TRANSPORTATION RESEARCH RECORD

No. 1320

Highway Operations, Capacity, and Traffic Control

Freeway Operations, Highway Capacity, and Traffic Flow 1991

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL
WASHINGTON, D.C. 1991

Transportation Research Record 1320
Price: $45.00

Subscriber Category
IVA highway operations, capacity, and traffic control

TRB Publications Staff
Director of Publications: Nancy A. Ackerman
Senior Editor: Naomi C. Kassabian
Associate Editor: Alison G. Tobias
Assistant Editors: Luanne Crayton, Norman Solomon
Graphics Coordinator: Diane L. Snell
Production Coordinator: Karen S. Waugh
Office Manager: Phyllis D. Barber
Production Assistant: Betty L. Hawkins

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data
National Research Council. Transportation Research Board.

Freeway operations, highway capacity, and traffic flow, 1991.
p. cm.—(Transportation research record, ISSN 0361-1981 ; no. 1320)
"A peer-reviewed publication of the Transportation Research Board."
Papers contained in this record are from the 1991 Annual Meeting of the Transportation Research Board.
ISBN 0-309-05153-3
1. Express highways—Congresses. 2. Highway capacity—Congresses. 3. Traffic flow—Congresses. 4. Traffic engineering—Congresses. I. National Research Council (U.S.). Transportation Research Board. Meeting (1991 ; Washington, D.C.) II. Series : Transportation research record ; 1320.
TE7.H4 no. 1320
[HE336.E94]
388 s—dc20
[388.3'14] 91-43674
CIP

Sponsorship of Transportation Research Record 1320

GROUP 1—TRANSPORTATION SYSTEMS PLANNING AND ADMINISTRATION
Chairman: Ronald F. Kirby, Metropolitan Washington Council of Governments

Multimodal Freight Transportation Section
Chairman: Richard A. Staley, Richard A. Staley Consulting

Committee on Urban Goods Movement
Chairman: Arun Chatterjee, University of Tennessee
Secretary: Frank Spielberg, SG Associates, Inc.
Marsha Dale Anderson, Siro De Gasperis, Richard De Neufville, S. Earl Dove, Peggy L. Drake, Gordon P. Fisher, Marilyn S. Gay, Philip A. Habib, George C. Harper, Peter J. Hills, Ann C. Ketter, Pamela M. Lebeaux, Herbert S. Levinson, William Marconi, Masamitsu Mori, Kenneth W. Ogden, Marion R. Poole, F. Gerald Rawling, Noreen Roberts, Richard A. Staley, Richard A. Stehr, Carol A. Walters, C. Michael Walton

GROUP 2—DESIGN AND CONSTRUCTION OF TRANSPORTATION FACILITIES
Chairman: Raymond A. Forsyth, Sacramento, California

Evaluations, Systems, and Procedures Section
Chairman: Terry M. Mitchell, Federal Highway Administration, U.S. Department of Transportation

Committee on Applications of Emerging Technology
Chairman: Chris T. Hendrickson, Carnegie-Mellon University
Secretary: Thomas A. Fuca, New Jersey Department of Transportation
Robert J. Benke, Bertell C. Butler, Jr., Wiley D. Cunagin, Carl Haas, Richard C. Hanley, Haris N. Koutsopoulos, Sue McNeil, A. Essam Radwan, Malcom Ray, Justin J. Rennilson, Earl C. Shirley, Kumares C. Sinha, Yorgos J. Stephanedes, Nick A. Theophilopoulos, K. Thirumalai, W. Wiercienski, Walter A. Winter

GROUP 3—OPERATION, SAFETY, AND MAINTENANCE OF TRANSPORTATION FACILITIES
Chairman: H. Douglas Robertson, University of North Carolina–Charlotte

Facilities and Operations Section
Chairman: Lyle Saxton, Federal Highway Administration

Committee on Freeway Operations
Chairman: Ronald C. Sonntag, Wisconsin Department of Transportation
Secretary: Jeffrey A. Lindley, Federal Highway Administration
B. Beukers, E. Nels Burns, Donald G. Capelle, Glen C. Carlson, Robert F. Dale, Conrad L. Dudek, Walter M. Dunn, Jr., P. Kay Griffin, Herman E. Haenel, Kim C. Henry, Dennis C. Judycki, Jack L. Kay, Robert J. Kelly, Louis E. Lipp, Adolf D. May, Jr., William R. McCasland, Joseph M. McDermott, Gordon F. Paesani, Gary C. Price, Colin A. Rayman, David H. Roper, James F. Shea, Thomas C. Werner, Philip Zove

Committee on Highway Capacity and Quality of Service
Chairman: Adolf D. May, Jr., University of California at Berkeley
Secretary: Wayne K. Kittelson, Kittelson & Associates, Inc.
Rahmi Akcelik, Ulrich Brannolte, Kenneth G. Courage, Rafael E. De Arazoza, Douglas W. Harwood, Paul P. Jovanis, Michael Kyte, Herbert S. Levinson, John Morrall, Barbara K. Ostrom, Ronald C. Pfefer, James L. Powell, William R. Reilly, Carlton C. Robinson, Roger P. Roess, Nagui M. Rouphail, Ronald C. Sonntag, Stan Teply, Pierre-Yves Texier, Thomas Urbanik II, Mark R. Virkler, Robert H. Wortman, John D. Zegeer

Committee on Traffic Flow Theory and Characteristics
Chairman: Carroll J. Messer, Texas A&M University
Secretary: Edmund A. Hodgkins, EAH and Associates
R. F. Benekohal, Gang-Len Chang, Nathan H. Gartner, Fred L. Hall, Douglas W. Harwood, Richard L. Hollinger, Reinhart Kuhne, Michael Kyte, Edward Lieberman, Henry Lieu, Feng-Bor Lin, David Mahalel, Hani S. Mahmassani, Panos G. Michalopoulos, Abbas Mohaddes, A. Essam Radwan, Ajay K. Rathi, Nagui M. Rouphail, Mitsuru Saito, James C. Williams, Sam Yagar

GROUP 5—INTERGROUP RESOURCES AND ISSUES
Chairman: William J. Harris, Jr., Texas A&M University

Committee on High-Occupancy Vehicle Systems
Chairman: Donald G. Capelle, Parsons Brinckerhoff
Secretary: Dennis L. Christiansen, Texas A&M University
Richard D. Bauman, Paul N. Bay, John Bonsall, Frank A. Cechini, Jr., Donald J. Emerson, Charles Fuhs, David R. Gehr, Alan T. Gonseth, Leslie N. Jacobson, Theodore C. Knappen, James R. Lightbody, Timothy J. Lomax, Brian Pearson, Allan E. Pint, Lew W. Pratsch, Gary C. Price, William T. Roach, Morris J. Rothenberg, Sheldon G. Strickland, David L. Tollett, Gary K. Trietsch, Katherine F. Turnbull, Leonard J. Whitlock

Richard A. Cunard, G. P. Jayaprakash, and Elaine King, 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, 1990.

