC	C(14)
2	V	NC	NC	NC	NC	NC	C(37)
3	1.4V	NC	NC	NC	NC	NC	NC
Mean Tolerable Schedule Delay E[IB] = 15 Min							
4	0.6V	NC	NC	NC	NC	C(50)	C(56)
5	V	NC	NC	NC	NC	C(64)	C(12)
6	1.4V	NC	NC	NC	NC	NC	NC
E[IB] = 10 Min							
7	0.6V	C(8)	C(8)	C(18)	C(7)	C(8)	C(7)
8	V	O(46)	O(46)	C(45)	C(34)	C(7)	C(7)
9	1.4V	NC	NC	NC	NC	C(44)	C(10)
E[IB] = 15 Min							
10	0.6V	C(7)	C(7)	C(7)	C(7)	C(7)	C(7)
11	V	C(14)	C(13)	C(14)	C(13)	C(8)	C(4)
12	1.4V	NC	NC	NC	NC	C(9)	C(9)

Note: NC = no convergence by Day 70; C(I) = convergence achieved after I days; O(I) = regular oscillations started after I days.

[a]V is the reference value of 420 vehicles per sector.

TABLE 7 State of System After 70 Simulation Days, Learning Behavior

Experiment No.	Usage Level[a]	Sector 1	2	3	4	5	6
No Indifference Band; E[IB] = 0							
1	0.6V	NC	NC	NC	NC	NC	C(32)
2	V	NC	NC	NC	NC	NC	C(55)
3	1.4V	NC	NC	NC	NC	NC	NC
Mean Tolerable Schedule Delay E[IB] = 5 Min							
4	0.6V	NC	NC	NC	NC	C(8)	C(4)
5	V	NC	NC	NC	NC	C(36)	C(6)
6	1.4V	NC	NC	NC	NC	NC	NC
E[IB] = 10 Min							
7	0.6V	C(9)	C(10)	C(10)	C(9)	C(9)	C(4)
8	V	NC	NC	NC	C(52)	C(6)	C(6)
9	1.4V	NC	NC	NC	NC	C(48)	C(9)
E[IB] = 15 Min							
10	0.6V	C(6)	C(6)	C(7)	C(5)	C(6)	C(4)
11	V	C(11)	C(12)	C(11)	C(12)	C(13)	C(4)
12	1.4V	NC	NC	NC	NC	C(21)	C(9)

Note: NC = no convergence by Day 70; C(I) = convergence achieved after I days; O(I) = regular oscillations started after I days.

[a]V is the reference of 420 vehicles per sector.

level would essentially reflect the users' flexibility and willingness to compromise. The tighter this tolerable amount, the less likely it is that users will be able to reach a stable state (or a repeatable daily routine). It can also be noted that, all else being equal, the time needed to reach a steady state in a given sector appears to increase as the tolerable schedule delay increases.

Effect of Usage Level

The data in Tables 6 and 7 also reveal that, for a given E[IB], the system is more likely to converge at lower than at higher demand levels. Furthermore, higher usage levels (number of trip makers per sector) require a longer period to converge. For instance, consider experiments 10, 11, and 12 in Table 7, corresponding to usage levels of 0.6V (V = 420 vehicles per sector) and 1.4V, respectively, with E[IB] = 15 min. At the highest demand of 1.4V, only the last two sectors appear to have reached a steady state. At the reference demand level V, all sectors converge to a steady state as of Day 13. On the other hand, at the lowest usage level considered, 0.6V, convergence to steady state is reached in all sectors as of Day 7. The same general effect of the usage level can be observed in all the other experiments summarized in Tables 6 and 7.

In summary, there appears to be a strong inter-

relation between the usage level and the amount of tolerable schedule delay in determining the ability of the system to converge to a steady state, as well as the duration needed to achieve this convergence. Heavy traffic levels require wider indifference bands on behalf of system users, reflecting greater tolerance for delay, for a given system to equilibrate. It can also be noted that sectors closer to the destination generally attain their steady state sooner than more distant sectors and at lower values of E[IB] for a given usage level. This implies that more distant trip makers must be prepared to tolerate greater delay for the system to equilibrate, which is important in the design and evaluation of demand-side congestion relief measures.

APPLICATION 2: INTEGRATION OF SURVEY TECHNIQUES AND COMPUTER SIMULATION

Instead of employing prespecified mathematical expressions to supply users' decisions in the simulations, a number of real commuters were asked to take part in an experiment in which they provided daily departure time choices in a given hypothetical situation. Detailed description and in-depth analysis of the experimental results can be found elsewhere (4). A summary of this research and highlights of its results, with emphasis on congestion dynamics and system convergence, are presented in this section.

Description of Experiment

Much like the example application described in the previous section, the commuting corridor considered here consists of a highway facility used by adjoining residents for their daily home-to-work commute to a single work destination (CBD). This facility of two lanes in each direction is subdivided into nine identical 1-mile sectors, with the common work destination at the end of the last sector. Sectors are numbered from 1 through 9, with Sector 1 being the farthest outbound and Sector 9 the closest to the destination. The parameters used in the MPSM were v_f = 40 mph, v_0 = 6 mph, k_0 = 180 vehicles per lane-mile, and the maximum entry rate is taken as a constant equal to 60 vehicles per minute.

A total of 400 trip makers was assumed in each of the five residential sectors (from Sector 1 through Sector 5). A total of 100 participants was involved in this experiment, with each participant in effect representing 20 vehicles (or one macroparticle in each lane, with m = 10) in the MPSM program.

Each participant was initially given information about the highway facility (free speed, number of lanes) and his location, as well as the familiar decision situation of starting work at 8:00 a.m. Under those conditions participants were asked to state their desired arrival time and the corresponding departure time. The departure time decisions of all participants were obtained daily and input into the MPSM. The latest outcome of their choices, namely the respective actual arrival time on the previous day, as determined by the simulation, was provided to each participant individually on the following day, before his choice for that day was made. To relate the experiment to participants' actual daily commute, the survey was administered 5 days per week during the entire survey period.

Note that, because departure times were determined directly by actual commuters, there was no need to specify a user-decisions subroutine in this case.

SUMMARY OF EXPERIMENTAL RESULTS

System Convergence

This experiment covered 24 days, at which time all sectors in the system had already attained a steady state. The day-to-day evolution of average travel time over the period of the experiment is shown in Figure 3. It can be seen that the system's performance did not change as of Day 21. The steady-state values were actually attained on Day 18, exhibiting a fluctuation on Day 20 due to a few participants' (in Sector 1) unsuccessful attempt to improve their outcome, after which they reverted back to their steady-state choices (see Figure 4).

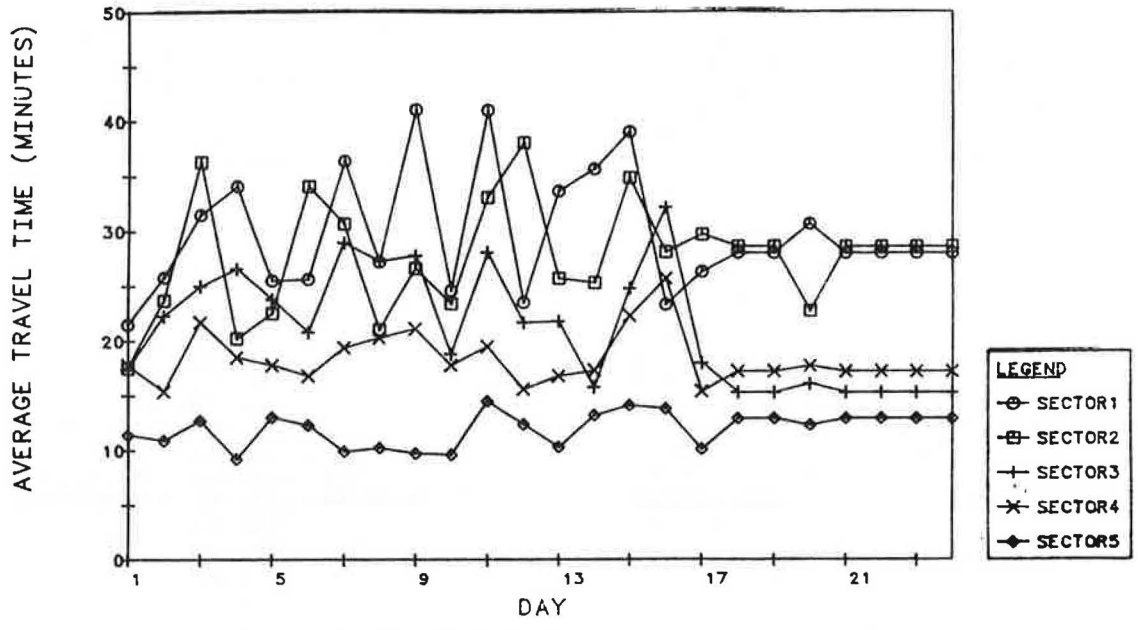


FIGURE 3 Evolution of average travel time for each sector.

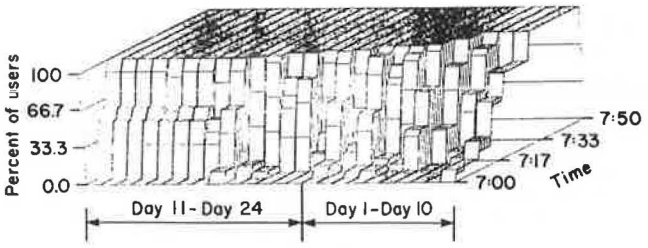


FIGURE 4 Cumulative departure pattern evolution for Sector 1.

An examination of the day-to-day evolution of the cumulative departure time distributions in each sector, considered separately, reveals that sectors with less travel distance tend to reach their steady-state values earlier than more distant sectors. This is exemplified by the day-to-day patterns for Sectors 1 and 5, respectively the farthest and closest, as shown in Figures 4 and 5. It can be seen in Figure 5 that all users in Sector 5 maintained their departure time choices as of Day 5, whereas Figure 4 indicates that the steady-state distribution in Sector 1 was completely reached as of Day 21. This phenomenon is consistent with the conclusions suggested by the simulation experiments described in the previous section.

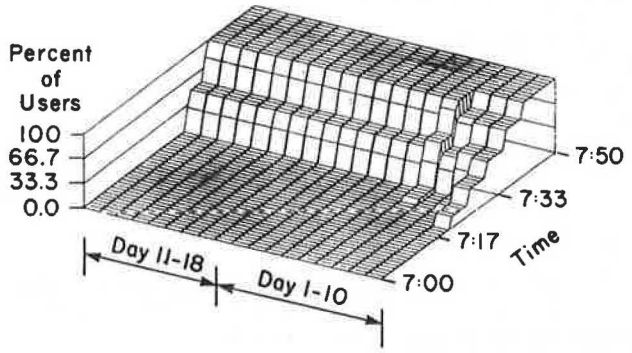


FIGURE 5 Cumulative departure pattern evolution for Sector 5.

The day-to-day variation of the average schedule delay of users in each sector is shown in Figure 6. It is worth noting that the average schedule delay, observed at steady state, increases in magnitude with increasing travel distance. In other words, commuters located in distant sectors have to tolerate longer schedule delays than those in near sectors in order to converge to a steady state. This interesting result is consistent with the conclusion of the simulation experiments described in the previous section.

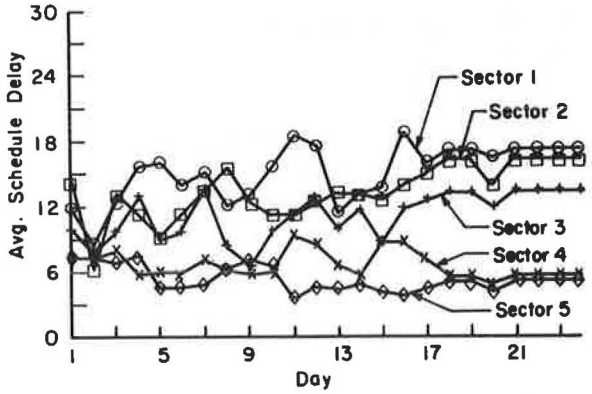


FIGURE 6 Evolution of average schedule delay for each sector.

Congestion Evolution and Travel Time Variability

The time-dependent pattern of traffic concentration in each sector is shown in Figures 7-11 for each of the first 5 days (first week), respectively. It can be seen that congestion was not severe on Day 1, with concentration in most sectors remaining below $2k_0/3$ (120 vehicles per lane-mile). A general worsening on Days 2 and 3 can be detected, with Sectors 3 through 5 and Sectors 7 through 9 manifesting

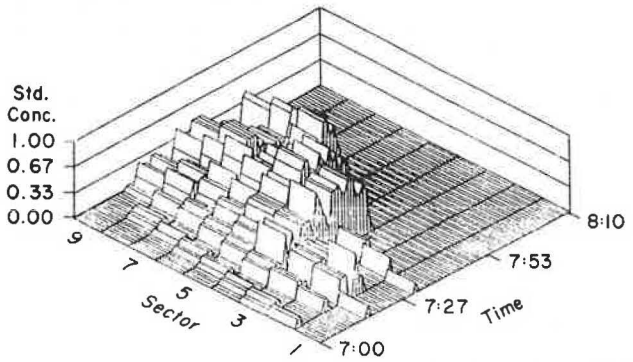


FIGURE 7 Time-dependent concentration pattern on Day 1, by sector.

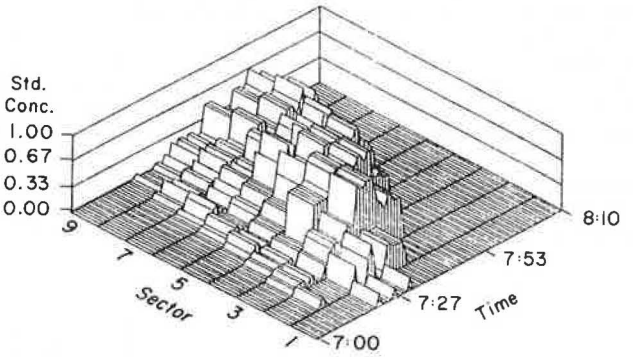


FIGURE 8 Time-dependent concentration pattern on Day 2, by sector.

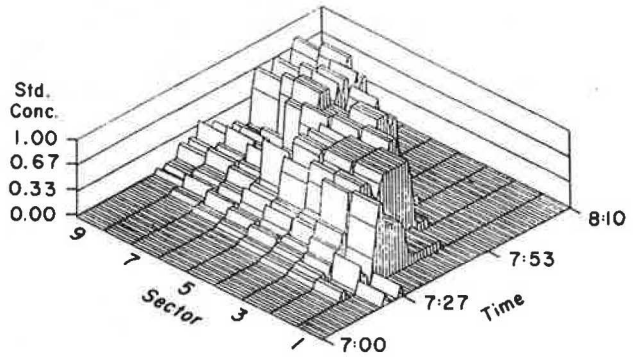


FIGURE 9 Time-dependent concentration pattern on Day 3, by sector.

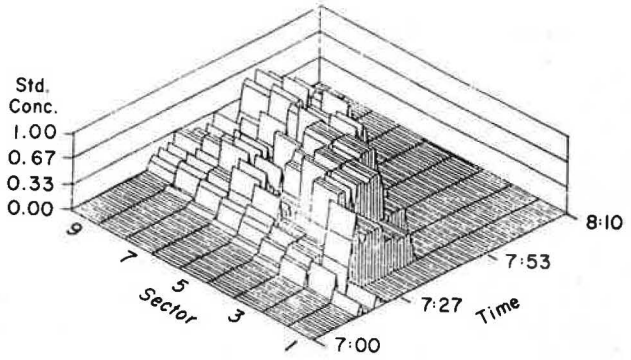


FIGURE 10 Time-dependent concentration pattern on Day 4, by sector.

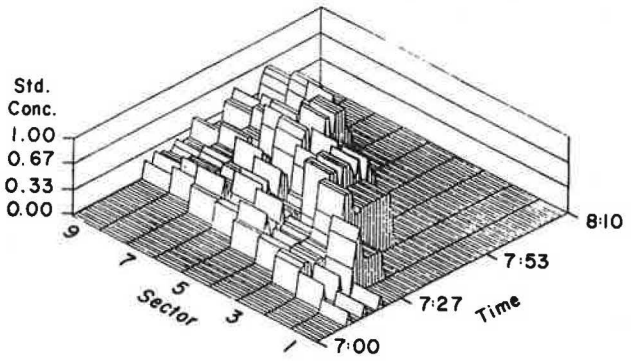


FIGURE 11 Time-dependent concentration pattern on Day 5, by sector.

relatively long periods of high concentration. Users' responses to these high levels led to a shift of the concentration distribution on Day 4 to an earlier time relative to Day 3 in all sectors; the overall pattern on Day 4 is quite similar to that of Day 2. Conversely, the temporal distributions of the concentration in all sectors moved forward on Day 5, resulting in a pattern that is similar to that of Day 3, except that the magnitude of the shift was less drastic on Day 5 than on Day 3. In other words, the range of the day-to-day fluctuation of concentration appeared to be narrowing until steady state was reached. This phenomenon can be attributed to the effect of commuters' previous experience with their departure time choices. Detailed discussion regarding commuters' departure time choice behavior and its interaction with congestion variation is outside the scope of the present paper and can be found elsewhere (10,14).

The steady-state time-dependent concentration pattern in each sector is shown in Figure 12, which reveals the high congestion periods for each sector. In addition, a typical example of the day-to-day evolution of the time-dependent concentration is shown in Figure 13 for Sector 3.