Transportation Research Record 1320

Contents

Highway Capacity

Traffic Flow and Simulation Models

Foreword

The papers contained in this Record are from the 1991 Annual Meeting of the Transportation Research Board and are related by their concern for freeway operations, traffic flow, and highway capacity.

Readers with a specific interest in freeway operations will find papers pertaining to centralized computer control of ramp metering systems, expert systems applied to freeway incident management and traffic surveillance data, data-screening techniques to improve the quality of data received from roadway detectors, the effects of restricting large vehicles from one or more lanes on freeways, determination of appropriate design-hour traffic volumes, and an integrated system of freeway corridor simulation models that extend the types of traffic management strategies that can be evaluated for a freeway corridor.

Those readers with an interest in traffic flow, capacity, and level-of-service measurements will find papers related to the capacity of a freeway and techniques for estimating it, inconsistency in the understanding of saturation flow and how it is measured, analysis of the application of the proposed revision to the procedure in the *Highway Capacity Manual* for multilane highways to freeway sections, capability of ramp metering to increase the capacity of freeway bottlenecks, methodology for evaluating the capacity and level of service at toll plazas, traffic flow and capacity characteristics of two-lane highways, capacity and delay at unsignalized intersections, validation and field testing of traffic simulation models, and traffic flow at freeway bottlenecks. In addition, papers on traffic operations of bicycle traffic and simulation of crowd behavior are also included.

The paper "Saturation Flow: Do We Speak the Same Language?" by Teply and Jones is of particular interest. This paper received the D. Grant Mickle award as the outstanding paper in the areas of operation, safety, and maintenance of transportation facilities at the 1992 Annual Meeting of the Transportation Research Board.

Benefits of Central Computer Control for Denver Ramp-Metering System

Louis E. Lipp, Lawrence J. Corcoran, and Gordon A. Hickman

An evaluation was made to determine the effects of providing central computer coordination on the Denver ramp-metering system. Information from the system was collected with the coordination algorithm operational and with it disabled. The data suggest that, when local demand-responsive ramp-metering control is unable to maintain freeway speeds at or near the posted limit of 55 mph, central coordination can be useful in reducing main-line delays. However, when speeds are near 55 mph with local ramp metering in operation, central coordination of the type used in the Denver system may be of little benefit.

In the 1970s the Denver metropolitan area experienced unprecedented growth in population and employment. For the motoring public, this growth resulted in increased congestion, delay, vehicle emissions, fuel usage, and frustration. Two studies, one conducted by the Denver Regional Council of Governments (*1*) and the other by the Colorado Department of Highways (*2*), recommended the implementation of ramp metering on selected portions of the freeway system to improve traffic flow and safety.

These studies led the Colorado Department of Highways to install ramp metering as a demonstration project on one of the most congested portions of the Denver freeway system. Metering began on March 3, 1981, during the morning peak period on five ramps of northbound I-25 from Hampden Avenue to Colorado Boulevard. Bus bypasses were built at two of these ramps to accommodate existing bus routes and to give preference to transit riders.

The initial system used stand-alone, demand-responsive controllers at each ramp metering location. Geometric improvements were made to bring acceleration lanes to standard lengths and to improve interchange designs. The demonstration project was an immediate technical and public relations success because it eliminated most stop-and-go traffic on this section of freeway. The project resulted in a 58 percent increase in freeway speed and a 37 percent reduction in vehicle-hours of travel (VHT) during the morning peak period (*3*).

Due to the success of the initial project, ramp metering was sanctioned as a transportation systems management strategy in the Denver area. The system was expanded in 1984 and in subsequent years to include a central computer with system coordination as well as additional meters on I-25, I-225, US-6, and I-270. In 1988, 24 ramps were under metered control with 22 operating during the morning peak period and 15 during the evening peak period (see Figure 1).

SYSTEM DESCRIPTION

The system is divided into six groups, with one to seven ramps per group (see Figure 1). Most ramps are striped for two lanes of vehicle storage. I-25, which includes Groups 1 and 2, is a six-lane facility in this part of the metropolitan area. Ramp control is operational for northbound traffic from County Line Road on the south to Colorado Boulevard on the north. Group 1 is separated from Group 2 by the I-225 interchange.

I-225 has four lanes north of the Parker Road interchange and six lanes between Parker Road and the junction of I-25. Group 3 includes all the southbound ramps, and Group 4 controls traffic on the northbound ramps. The northbound Yosemite Street onramp is the only nonmetered location on I-225. The two-lane I-225 and Parker Road southbound entrance ramp is unique within the system in that it allows two vehicles in each lane to enter the freeway during each metering cycle. This ramp also includes a high-occupancy vehicle (HOV) bypass lane with bus-only access to a park-and-ride lot.

Groups 5 and 6 are single-ramp systems on US-6 and I-270, respectively. These two groups were not included in the evaluation. Additional metering locations are planned for I-25 southbound and I-70 westbound between I-225 and I-270. Communications between the local ramp controllers and the central computer are handled by a combination of leased telephone lines and state-owned cable.

RAMP CONTROL STRATEGIES

The ramp controllers are programmed to operate only during weekdays for one or both of the peak periods. During metering periods, each ramp meter selects one of six metering rates on the basis of local traffic conditions (see Figure 2). Main-line primary and secondary detectors are used to determine the volume, occupancies, and speeds in each lane. The secondary detectors function as a backup for the primary detectors. The ramp presence and passage detectors inform the controller when a vehicle is waiting to be metered and when it has passed the metering signal. A 2-sec delay is programmed into the controller to prevent a vehicle from receiving an immediate green indication. Rate selections are made every 20 sec according to the information provided by the main-line detectors. An exponential smoothing function is included in the algorithm to prevent rapid switching between rates.

Queue detectors are installed near the entrance of the ramp to sense when vehicles are backing toward the cross street.

Colorado Department of Highways, 2000 South Holly Street, Denver, Colo. 80222.