Table 8 gives a further summary of congestion patterns; of concern are both the occurrence of high concentration periods and their duration. Operationally, and somewhat arbitrarily, high congestion in a given sector was defined as the occurrence of 3 or more minutes within which concentration was continuously greater than or equal to $2k_0/3$. The duration of periods in which this criterion held, in each sector, is given on a daily basis in Table 8. Apparently, Sectors 3 through 5 experience the longest high congestion periods at steady state as well as

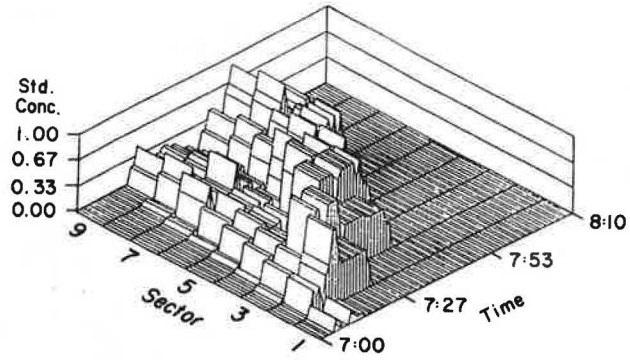


FIGURE 12 Time-dependent concentration pattern at steady state, by sector.

throughout the duration of the experiment. This congestion, with a time lag, spreads to downstream sectors, which experience shorter congestion periods overall.

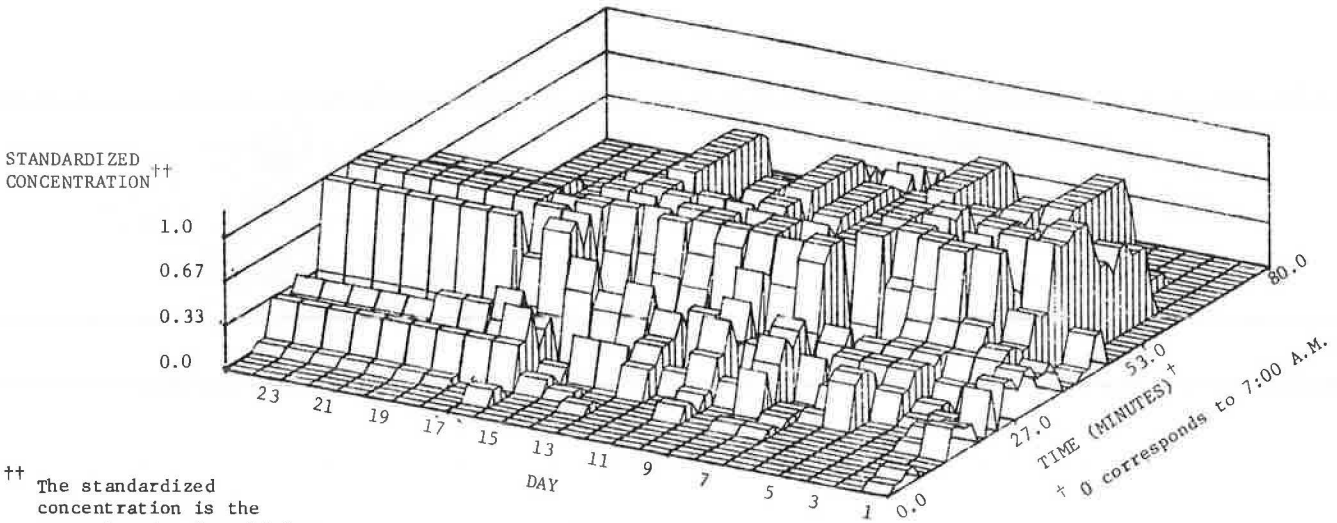
The record of travel time fluctuation for different departure times in each sector is one of the principal results of the MPSM. The steady-state pattern of travel time versus departure time [Figure 14 (d)] exhibits clear peaking characteristics with more distant sectors experiencing higher peaks than closer sectors. The peaking phenomenon, initially not very pronounced on Day 1 [Figure 14(a)], becomes more distinct as the system evolves, as illustrated by the patterns for Days 1, 6, and 11, and at steady state [Figure 14(a), (b), (c), and (d), respectively].

The extent of the sensitivity of travel time to the choice of departure time can also be seen from these patterns. For instance, if a commuter were to depart (at steady state) from Sector 1 at 7:10 a.m. instead of 7:05 a.m., he would experience as much as a 30-min increase in travel time. Similar situations also exist in other sectors, although sensitivity to departure time choice is somewhat less drastic in closer sectors.

TABLE 8 High Congestion Duration in Each Sector[a] (min)

	Sector								
Day	1	2	3	4	5	6	7	8	9
1	0	0	0	22	23	8	0	4	0
2	0	0	14	23	23	11	7	0	0
3	0	0	21	20	24	13	17	7	4
4	0	10	15	17	18	11	3	3	0
5	0	0	16	15	9	21	15	8	4
6	0	13	16	16	22	7	7	7	4
7	0	19	21	20	26	7	17	10	6
8	0	0	20	25	21	3	0	0	0
9	0	21	25	26	24	6	4	5	0
10	0	0	19	13	12	3	6	0	0
11	0	14	27	27	27	22	20	0	0
12	0	13	16	17	15	17	4	0	0
13	0	14	17	20	19	11	8	3	0
14	0	12	8	14	13	16	7	3	0
15	0	11	25	25	28	7	4	6	4
16	0	0	14	22	24	16	7	5	0
17	0	7	8	19	15	10	7	8	4
18	0	5	22	17	17	16	10	7	5
19	0	4	22	17	17	8	7	10	5
20	0	13	19	18	14	8	15	7	3
21[b]	0	4	22	17	17	8	7	10	5

[a]High congestion is defined as the occurrence of $k \geq 2k_0/3$ for more than 3 min.
[b]Days 22-24 are identical to Day 21.



†† The standardized concentration is the concentration in vehicles per lane-mile normalized by the "jam" concentration.

FIGURE 13 Evolution of time-dependent concentration pattern, Sector 3.

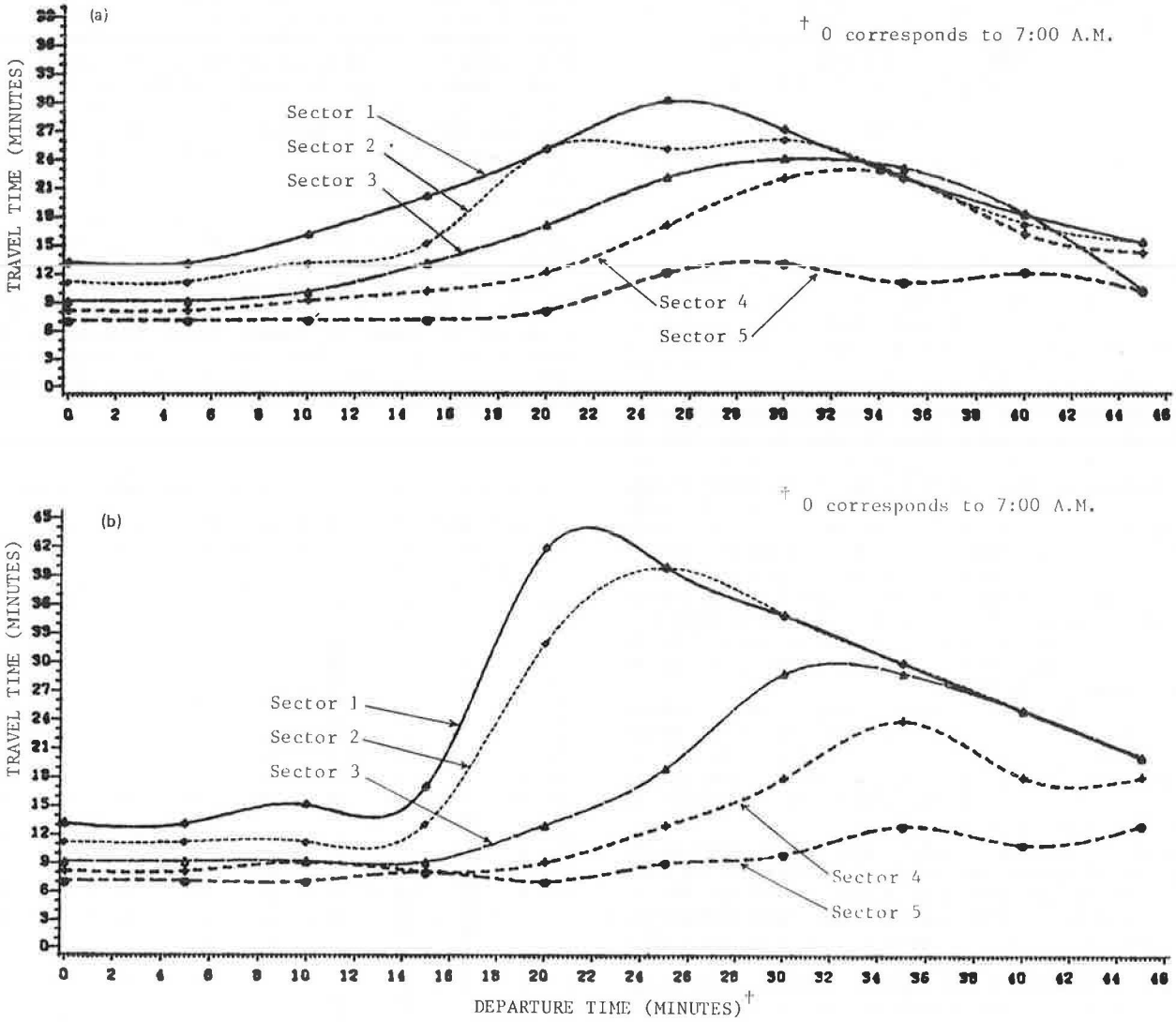


FIGURE 14 Travel time versus departure time on selected days, by sector: (a) Day 1; (b) Day 6; (c) Day 11; (d) Day 24, steady state.