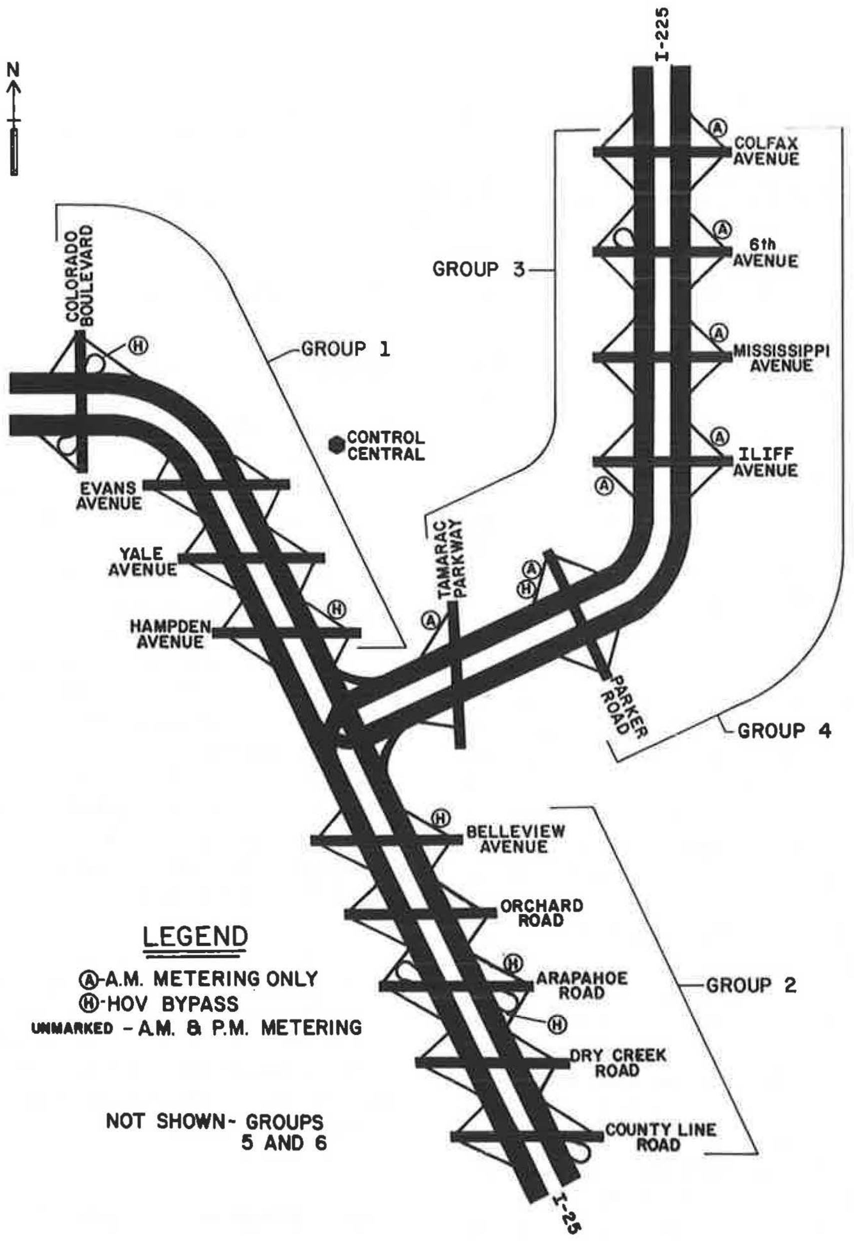


FIGURE 1 Denver area ramp-metering system.

When the occupancies of the queue detectors exceed an established threshold, the controller overrides the normally selected metering rate. The controller then initiates less restrictive rates until the backup is reduced to an acceptable level. The HOV detector, when actuated, extends the red signal time by a preset amount to help avoid conflicts in the merge area.

CENTRAL CONTROL STRATEGIES

Every 20 sec the central computer collects detector and metering rate data from each ramp. If a ramp is in the most restricted rate (freeway congested) or in queue override (ramp congested), the ramp is defined as critical. A system coordination plan is then put into effect. This plan calculates the travel time between ramps and, after that time, forces a more restrictive metering rate on the next upstream ramp.

If the ramp remains in a critical condition during the next sampling period, the rates of the next two upstream ramps are forced one rate more restrictive. The system continues to add ramps to the coordination plan during each sampling cycle until all ramps in the group are under central control. If more than one ramp becomes critical, multiple plans are put into effect.

When the last ramp in a group is put under the system coordination plan (see Figure 1), the central computer begins implementing coordination in the first ramp of the next upstream group that feeds the critical ramp. The system coordination plan continues until the ramps return to a noncritical condition. The plan returns the local controllers to normal operation one rate at a time at the 20-sec sampling intervals.

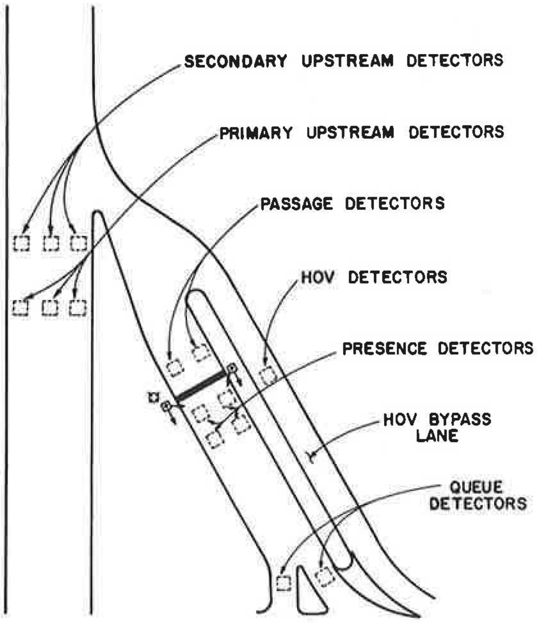


FIGURE 2 Typical metered freeway ramp.

The central computer also monitors the traffic just before and after the peak periods to determine if metering should occur earlier or later than the core metering times. If conditions are favorable for ramp control, the central computer allows the local controllers to begin metering up to 20 min before and to remain on 30 min after the mandatory metering times.

DATA COLLECTION

Data on the effects of the system coordination algorithm were collected during September, October, and November of 1988. This time of the year was selected because of the stability of the daily traffic patterns. For similar reasons, only information from the morning peak period on Tuesdays, Wednesdays, and Thursdays was used to evaluate the benefits of coordination.

A microcomputer was used to collect and summarize data from the ramp-metering central computer. The microcomputer was connected to the central computer using a readily available communications software package. Volume and speed data for each ramp in the four groups were obtained for the hours of 7:00, 8:00, and 9:00 a.m. This information was transferred to a spreadsheet program that contained distances between the ramps. VHT and vehicle-miles of travel (VMT) were calculated for each group by adding VMT and VHT between each ramp. Data from each day were examined for indications of inclement weather, accidents, or equipment malfunctions, and any suspect information was rejected.

Six days' worth of acceptable data was collected between September 28 and October 13 with the system in normal operation (coordination on). On October 18 the system coordination algorithm was disabled, and 8 days' worth of information was gathered (coordination off). Data for an additional 2 days of normal operation were obtained on November 8 and 10.

EVALUATION

A summary of the data is presented in Tables 1 and 2. For all groups there was a slight reduction in VMT for the coordinated condition compared with the noncoordinated condition. VHT reductions were 13.1 and 3.9 percent for Groups 1 and 4, respectively, with only slight reductions indicated for Groups 2 and 3.