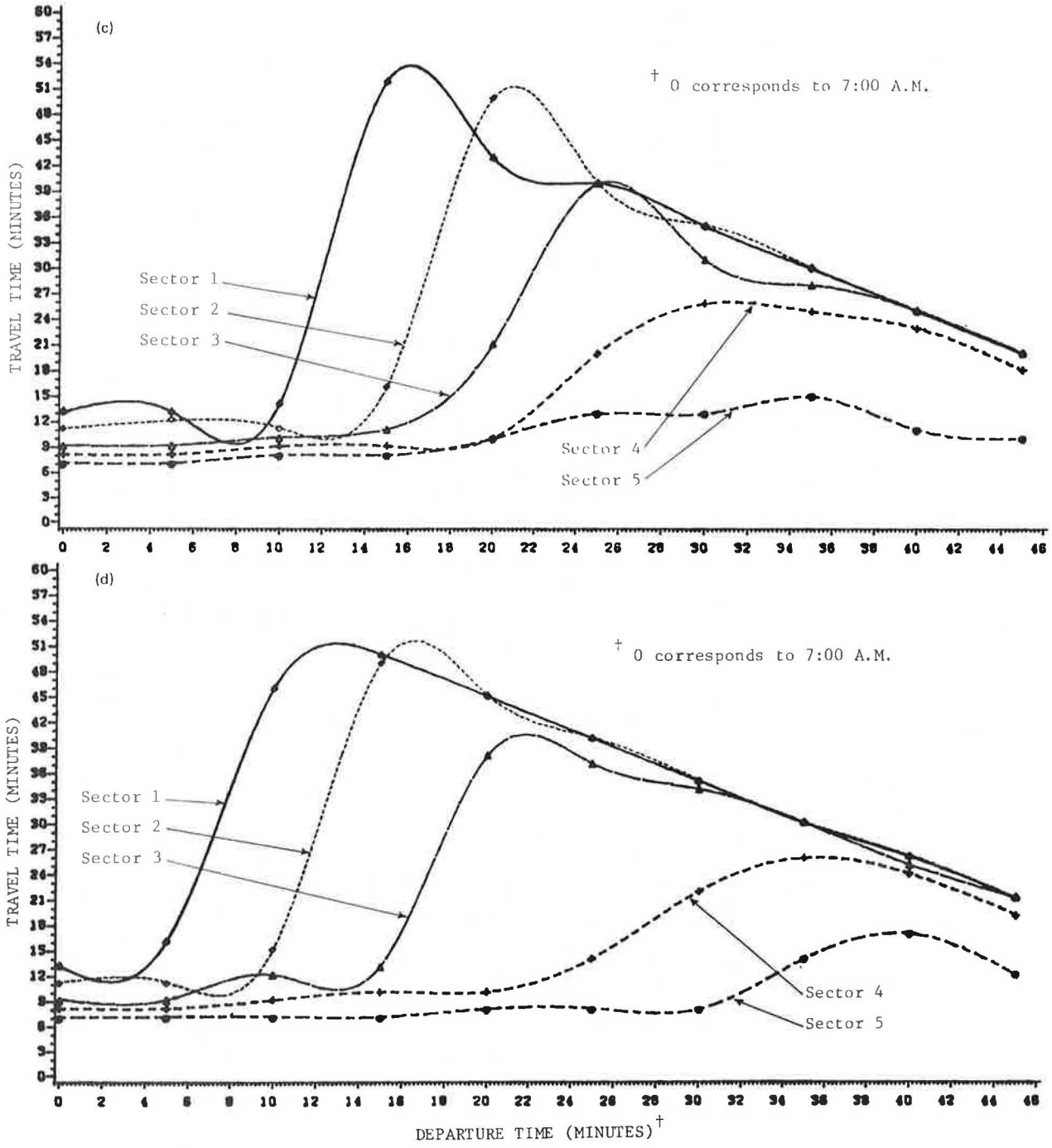


FIGURE 14 continued.

CONCLUSION

A special-purpose macroscopic model for the simulation of highway corridor traffic has been presented. The model was developed in conjunction with studies of the dynamics of peak-period traffic congestion in commuting systems and the interrelation between user decisions and congestion formation and dissipation. The model moves vehicles in small bunches or macroparticles using local speeds defined by the speed-concentration relation resulting in high computational efficiency, which is essential for the type of application for which the model is intended.

The details of merging and passing maneuvers are not captured by the simulation. However, the simulated system's behavior has been found adequate to pursue the investigation of peak-period congestion dynamics. Two such applications of the model have been presented: In the first, users' daily departure time decisions were simulated by the model using a set of specified dynamic decision rules. In the other application, a number of real commuters participated in a novel interactive experiment in which the traffic simulation model was used to determine each participant's arrival time, given all participants' departure time decisions for that day.

To further support this investigation of the dynamics of peak-period congestion, the following extensions and modifications of the MPSM are under way: (a) inclusion of more than one route, whereby

users can select both their departure time and their route; (b) better representation of traffic control devices, particularly at the entry points, to enhance the model's capability to evaluate traffic-based congestion relief measures; and (c) incorporation of the commuter's search for a parking space at the work destination, which is of particular relevance for downtown-oriented commuting in many large cities.

ACKNOWLEDGMENTS

The work reported herein was partly funded by a grant (to H. Mahmassani) from the Bureau of Engineering Research at the University of Texas at Austin. Computer funds made available by the Department of Civil Engineering, the Center for Studies in Statistical Mechanics, and the Physics Department at the university are gratefully acknowledged. Initial work on the development of the traffic simulation model was performed by Sonia Claudet, formerly a research scientist in the Center for Studies in Statistical Mechanics. In addition, the authors would like to thank T. Tajima, Department of Physics at the University of Texas at Austin, for alerting them to his work on the magnetohydrodynamic particle code for the simulation of plasmas. Finally, the assistance of Alfredo Hinojos in the preparation of the graphic material is appreciated. Support for the revision and continuation of this research is provided by NSF Grant CEE-8400306.

REFERENCES

1. C. Hendrickson and G. Kocur. Schedule Delay and Departure Time Decisions in a Deterministic Model. Transportation Science, Vol. 15, No. 1, 1981, pp. 62-77.
2. C. Hendrickson, D. Nagin, and E. Plank. Characteristics of Travel Time and Dynamic User Equilibrium for Travel-to-Work. In Proceedings of the Eighth International Symposium on Transportation and Traffic Theory, V.F. Hurdle, E. Hauer, and G.N. Steuart, eds., University of Toronto Press, Canada, 1983, pp. 321-347.
3. P.H. Fargier. Effects of the Choice of Departure Time on Road Traffic Congestion: Theoretical Approach. In Proceedings of the Eighth International Symposium on Transportation and Traffic Theory, V.F. Hurdle, E. Hauer, and G.N. Steuart, eds., University of Toronto Press, Canada, 1983, pp. 223-263.
4. A. De Palma, M. Ben-Akiva, C. Lefevre, and N. Litinas. Stochastic Equilibrium Model of Peak Period Traffic Congestion. Transportation Science, Vol. 17, No. 4, 1983, pp. 430-453.
5. H. Mahmassani and R. Herman. Dynamic User Equilibrium Departure Time and Route Choice on Idealized Traffic Arterials. Transportation Science, Vol. 18, No. 4, 1984, pp. 362-384.
6. D.N. Goodwin, S.D. Miller, and H.J. Payne. MACK: A Macroscopic Simulation Model of Freeway Traffic. Technology Service Corp., Santa Monica, Calif., 1974.
7. H.J. Payne. FREFLO: A Macroscopic Simulation Model of Freeway Traffic. In Transportation Research Record 722, TRB, National Research Council, Washington, D.C., 1979, pp. 68-75.
8. E.B. Lieberman. Dynamic Analysis of Freeway Corrdior Traffic. Presented at the Joint Transportation Engineering Conference, American Society of Mechanical Engineers, Chicago, Ill., 1970.
9. E.B. Lieberman. Simulation of Corridor Traffic: The SCOT Model. In Highway Research Record 409, HRB, National Research Council, Washington, D.C., 1972, pp. 34-45.
10. H.S. Mahmassani and G.-L. Chang. Experiments with Departure Time Choice Dynamics of Urban Commuters. Transportation Research B (in press).
11. A.D. May. Models for Freeway Corridor Analysis. In TRB Special Report 194: The Application of Traffic Simulation Models, TRB, National Research Council, Washington, D.C., 1981, pp. 23-32.
12. N.A. Derzko, A.J. Ugge, and E.R. Case. Evaluation of a Dynamic Freeway Flow Model Using Field Data. Presented at 62nd Annual Meeting of the Transportation Research Board, Washington, D.C., 1983.
13. J.N. Leboeuf, T. Tajima, and J.M. Dawson. A Magnetohydrodynamic Particle Code for Fluid Simulation of Plasmas. Journal of Computational Physics, Vol. 31, No. 3, June 1979, pp. 379-408.
14. H.S. Mahmassani, G.-L. Chang, and R. Herman. Individual Decisions and Collective Effects in a Simulated Traffic System. Transportation Science (in press).