A statistical test was performed to determine if the differences between the two conditions were significant given the sample size and the magnitude of the differences. The criterion used was the *t* distribution. The required *t* values for several levels of significance are as follows:

Minimum t Value	*Level of Significance*
1.34	0.2
1.76	0.1
2.14	0.05
2.62	0.02
2.98	0.01

Calculated *t* values for the ramp metering data are as follows:

Group	*VMT*	*VHT*
1	0.37	3.62
2	0.82	0.18
3	1.79	0.29
4	0.91	1.50

From this information, it appears that three values are significant: (a) the difference in Group 1 VHT is significant at the 0.01 level; (b) the difference in Group 3 VMT is significant at the 0.1 level; and (c) the difference in Group 4 VHT is significant at the 0.2 level. The rest of the data indicates no statistically significant change between the two conditions.

Because only minor benefits were found during coordination (with the exception of Group 1 VHT), further examination of the data was performed. Traffic volumes and speeds at the ramps closest to the bottleneck locations and average group speeds were reviewed (see Tables 3–5). For Groups 2, 3, and 4, the ramp and average group speeds were near or at the posted speed limit of 55 mph. For Group 4, the speed at Colfax and I-225 increased by 2 mph under coordination.

TABLE 1 VEHICLE-MILES OF TRAVEL

GROUP	WITHOUT COORD.	WITH COORD.	% DIFF.
1	21498	21360	-0.6
2	32857	32503	-1.1
3	33906	33451	-1.3
4	31862	31514	-1.1

TABLE 2 VEHICLE-HOURS OF TRAVEL

GROUP	WITHOUT COORD.	WITH COORD.	% DIFF.
1	603	524	-13.1
2	570	567	-0.5
3	604	601	-0.5
4	564	542	-3.9

TABLE 3 TRAFFIC VOLUMES AT RAMPS CLOSEST TO BOTTLENECK

GROUP	LOCATION	VOLUMES WITHOUT COORD.	VOLUMES WITH COORD.	% DIFF
1	I-25 AND COLORADO	7267	7199	-0.9
2	I-25 AND BELLEVIEW	7010	6881	-1.8
3	I-225 AND TAMARAC	5246	5185	-1.2
4	I-225 AND COLFAX	4166	4108	-1.4

TABLE 4 TRAFFIC SPEEDS AT RAMPS CLOSEST TO BOTTLENECK

GROUP	LOCATION	SPEEDS WITHOUT COORD.	SPEEDS WITH COORD.	% DIFF
1	I-25 AND COLORADO	31	42	+35.5
2	I-25 AND BELLEVIEW	57	56	-1.8
3	I-225 AND TAMARAC	55	55	0.0
4	I-225 AND COLFAX	54	56	+3.7

TABLE 5 AVERAGE GROUP SPEEDS

GROUP	SPEEDS WITHOUT COORD.	SPEEDS WITH COORD.	% DIFF.
1	37	42	+13.5
2	58	57	-1.7
3	56	55	-1.8
4	57	58	-1.8

The other ramp and group speeds either did not change or decreased by 1 mph during coordination.

CONCLUSIONS

The data suggest that, when local demand-responsive ramp metering control is unable to maintain freeway speeds near the posted limit of 55 mph, central coordination can be useful in reducing main-line delays. However, when speeds are near 55 mph with local ramp metering in operation, central coordination of the type used in the Denver system may be of little benefit.

REFERENCES

1. *Managing the Transportation System in the Denver Region*. Denver Regional Council of Governments, Denver, Colo., March 1980.
2. *I-25 TSM Study, Ramp Metering Feasibility Analysis*. Colorado Department of Highways, Denver, Aug. 1979.
3. L. J. Corcoran and G. A. Hickman. *Freeway Ramp Metering Effects in Denver*. July 1990.

Publication of this paper sponsored by Committee on Freeway Operations.

Real-Time Expert System Approach to Freeway Incident Management

STEPHEN G. RITCHIE AND NEIL A. PROSSER

Fundamental to the operation of most intelligent vehicle-highway system (IVHS) projects are advanced systems for surveillance, control, and management of integrated freeway and arterial networks. A major concern in the development of so-called "smart roads" is the provision of decision support for traffic management center personnel, particularly for addressing nonrecurring congestion in large or complex networks. Decision support for control room staff is necessary to detect, verify, and develop effective response strategies for traffic incidents. These incidents are events that disrupt the orderly flow of traffic and cause nonrecurring congestion and motorist delay. Nonrecurring congestion can be caused by accidents, spilled loads, stalled or broken-down vehicles, maintenance and construction activities, signal and detector malfunctions, and special or unusual events. An attempt was made to implement a novel, artificial intelligence–based approach to the problem of providing operator decision support in integrated freeway and arterial traffic management systems, as part of a more general IVHS. The development of the Freeway Real-Time Expert System Demonstration (FRED), a component prototype real-time expert system for managing nonrecurring congestion on urban freeways in southern California, is discussed. The application of FRED to a section of the Riverside Freeway (SR-91) in Orange County is presented as a case study and illustrates the current capabilities of the system.

As a means to improve road-based mobility and safety, with decreased economic and environmental impacts from traffic, the concept of an intelligent vehicle-highway system (IVHS) is evoking substantial interest in Europe, Japan, and the United States. Relying on advances in electronics, communications, and computing, IVHS technologies would create so-called "smart cars" and "smart roads" to achieve significant areawide traffic operation improvements. A recent report (*1*) classified IVHS technologies in four categories:

1. Advanced traffic management systems,
2. Advanced driver information systems,
3. Freight and fleet control systems, and
4. Automated vehicle control systems.

The following paragraphs focus on advanced traffic management systems because advanced systems for surveillance, control, and management of integrated freeway and arterial networks are fundamental to the operation of most IVHS projects. In addition, implementation of these concepts is beginning. In Los Angeles, for example, the Santa Monica Freeway Smart Corridor Demonstration Project is under way, and other smart corridor projects are likely to follow.

However, a major concern in the development of such smart roads is the provision of decision support for traffic management center personnel, particularly for addressing nonrecurring congestion in large or complex networks. Decision support for control room staff is necessary to detect, verify, and develop effective response strategies for traffic incidents. These incidents are events that disrupt the orderly flow of traffic and cause nonrecurring congestion and motorist delay. Nonrecurring congestion can be caused by accidents, spilled loads, stalled or broken-down vehicles, maintenance and construction activities, signal and detector malfunctions, and special or unusual events. The ultimate objective of this research was to implement a novel, artificial intelligence (AI)–based approach to the problem of providing operator decision support in integrated freeway and arterial traffic management systems, as part of a more general IVHS. Although it is envisioned that for some time vehicles will operate mostly under driver control, in the future automated lateral and longitudinal control of vehicles may be possible. New vehicles, facilities, and vehicle and system control strategies may also be used. Nevertheless, advanced decision support capabilities similar to the concepts being developed in this research are likely to be important to the operation of future IVHS projects.