APPENDIX: EQUATIONS FOR USER-DECISIONS SUBROUTINE

Two basic equations govern the dynamics of user departure time decisions in the commuting corridor. The first equation simply determines acceptability of a given departure time on the basis of its outcome. Let $D_{j,t}$ and $AT_{j,t}$ denote user j's departure time and the resulting actual arrival time at work, respectively, on day t. The mechanism for the acceptability of this outcome used in the simulation experiments, and based on the notation of indifference band, states that

$$\delta_{j,t} = \begin{cases} 1 & \text{if } |AT_{j,t} - DAT_j| \leq IB_j \\ 0 & \text{otherwise} \end{cases}$$

where $\delta_{j,t}$ is a binary variable that takes the value 1 if the actual arrival time on day t is acceptable to user j and 0 otherwise, DAT_j is the desired arrival time of user j, and IB_j is the indifference band, or tolerable schedule delay, of user i. Note that different indifference bands could be specified for earliness as opposed to lateness (10); however, this distinction was not made in the experiments reported in this paper, for the sake of simplicity. Furthermore, IB_j could be specified as varying from day to day, though it was treated as constant, for a given individual in these experiments, for lack of a clear explanatory mechanism for this daily variation. This aspect is the subject of ongoing research by the authors.

The second equation governing user behavior is the departure time readjustment whereby $D_{j,t+1}$ is determined given that $\delta_{i,t} = 1$. Two alternative rules were described in the third section: a myopic rule and a learning-based rule. In both cases, the commuter can be viewed as setting $D_{j,t+1}$ by subtracting his anticipated travel time $ATT_{j,t+1}$, on day t+1, from his desired arrival time, as follows:

$$D_{j,t+1} = DAT_j - ATT_{j,t+1}$$

The myopic and learning rules differ in how $ATT_{j,t+1}$ is obtained given user j's prior experience.

Rule 1: Myopic Adjustment

$$ATT_{j,t+1} = TT_{j,t} + a\gamma^{e}_{j,t}(AT_{j,t} - DAT_j) + b\gamma^{\ell}_{j,t}(AT_{j,t} - DAT_j)$$

where

$TT_{j,t}$ = actual trip time experienced by user j on day t;
$\gamma^{e}_{j,t}$ = a binary variable equal to -1 if $(AT_{j,t} - DAT_j) < 0$ (i.e., user j is early on day t) and 0 otherwise;
$\gamma^{\ell}_{j,t}$ = a binary variable equal to -1 if $(AT_{j,t} - DAT_j) > 0$ (i.e., user j is late on day t) and 0 otherwise, and
a,b = two parameters in the interval [0,1].

The values a = 0.5, b = 0 were used in the experiments described in the third section of this paper. Sensitivity analysis with respect to these parameters as well as an in depth discussion of the embedded behavioral assumptions can be found elsewhere (10).

Rule 2: Learning-Based Adjustment

Here $ATT_{j,t+1}$ is a function of all prior experience with the system, as follows:

$$ATT_{j,t+1} = \sum_{\ell=t_o}^{t} w_\ell \, (TT_{j,\ell})$$

where, w_ℓ, $\ell = t_o,\ldots,t$ denotes a set of non-negative weights attached to each day, starting with the initial day t_o. It is expected that users attach greater weight to more recent days than to earlier ones. Therefore, the following special form of the equation was used in the experiments:

$$ATT_{j,t+1} = [(1 - w_t)/(t - t_o)] \left(\sum_{\ell=t_o}^{t} TT_{j,\ell} \right) + w_t \, TT_{j,t}$$

where $0 < w_t \leq 1$. Thus all days before the last one are given a total weight of $(1 - w_t)$, equally allocated among all prior days; the last day is given a weight w_t. In the experiments reported in the third section of this paper, a value of 0.5 was used for w_t. The effect of the value of this parameter on system behavior does not, however, affect the general conclusions of the third section, as shown elsewhere (10).

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

Creation of Data Sets To Study Microscopic Traffic Flow in Freeway Bottleneck Sections

STEVEN A. SMITH and MARK E. ROSKIN

ABSTRACT

The methodologies employed in an FHWA research study entitled "Freeway Data Collection for Studying Vehicle Interactions" are described. The purpose of this study was to develop a series of data sets on microscopic vehicular traffic flow for selected types of freeway sections. The methodology used to develop these data sets involved digitizing vehicle positions from time-lapse aerial photographs of a series of freeway sites with various geometric configurations. Six types of freeway geometry were of interest: ramp merges, weaving sections, upgrade sections, reduced-width sections, lane drops, and horizontal curves. The aerial photography involved the use of a full-frame 35-mm motion picture camera operating in time-lapse mode mounted in a fixed-wing, short-takeoff-and-landing (STOL) aircraft. The sites were filmed at one frame per second with the aircraft flying clockwise at a slow speed around each site at altitudes ranging between 2,500 and 4,500 ft. Data were reduced to 1 hour of film (3,600 frames) of each site. Sites ranged between 1,200 and 3,200 ft in length. The data reduction method involved a microcomputer-based digitizing system. The most important components of the system were the mathematical techniques for computing vehicle position and the method of vehicle matching, which yielded complete vehicle trajectories for all vehicles passing through the sections studied. The data sets are expected to be useful both for empirical research on freeway traffic flow and for the validation of freeway simulation models. The data sets are being made available to those conducting research in these areas.

The methodologies used in an FHWA research study entitled "Freeway Data Collection for Studying Vehicle Interactions" (Contract DTFH61-82-C-00001) are described. The purpose of this study was to develop a series of data sets on microscopic vehicular traffic flow for selected types of freeway sections. The methodology used to develop these data sets involved digitizing vehicle positions from time-lapse aerial photographs of a series of freeway sites with various geometric configurations. Six types of freeway geometry were of interest:

- Ramp merges,
- Weaving sections,
- Upgrade sections,
- Reduced-width sections,
- Lane drops, and
- Horizontal curves.

There were several reasons for selecting these types of sections for analysis. These types of geometric configurations are the most frequent causes of bottlenecks or recurrent congestion points on freeways, and therefore a study of these types of sections is likely to provide the greatest benefit to overall freeway operations. In addition, these types of sections represent some of the more difficult situations for which to accurately simulate freeway traffic flow in mathematical models. Data on microscopic vehicle movements through such sections are expected to be useful in enhancing freeway simulation models as well as in direct empirical research.

The objective of the study was to develop data sets that could be used in the study of traffic flow in these sections. No actual analysis of the data was to be performed in the study. The data sets are being made available to other researchers who will use the data to study particular aspects of traffic flow that are of interest to them. Information on how to request a copy of a data set is presented later in this paper. The following sections discuss the operational characteristics of freeway bottleneck sections and the current need for data on vehicle interactions, as developed in this study.

OVERVIEW OF METHODOLOGY

The collection of data for studying microscopic traffic flow in freeway bottleneck sections implies the need for detailed data on vehicle trajectories. This is the most difficult type of traffic data to obtain because the position of all vehicles must be known for short time increments (1 to 3 sec) and vehicles must be completely traced through the section.

There are two basic approaches to developing detailed vehicle trajectory information. The first approach involves the placement of closely spaced pairs of axle detectors on the roadway. The detectors are used to record the exact time of each axle crossing. This enables the identification of vehicles by axle characteristics and the matching of these characteristics from one detector to the next, thus tracing the vehicles through the section. Devices such as FHWA's Traffic Evaluator System (TES) have been used to develop vehicle trajectories using this method. However, the number of detectors required to track vehicles through a section in short time increments is not practical for the lengths of sections required in this study. In addition, the TES software has not been designed for vehicle matching under congested flow conditions.

The second approach to developing vehicle trajectory data involves the photographic tracing of vehicles. In this approach vehicles are tracked through the section of interest and their positions recorded at discrete points in time. This approach has been attempted in the past with varying degrees of success. One of the methods used for matching vehicles in these photographic methods has been to record the position of vehicles at short time increments (e.g., 1 sec) and to match vehicles from one frame to the next using a computer algorithm based on the position of vehicles and on known speed and lane-changing behavior. This method was used in a study by UCLA and System Development Corporation (1) to reduce vehicle trajectory data from aerial 70-mm films. Data were collected at three sites, and the data were reduced by the transformation of vehicle positions into digitized form. However, numerous problems were encountered in the vehicle matching algorithm, and an evaluation of the UCLA/SDC experience by Raudseps (2) was not terribly optimistic about this approach. This technique was also used in a study by Garner and Mountain in the United Kingdom (3), but an approximate 85 percent matching rate was reportedly all that could be achieved.

The data collection and reduction method adopted for this study involved an aerial photographic approach in conjunction with a microcomputer-based digitizing system. The key to the success of the system was the method of vehicle matching, which yielded complete vehicle trajectories for all vehicles passing through the sections studied. The matching method involved the digitizing of all vehicles within defined section limits on each successive frame of film, relying on the operators' ability to match vehicles on the basis of four key characteristics:

- Order of vehicles in the previous frame,
- Color of vehicle,
- Type of vehicle, and
- Lane vehicle was in in the previous frame.

The operator was assisted in the matching process by error checks built into the computer software. The operator matching approach (with computer assist) is superior to the computer matching approach because the operator-computer interaction allows many potential errors to be caught before entry of data into the working data files.

Some researchers have suggested that a fully automated system might be employed using the image recognition capabilities of video-recording methods. Although this method is conceptually attractive, both eliminating the need for operator matching and potentially accelerating the data reduction process, major advances need to be made in video resolution, image recognition, and matching algorithms before this technique will be feasible for the lengths of freeway sections studied here. It is possible that video methods may be developed for applications that require considerably larger-scale images.

The aerial photography involved the use of a full-frame 35-mm motion picture camera operating in the time-lapse mode mounted in a fixed-wing, short-takeoff-and-landing (STOL) aircraft. The sites were filmed at one frame per second with the aircraft flying clockwise at a slow speed around each site at altitudes ranging between 2,500 and 4,500 ft. Data were reduced to 1 hour of film of each site. Sites ranged between 1,200 and 3,200 ft in length. The sites included all six of the types of sections discussed previously. The following sections describe the filming and data reduction procedures in detail.

AERIAL PHOTOGRAPHY

Initial Experiments

Aerial photography proved to be one of the more difficult aspects of the study, given the stringent filming requirements and the budget limitations. A pilot study was conducted in which experiments were made using various combinations of film formats and aircraft until an optimum filming method was achieved. Requirements for the filming were as follows:

- The same section had to be filmed continuously at one frame per second for more than 1 hour,
- The camera angle had to be as nearly vertical as practical to maximize the accuracy of measurements, and
- The film format had to be as small as possible for reasons of economy and yet of high enough resolution for all vehicles to be distinguished.

Experimentation with the photography was begun with a light plane circling in a tight radius and testing both 16-mm and half-frame 35-mm film formats. (Note that a half-frame 35-mm film frame is half the size of a standard 35-mm slide and a full-frame 35-mm film frame is the same size as a 35-mm slide.) A camera mount was constructed and affixed to the floor of the aircraft enabling the camera to be angled toward the ground as the aircraft was banked in a continuous circle around the freeway section being photographed. In this first experiment it was found that the aircraft (a Cessna 206) did not provide a stable enough platform and could not fly sufficiently slowly to enable the photographer to keep the camera continuously on the section. The 16-mm film format was clearly unacceptable for the length of sections studied. The 35-mm half-frame format, although considerably better, was still insufficient to obtain the resolution required.

A brief experiment was also conducted using a hovering helicopter and 35-mm half-frame format. This configuration was also found to be impractical. Although 1 hour of film could conceivably be obtained with this method, ideal conditions of wind velocity and direction would be needed at every site to avoid overheating of the helicopter engine. These highly restrictive requirements and the cost of helicopter rental eliminated this method from further consideration.

The final experiment involved the use of a Heli Porter STOL aircraft and a full-frame 35-mm film format. This configuration proved to meet all of the filming requirements and was therefore adopted for use. Enough test film was shot in the fall of 1982 to allow further development of the data reduction process. Although 70-mm film was considered in these experiments, the cost of the film and of projection equipment was substantially higher than for 35-mm film, and only 40 min of continuous 70-mm filming could be obtained with the available film magazine size.

Filming of Freeway Sections

The actual filming was undertaken in the spring of 1983 on freeway sections in both the Washington, D.C., and the Los Angeles metropolitan areas. A Helio-Courier STOL aircraft was used along with a 35-mm full-frame camera and a 1,000-ft film magazine. This magazine size enabled two sites to be filmed back-to-back in one flight, speeding up the filming process and reducing filming costs considerably. The camera was a Flight Research Model 207 equipped with a high-resolution Nikkor 35-mm focal length lens and automatic exposure control. The automatic exposure control was an important feature because camera angle with respect to the sun was continuously changing. The frame number was superimposed on the film image. The camera was angled out the rear right-side door, and the camera operator was responsible for keeping the freeway section continuously in view. Figure 1 shows the aircraft and Figure 2 shows the camera mounted in the aircraft.

Filming was done primarily in the evening peak period and sometimes in midafternoon, depending on traffic conditions. The characteristics of both the

FIGURE 1 STOL aircraft.

FIGURE 2 35-mm camera mounted in aircraft.

site itself and the traffic conditions on the site were critical. Some of the site requirements included the following:

- Traffic levels of service should be in the C to E range, depending on the type of site. For most sites an effort was made to include the transition period from uncongested flow to congested flow.
- There must be no condition downstream of the site that influences congestion within the site. In other words, there can be no backup from a downstream location into the site area.
- There should be few or no bridge structures passing over the site, and those that are present should not be so wide as to obscure vehicles from view in the films.
- For best film resolution, the site should generally be no longer than 3,000 ft. The average site was approximately 1,800 ft long.
- The peak traffic time to be filmed must occur when light conditions are favorable for filming. Image quality significantly deteriorates with low sun angles.
- Suitable locations must exist for establishing and placing control points.
- The site should not be so incident prone that obtaining satisfactory film footage of incident-free traffic flow is unlikely.

No sites were filmed in the morning peak period because of low sun angles that existed at the times the filming would have had to be begun. Other than this, the traffic flow requirement was the most difficult criterion to satisfy in the site selection process.

A total of 18 sites was filmed, 8 in Washington and 10 in Los Angeles. The following numbers of sites were obtained:

- Weaving sections (7 sites, not all under the preferred traffic conditions);
- Ramp merges (3 sites);
- Reduced-width sections (2 sites);
- Upgrades (2 sites);
- Horizontal curves (2 sites); and
- Lane drops (2 sites, not all under the preferred traffic conditions).

Before filming, a set of targets, which would be visible in the film, was set out on the right shoulders of each direction of travel on the freeway sections to filmed. The targets consisted of 3- to 4-ft squares of dayglow orange plastic material with nylon mesh and were nailed to the pavement the day of or the day before filming. Typically four pairs of targets were laid out on each section to be filmed. The relative position of the targets was established by a ground survey. These positions served as control points and were used to establish a known ground coordinate system for the data reduction process.

DATA REDUCTION

Digitizing

The data reduction system consisted of a digitizing tablet and processor, a microcomputer and terminal, a voice synthesizer, and a 35-mm full-frame filmstrip projector. The system components are shown in Figure 3. A SAGE IV microcomputer with 512K bytes of memory and an 18-megabyte hard disk was used, and UCSD Pascal was employed as the programming language. The microcomputer was operated under a multiuser configuration, enabling two digitizing systems to be operated simultaneously.

The digitizing process is controlled by the operator primarily through the digitizer keypad. The keypad can be used to generate and transmit to the computer the X and Y coordinates of any point on the active area of the digitizing surface and to transmit numerical codes entered by the operator. Computer algorithms are used to transform digitized positions on the film to meaningful information on vehicle positions. The voice synthesizer is used to prompt the operator with audible feedback from the computer.

The key components of the data reduction process are the algorithms used to compute vehicle position and the method of tracking vehicles through the section. A two-stage process is used to compute vehicle position. First, a photogrammetric technique is used to translate the X-Y coordinates digitized from the projected film image (which is in a perspective view) into the coordinate system of the ground established with the control points (plan view). The specific technique used is termed "projective transformation" and involves the use of eight simultaneous equations to compute the coefficients of equa-

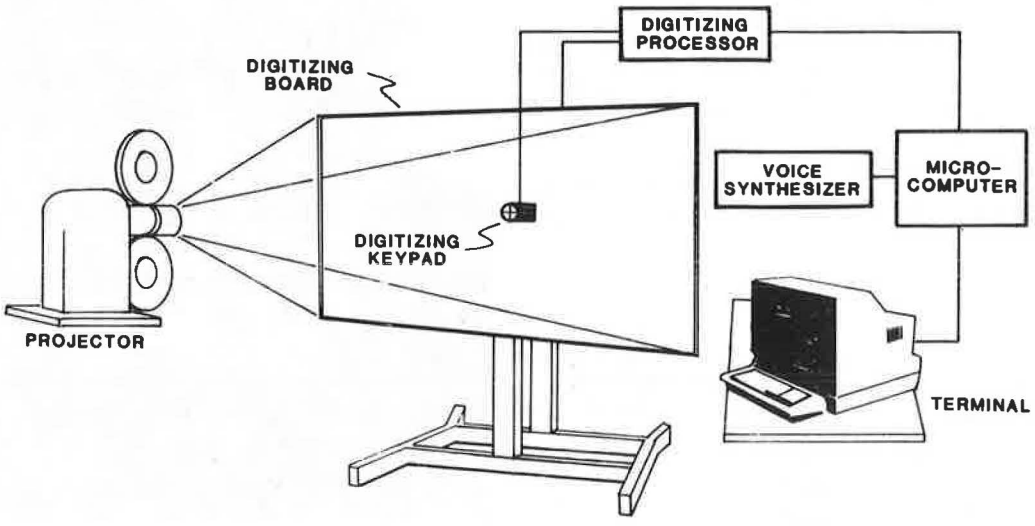


FIGURE 3 Schematic of digitizing system components.

tions that are, in turn, used to compute the X and Y coordinates of the ground plane for any digitized point on the film. This technique has been well described in previous publications (3,4) and the details need not be repeated here.