In previous research (*2*), a conceptual AI–based design was developed. The approach involved a hierarchically defined set of decision support modules within a distributed blackboard framework, emphasizing the use of real-time knowledge-based expert systems (KBESs). In practice these KBESs could be associated with multiple computers, traffic control centers, transportation agencies, and traffic subnetworks, even in one corridor.

The development of the Freeway Real-Time Expert System Demonstration (FRED), a component prototype real-time expert system for managing nonrecurring congestion on urban freeways in southern California, is discussed. The application of FRED to a section of the Riverside Freeway (SR-91) in Orange County is presented as a case study and illustrates the current capabilities of the system.

SYSTEM DEVELOPMENT

Nature of the Domain

In describing the operation of the freeway traffic operations center (TOC) in Los Angeles, the California Department of Transportation (Caltrans) states that the basic goal is to "know what's happening on the freeway system and to get infor-

Institute of Transportation Studies and Department of Civil Engineering, University of California, Irvine, Calif. 92717.

mation out to the motoring public." In conjunction with the California Highway Patrol, the TOC currently disseminates information by commercial radio stations and changeable message signs (CMSs) adjacent to the freeway. The TOC can dispatch an operational unit called the "major incident traffic management team" (MITMT) to incident locations, while providing continuous monitoring and coordination functions. The semi-automatic traffic management system (SATMS) includes approximately 700 directional freeway miles, 934 instrumented locations or stations (typically involving a full set of loops across the pavement, plus those at on- and offramps), and about 5,000 detectors providing 30-sec occupancy and volume data.

The city of Los Angeles also maintains a TOC for the signalized surface street system, called the "automated traffic surveillance and control system" (ATSAC). Currently, the system monitors approximately 400 system detectors. The Santa Monica Freeway Smart Corridor Demonstration Project area encompasses over 400 signalized intersections and will likely add 1,000 detectors to the ATSAC system. Inclusion of additional areas may add hundreds of intersections and thousands of detectors to ATSAC in the future.

As the breadth and scope of these systems continue to expand, particularly in conjunction with smart corridor and other IVHS concepts and requirements, the amount of incoming TOC data and the complexity of both the networks and incident management and response functions will make it increasingly difficult, if not impossible, for human operators to function effectively without automated assistance.

Real-time KBESs address situations such as those in which human operators suffer from cognitive overload in time-sensitive environments. In a smart corridor context, a KBES could filter low-level (but voluminous) detector data and present the operator with fewer high-level analyses and recommendations concerning incident detection, verification, and response. Use of the system would reduce the operator involvement needed to focus on true operational problems, permit rapid development of optimal and consistent response plans, and facilitate coordination among all relevant agencies, thereby reducing delays associated with nonrecurring congestion.

System Functions

The objective of the envisioned system is to provide decision support to TOC staff in the traffic surveillance and control functions required in a smart corridor context, potentially as part of a more general IVHS. Five integrated modules are proposed (*2*):

1. Incident detection,
2. Incident verification,
3. Identification and evaluation of alternative responses,
4. Implementation of selected responses, and
5. Monitoring recovery.

Emphasis is placed on the initial development of FRED, which, as a component of an overall decision support system, is limited to a freeway TOC to assist in managing nonrecurring congestion on urban freeways. FRED is currently being developed for a 6-mi section of the Riverside Freeway (SR-91) in Orange County, California. To assist in the development of FRED, detector data containing several major incidents have been supplied by Caltrans for this section of freeway, which is located between two other major freeways (I-5 and SR-57).

The overall functions of FRED are presented in Figure 1. In this figure Smart Central, or SCC, refers to a proposed real-time KBES that would attempt to optimize corridor or areawide traffic conditions and would coordinate response actions among all relevant agencies. Associated with Smart Central would be a major relational data-base system to facilitate the networked linking of all agencies and their control systems. Further details are provided by Ritchie (*2*).

Hardware and Software

FRED is being developed using G2 real-time expert system development software (*3*). G2 has been designed specifically for real-time applications and provides a powerful software development environment. In FRED external functions to G2 are being written in C (G2 is LISP-based). G2 also permits a highly graphical and easy-to-use window-based operator interface to be constructed. The hardware platform being used is a color Sun SPARCstation 1 workstation and a RISC-based Unix machine with 16 Mb of random access memory.

Knowledge Acquisition

To date, the knowledge embodied in FRED has been acquired from the authors, from a variety of professional papers and

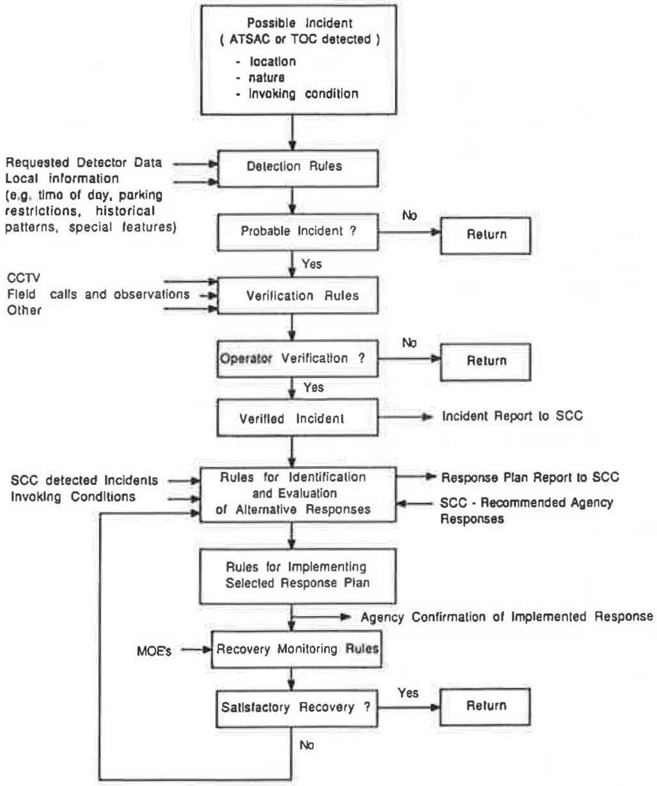

Possible Incident
(ATSAC or TOC detected)
- location
- nature
- invoking condition
Requested Detector Data
Local information
(e.g. time of day, parking restrictions, historical patterns, special features)
Detection Rules
Probable Incident ?
No
Return
Yes
CCTV
Field calls and observations
Other
Verification Rules
Operator Verification ?
No
Return
Yes
Verified Incident
Incident Report to SCC
SCC detected Incidents
Invoking Conditions
Rules for Identification and Evaluation of Alternative Responses
Response Plan Report to SCC
SCC - Recommended Agency Responses
Rules for Implementing Selected Response Plan
Agency Confirmation of Implemented Response
MOE's
Recovery Monitoring Rules
Satisfactory Recovery ?
Yes
Return
No

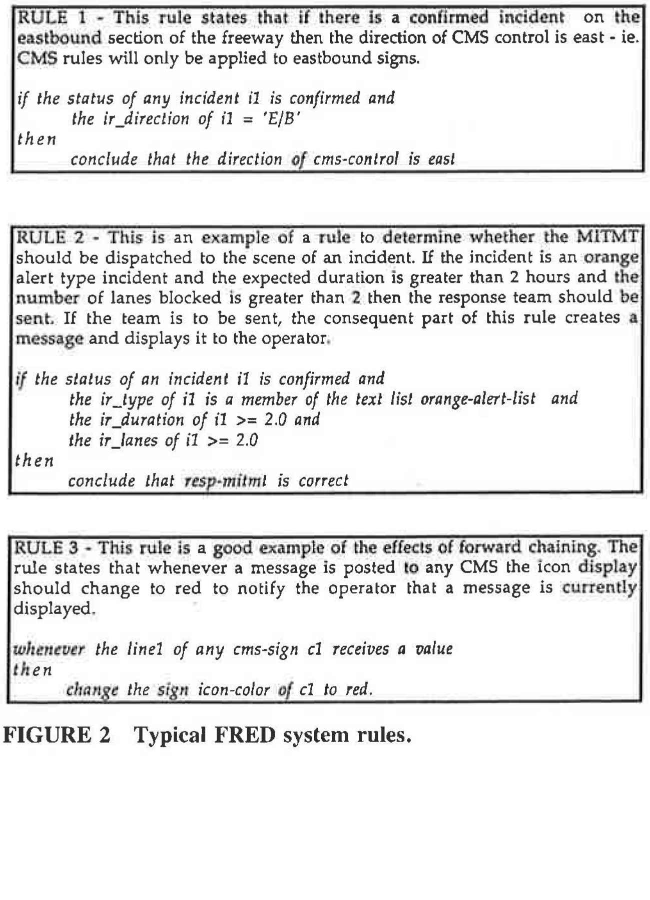
FIGURE 1 Freeway TOC overview.

reports (*4,5*), from Caltrans traffic operations specialists in Los Angeles and Orange County (Districts 7 and 12, respectively), and from many individuals and colleagues involved in the Santa Monica Freeway Smart Corridor Demonstration Project. Further research is clearly required to develop fundamental traffic operations and control system knowledge to be captured in FRED for identifying optimal control and motorist information response strategies in an integrated freeway and arterial traffic system. Such research could be pursued parallel to the development of tools such as FRED.

Knowledge Representation and Inference

Any expert system, real-time or otherwise, requires knowledge and data. Knowledge is usually embodied as a set of rules that act on data and facts to accomplish the objectives of the system. Real-time expert systems differ from conventional static expert systems in that they must respond to data that are continually changing. The nature of traffic is such that its behavior can change rapidly, particularly if an incident occurs. A real-time expert system for traffic monitoring and control must respond reliably and quickly to changes in incoming data.

In FRED knowledge is represented as a series of production rules, examples of which are given in Figure 2. An English-style format is used to make writing and interpreting rules less taxing. Each rule has an antecedent and a consequent. If the conditions embodied in the antecedent are satisfied, then the actions within the consequent are executed. Actions encompass a whole range of tasks that can be performed by the system, including graphic displays, posting of messages to the operator and external systems, setting of attributes, and so on. To behave as a real-time system, FRED examines all active rules each second. Thus, changing data can be responded to every second, which is sufficiently fast for traffic control conditions, particularly when loop data are often only available every 30 sec. The structure of the rules depends greatly on the structure of the data, of which a brief description follows.



RULE 1 - This rule states that if there is a confirmed incident on the eastbound section of the freeway then the direction of CMS control is east - ie. CMS rules will only be applied to eastbound signs.

```
if the status of any incident i1 is confirmed and
       the ir_direction of i1 = 'E/B'
then
       conclude that the direction of cms-control is east
```

RULE 2 - This is an example of a rule to determine whether the MITMT should be dispatched to the scene of an incident. If the incident is an orange alert type incident and the expected duration is greater than 2 hours and the number of lanes blocked is greater than 2 then the response team should be sent. If the team is to be sent, the consequent part of this rule creates a message and displays it to the operator.

```
if the status of an incident i1 is confirmed and
       the ir_type of i1 is a member of the text list orange-alert-list  and
       the ir_duration of i1 >= 2.0 and
       the ir_lanes of i1 >= 2.0
then
       conclude that resp-mitmt is correct
```

RULE 3 - This rule is a good example of the effects of forward chaining. The rule states that whenever a message is posted to any CMS the icon display should change to red to notify the operator that a message is currently displayed.

```
whenever the line1 of any cms-sign c1 receives a value
then
       change the sign icon-color of c1 to red.
```

FIGURE 2 Typical FRED system rules.

Data in FRED are organized as sets of objects, and in this sense the system can be described as object-oriented. Each object contains a set of attributes in which data are stored. For example, in FRED each incident is represented as an object with such attributes as location, type, and expected duration. All objects belong to an object class, and all classes are arranged in a hierarchy that incorporates downward inheritance. A characteristic of real-time expert systems in general, and of FRED in particular, is that of transient (or dynamic) objects. When the need arises to store data, an object of the appropriate predefined class can be created and data stored in its attributes. Rules can then operate on this object. In FRED, this idea is appropriate for the various stages of an incident. Once an incident is detected, an incident object is created and its attributes are assigned values. When an incident has terminated, its corresponding object is deleted. By this method, any number of incident objects corresponding to multiple freeway incidents can exist at any one time, and rules can operate on all of them together. For example, Rules 1 and 2 in Figure 2 have in their antecedents the clause "if the status of an incident i1 is confirmed. . ." This clause translates as "if the status attribute of any incident object has the value confirmed." Here the incident is the object class, of which there may be any number of instances, each representing a different freeway incident.

Having established how knowledge and data are represented, the inference process, that is, the procedure for examining and executing (firing) rules, can be considered. Most expert systems incorporate either forward or backward chaining, and usually both. Most real-time expert systems would rely on forward chaining to respond to events, and FRED is no different. Forward chaining is an inference method that attempts to match the antecedents of rules against available facts (or events) to establish new facts that will eventually lead to a goal or conclusion. The major event in the FRED system is the confirmation, by the operator, of an incident that, by forward chaining, fires a series of rules designed to formulate responses.

An important aspect of real-time expert system development is to avoid unnecessary processing so as to maintain a fast system. FRED uses the capabilities of the G2 system to invoke a set of rules only when they are needed. The system completely ignores rules that are not invoked, so the invoking of rules is a means of controlling the amount of knowledge that needs to be applied to any particular stage of the overall incident management process.