Figure 4 shows the relationship between the plane of the film and the plane of the ground and how the digitized position of a vehicle would be translated from one plane to another. The primary rule governing the establishment of the control points is that they lie in the same plane and that that plane follow the vertical alignment of the highway as closely as possible. All vehicles or points digitized will thus take on the coordinates of that plane. If changes in vertical alignment occur within a section, separate planes, with separate sets of control points, are needed to adequately carry out the transformation to ground coordinates. Any points digitized that are not in that plane will introduce a parallax error if the point is not viewed directly from above. This was particularly important for this study because all highway sections were photographed at an oblique angle.

Other rules governing the use of control points included the need to maintain all of the internal angles of the quadrilateral reasonably close to right angles and to make all sides of the quadrilateral of sufficient length. It was found that the width of the highway did not generally provide sufficient lateral distance between control points to enable stability to be achieved in the film-to-ground coordinate transformation. Therefore, a new set of control points was often established on one side of the highway using points outside the highway right-of-way that were also visible on the film (e.g., corners of rooftops) and were roughly within the plane of the highway. Other rules were also applied to minimize transformation errors.

The second coordinate transformation process involved translating the coordinate system of the ground into the coordinate system of the highway. The highway coordinate system was established to follow the horizontal alignment of the highway. The longitudinal axis was parallel to the highway centerline, coinciding with the right edge of the main line, and the lateral axis was always perpendicular to the longitudinal axis. The longitudinal coordinates began with zero at the upstream end of the section and ended at the downstream end of the section. The lateral coordinates were measured from the baseline at the right edge of the main line. The positive direction was defined as being to the left and the negative direction was to the right, so that the positions of any vehicles on an off-ramp or on-ramp took on a negative value.

The mathematics used to translate ground coordinates to highway coordinates were based on standard trigonometric equations used in highway design. More details on the mathematics are available in the final report (5). Subsections were defined by highway sections with homogeneous geometry. A number of the sections were either entirely tangent or entirely curved throughout, whereas others required up to three subsections.

The digitizing process is shown in simplified form in Figure 5. The figure shows the basic iterations involved but does not show the error checking and correction features, the initial creation of the geometric profile, and other more detailed aspects of the process. The digitizing of each frame of film begins with entering the next frame to be digitized. The four control points are then digitized and, following a computer calibration, a fifth point with known position is used to check the calibration.

Following a successful calibration, the computer recalls the most downstream vehicle digitized in the previous frame and, through the voice synthesizer, prompts the operator with several items of information: (a) the color of the vehicle, (b) the vehicle type, and (c) the number of the lane in which the

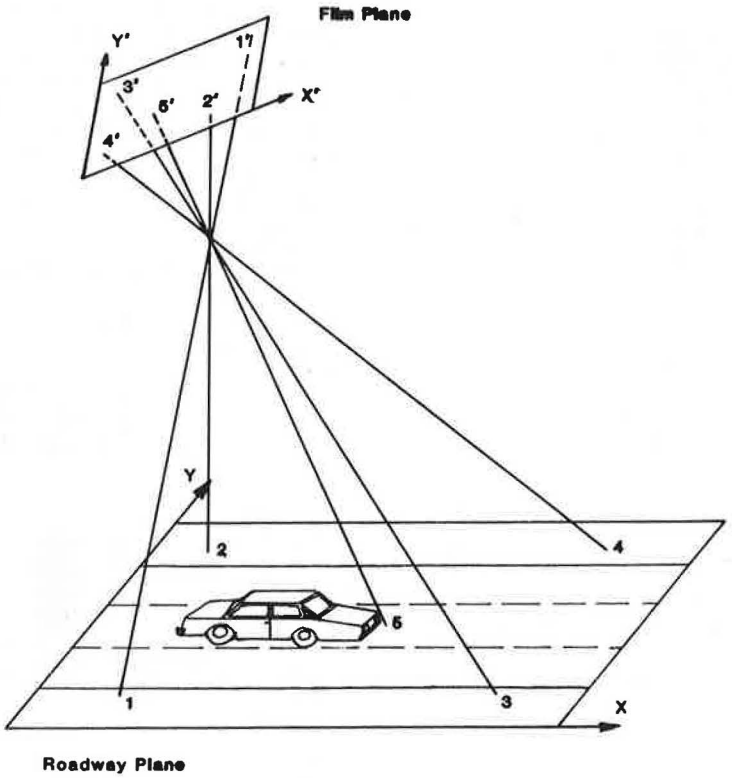


FIGURE 4 Relationship of points in roadway plane to points in film plane.

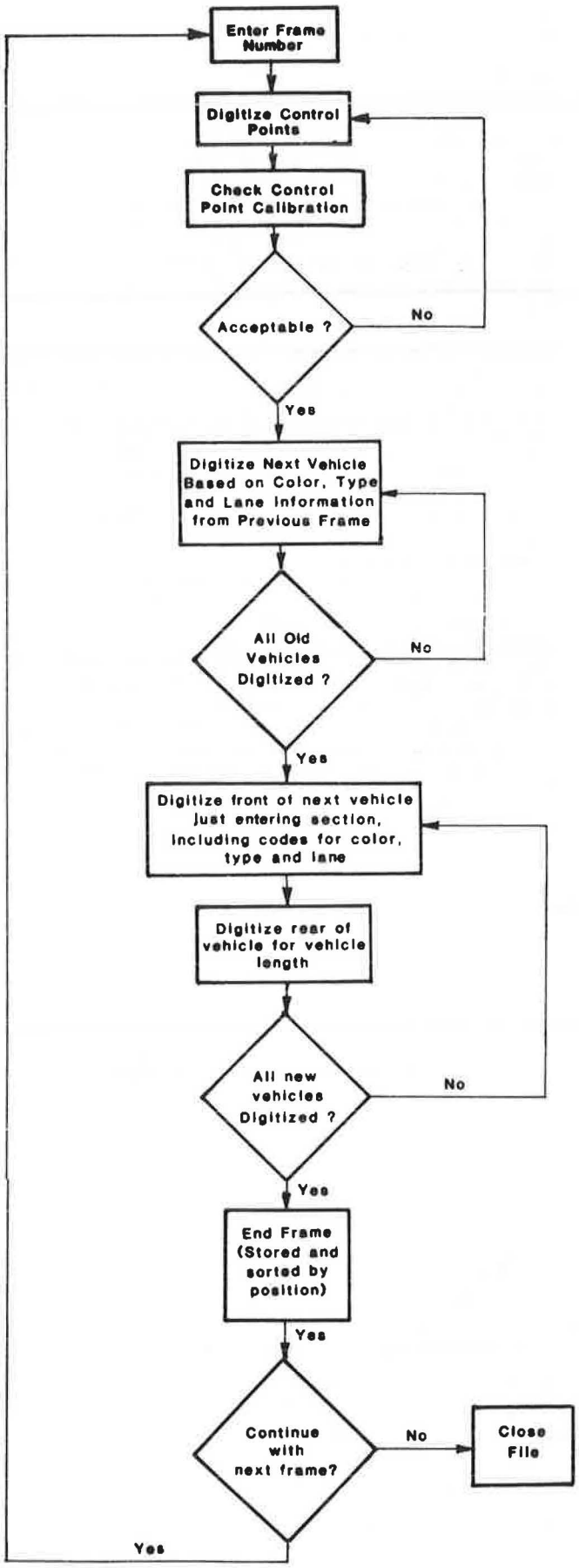


FIGURE 5 Overview of digitizing process.

vehicle was positioned in the previous frame. On the basis of this information, the operator finds the vehicle matching that description, beginning the search at the downstream end of the section. The cross hairs of the digitizer's cursor pad are placed over the center of the front bumper of the vehicle, and the key corresponding to the current lane number of that vehicle is depressed. The depression of the key on the cursor pad transmits the X-Y coordinates of that point on the projected film image and the lane number to the computer. A vehicle may have changed lanes since the last frame, so the lane number may have changed since the previous frame. The computer then performs the film-to-ground and ground-to-highway coordinate transformations and a series of checks to screen out possible errors made by the operator (particularly digitizing the wrong vehicle). Provided the computer finds no errors in the operator's digitizing of that vehicle, a new record is created in the vehicle file. The new record includes the frame number, a unique identification for that vehicle, the vehicle's lateral and longitudinal position, and other data such as the vehicle's color, type, and length. The computer then prompts the operator with the next upstream vehicle, which the operator then locates in the current frame, just as was done with the first vehicle. The digitizing proceeds in the upstream direction until all of the vehicles that appeared in the previous frame are digitized in the current frame. Vehicles that are entering the section for the first time are then digitized. The vehicle color and type are entered along with the lane number, using the digitizer's cursor pad. Vehicle length is also obtained by digitizing the rear of the vehicle the first time it enters the section.