PROTOTYPE FEATURES

System Structure

Figures 3 and 4 show, respectively, the external and internal layout of the FRED system. External system components include incident detection algorithms and a communications

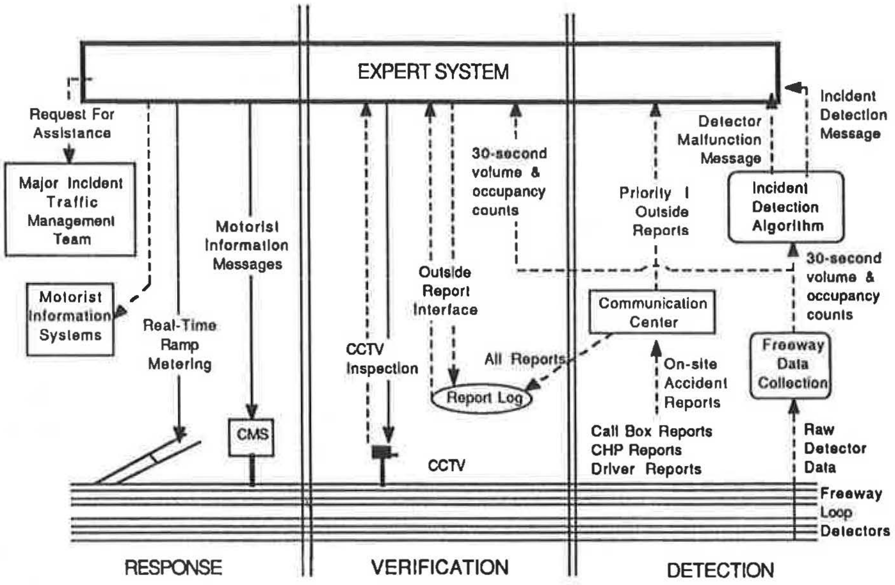


FIGURE 3 FRED external system overview.

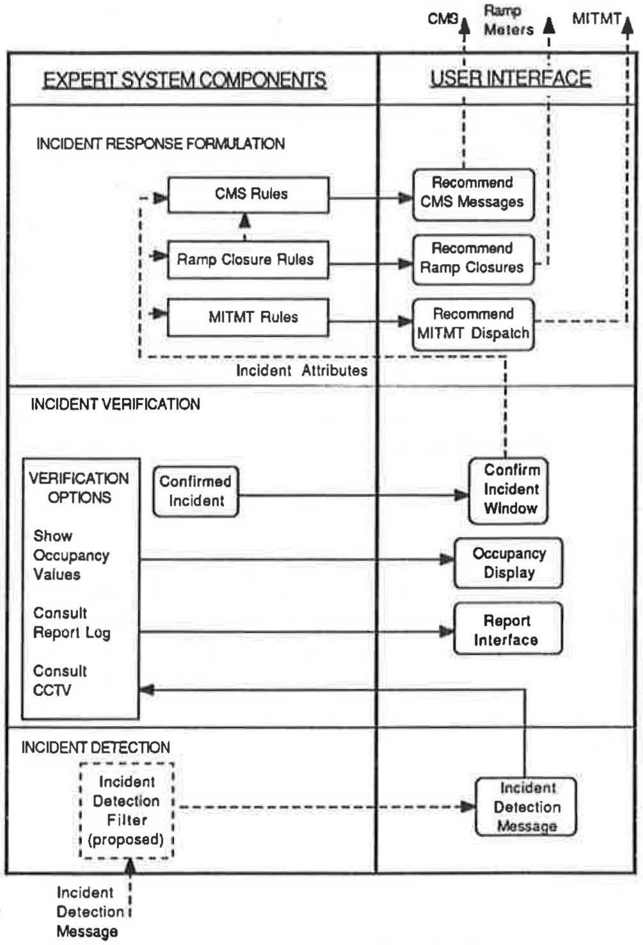


FIGURE 4 FRED internal system overview.

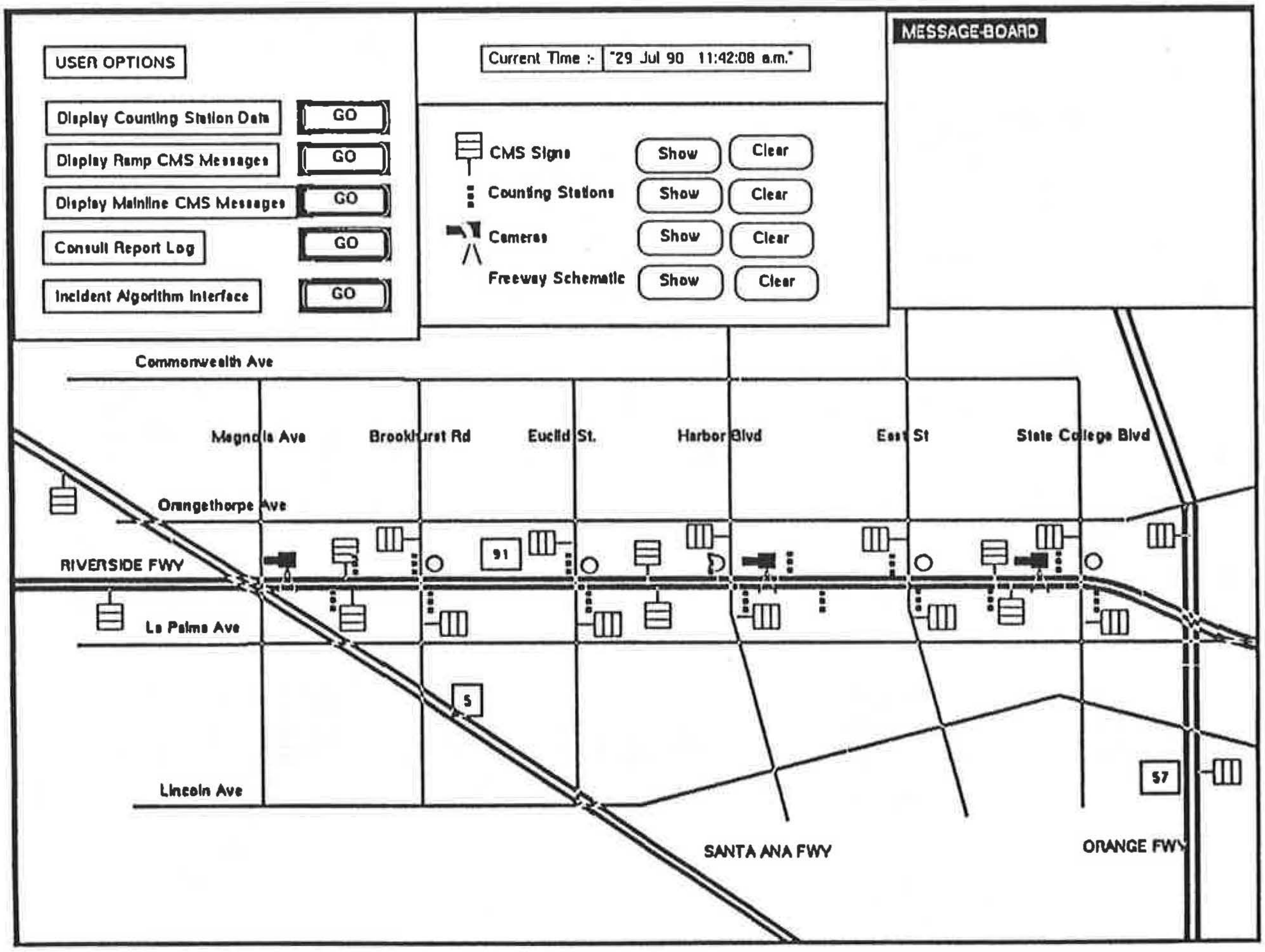
center to aid in the detection of incidents, as well as various incident response mechanisms. The role of both the internal and external components is discussed in later sections.