The accuracy of the longitudinal and lateral position is dependent on a multitude of factors. The most important control on errors is the placement of the control points and the accuracy of the ground measurements taken. Ideal control point configurations are not always possible. As the points being digitized become farther from either of the control point pairs, the higher the error is likely to be. Points digitized close to a control point will be subject primarily to errors by the operator in placing the cross hairs of the cursor over the true position of the vehicle. These errors are typically 2 ft or less. Errors arising from control point calibration are typically greater in the longitudinal dimension than in the lateral dimension because the distance from a control point is likely to be greater in the longitudinal than in the lateral dimension. The overall positional error is estimated to be within ±5 ft in the longitudinal direction and ±3 ft in the lateral direction.

The prompting of the operator by the computer is done primarily through a general purpose speech synthesizer. The color, type, and lane number of the next vehicle to digitize are given audibly, enabling the operator to continually keep his or her eyes on the projected film image instead of having to look back and forth between a CRT screen and the digitizing surface. When an error or special situation is encountered, a message is sent to the operator through the speech synthesizer to look at the CRT screen where additional information is displayed. The speech synthesizer not only saves substantial amounts of time in the digitizing process but also reduces both operator fatigue and the probability of error.

The digitizing rate varied with the quality of the film. Sections with asphalt pavement tended to be more difficult to digitize because of the greater difficulty in seeing dark-colored vehicles on the dark pavement. Better contrast was achieved on con-

crete pavements, especially those that had been heavily traveled and had both dark and light tones. Digitizing vehicles that had been previously digitized could be accomplished at a rate of approximately 5 sec per vehicle. New vehicles just entering the section could normally be digitized in 20 sec, including color, type, and length information. Entering the control points at the beginning of each frame could be done in approximately 30 sec. Thus the time to digitize one frame would depend primarily on the number of vehicles within each frame, which, in turn, is dependent on the section length, number of lanes, and traffic density. The time to digitize one frame of film would ordinarily be between 4 and 12 min, depending on these factors and assuming minimal need for error correction. Error correction is handled through an error correction menu on the CRT screen, which gives the operator the options of modifying an entry, restarting the frame at any point, or deleting an entry, among others. The number of errors and the time to correct an error are dependent on a number of factors, primarily image quality and operator proficiency.

Format of Data Files

Figure 6 shows the format of the data file created as a result of the digitizing process. The section shown is from an upgrade section with five 11-ft lanes, filmed in Los Angeles. The distance of each vehicle from the beginning of the section is shown in field 6, and the lateral position is shown in field 7. If vehicles are traveling in the center of the lane, the value in field 7 should be approximately equal to 11 times the lane number minus 5 or 6 ft (half a lane). Speeds are shown in field 5. Records with a zero speed represent vehicles that had only been digitized once at this point in the file and therefore could not have developed a speed history. As indicated, the file is organized by frame number with vehicles ordered sequentially from the downstream to the upstream end.

In each film there are approximately 3,600 frames to be digitized (1 hr at one frame per second). The number of vehicles within the section may range between 25 and 130, depending on the section and traffic characteristics, but will typically be in the 50 to 60 vehicle range. Thus, the average size of a data set may be nearly 200,000 records, but certain data sets may be half that size and others may be twice that size.

The data files are assembled in stages, beginning with the individual microcomputer files. The microcomputer files are limited in size so that final editing can be accomplished on the microcomputer itself. One microcomputer file might consist of between 5 and 30 frames of data. These files are concatenated in the process of being telecommunicated to a mainframe computer.

The validity of the data is dependent primarily on the validity of the calibration of each frame using the control points. Because eight pairs of control points were generally available at each site, it was possible to use several of the points other than the four used in the calibration itself to check the calibration. Although no ground-based data were collected at the same time as the film data, careful calibration of the film frames ensures that the data adequately represent traffic flow within the general tolerances indicated earlier. Because the vehicle positions are not perfectly precise, there will be some degree of jerkiness associated with vehicle movements, especially if

FIELD NUMBERS

1	2	3	4	5	6	7	8	9
53	125	1	14	59	234	50	3	5
53	121	1	17	46	199	7	7	1
53	126	1	15	48	146	6	6	1
53	127	1	18	46	106	19	6	2
53	128	4	19	0	28	8	1	1
54	82	1	12	48	1248	6	3	1
54	86	1	13	47	1240	25	6	3
54	92	1	18	57	1234	36	4	4
54	93	1	16	52	1165	48	2	5
54	94	1	14	52	1150	30	2	3
54	95	4	13	55	1111	49	9	5
54	91	1	17	46	1074	26	7	3
54	90	1	12	46	1055	4	1	1
54	97	1	16	55	1035	38	8	4
54	96	1	20	48	964	27	2	3
54	98	1	14	52	946	15	3	2
54	99	1	15	52	941	48	2	5
54	101	4	19	52	900	16	9	2
54	105	1	15	61	870	36	8	4
54	100	1	20	52	847	16	2	2
54	103	1	12	52	846	5	7	1
54	104	1	19	55	843	50	2	5
54	102	1	18	48	790	26	2	3
54	106	1	19	55	782	49	7	5
54	108	4	22	52	683	26	4	3
54	111	1	13	59	677	49	2	5
54	107	1	19	47	667	16	2	2
54	110	1	19	50	609	24	4	3
54	115	1	16	59	589	39	2	4
54	109	1	13	44	587	16	2	2
54	113	1	15	50	540	28	3	3
54	114	4	16	52	509	5	9	1
54	112	1	15	46	492	16	3	2
54	116	1	17	50	444	5	4	1
54	118	1	12	55	431	48	7	5
54	119	1	16	48	347	4	1	1
54	124	1	16	55	335	36	2	4
54	125	1	14	58	319	49	3	5
54	120	1	11	48	316	27	7	3

KEY TO FIELD NUMBERS

1 - FRAME NUMBER

2 - VEHICLE ID

3 - VEHICLE TYPE CODE

4 - VEHICLE LENGTH (FEET)

5 - SPEED (MPH)

6 - DISTANCE FROM BEGINNING OF SECTION TO FRONT OF VEHICLE

7 - DISTANCE FROM RIGHT EDGELINE TO MIDDLE FRONT OF VEHICLE

8 - VEHICLE COLOR CODE

9 - LANE NUMBER (RIGHT LANE=LANE 1)

FIGURE 6 Sample record format of digitized data file.

viewed in the unsmoothed 1-sec increments. The vehicle movements could be smoothed, if desired, to eliminate some of the jerkiness of movement. Studying movements over time increments longer than 1 sec will proportionately reduce this effect as well.

Availability of Data Sets

Each data set is available on 9-track magnetic tape from the FHWA Office of Research, HSR-20, Turner-Fairbank Highway Research Center, 6300 Georgetown Pike, McLean, Virginia 22101. Fourteen data sets were reduced in the study, including six weaving sites, three ramp merges, two grades, one curve, one lane drop, and one reduced-width section. In addition to the vehicle data file, each magnetic tape contains a file of geometric data for the section, coded in accordance with the conventions used in the INTRAS freeway simulation model. More information on the data sets is available elsewhere (5).

ACKNOWLEDGMENTS

The authors would like to acknowledge and thank Stephen Cohen, FHWA's contract manager, for his guidance and direction throughout the course of the project. The photography was performed by Photoscience, Inc., of Gathersburg, Maryland. Special thanks are also expressed to the departments of transportation of the states of Virginia, Maryland, and California and of the District of Columbia for their assistance in site selection and permitting the filming to be done.

REFERENCES

1. S.E. Charles et al. Exit Ramp Effects on Freeway System Operations and Control. ITE (UCLA); FHWA, U.S. Department of Transportation, 1971.
2. J.F. Raudseps. Automatic Extraction of Highway Traffic Data from Aerial Photographs. Transportation Systems Center, U.S. Department of Transportation, Cambridge, Mass., 1975.
3. J.B. Garner and L.J. Mountain. Traffic Data Collection--An Alternative Method. Traffic Engineering and Control, Vol. 19, No. 10, Oct. 1978, pp. 451-454.
4. R.L. Bleyl. Traffic Analysis of Time-Lapse Photographs Without Employing a Perspective Grid. Traffic Engineering, Vol. 42, No. 11, Aug. 1972, pp. 29-31.
5. S.A. Smith. Freeway Data Collection for Studying Vehicle Interactions. Final Report. FHWA, U.S. Department of Transportation, May 1985.

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