A vital requirement of any system designed for operator support is the user interface. FRED uses the sophisticated built-in screen management facilities of G2 to provide the operator with a series of separate windows containing either graphic data displays or messages. The operator interacts with the expert system by way of mouse-driven action buttons or keyboard-driven type-in boxes.

Figure 5 shows the screen that is presented to the operator when no incidents have been detected (actual screens employ different colors that do not appear in the figures). The central part of the display is the location map, depicting the section of SR-91 under study, along with two adjacent freeways (I-5 and SR-57) and major arterial streets. At the top center of the screen is a panel of display action buttons, allowing the operator to selectively view the location of counting stations, CMSs, and closed circuit television (CCTV) cameras. In Figure 5 all these symbols are displayed. Figure 5 shows the location of hypothetical CCTV cameras placed above this section of SR-91 to provide the operator with visual images of traffic conditions, as well as CMSs.

Action buttons perform a prescribed action when the operator clicks the mouse on the button icon. For example, Figure 6 shows a message that appears when an incident is detected. Next to the message is a CONFIRM action button. When the operator clicks the mouse on this button, the system records that the operator received the message.

Another visual aid to the operator is shown in Figure 6. A schematic of the freeway layout appears at the bottom of the screen, showing lane configurations, ramp designs, and the precise locations of counting stations and CMSs. The icons representing CMSs and ramp meters can be changed to denote



Current Time :- "29 Jul 90 11:42:08 a.m."

FIGURE 5 System display board.

different operating conditions. For example, the two lights on a ramp meter icon are changed to red if the corresponding ramp is closed in response to an incident.

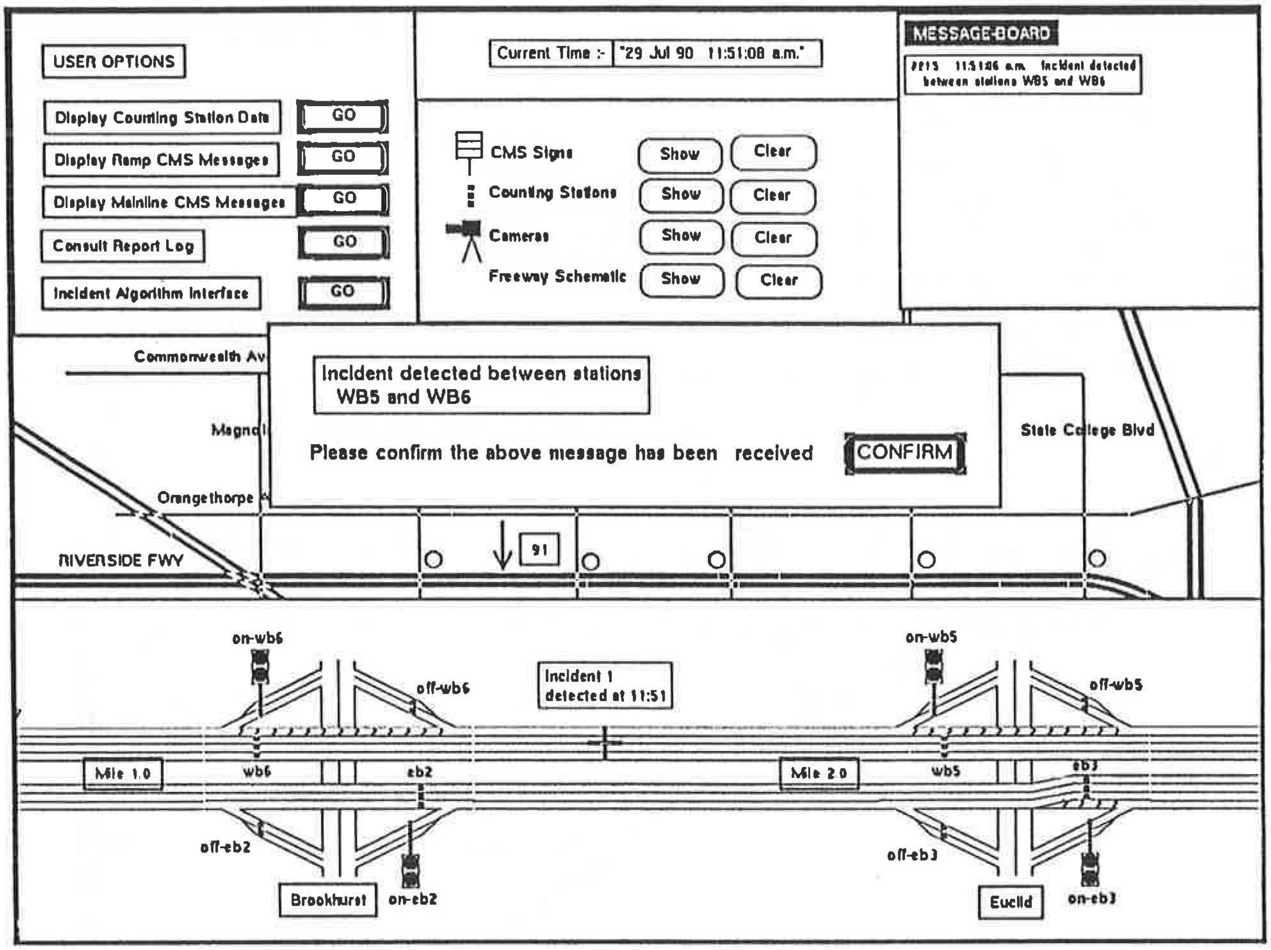
Incident Detection

The first step in incident management is the detection phase. Attempts to manage incidents are ineffective if incidents are not detected quickly and reliably. In the FRED environment, incidents are detected in two ways: (a) from applying an incident detection algorithm to loop detector data and (b) from outside reports.

Considerable research has been undertaken on developing effective computer algorithms for the detection of freeway incidents using main-line loop detector data. In the FRED system, a version of the California algorithm reported by Payne and Knobel (*6*) is implemented. The algorithm reads 30-sec occupancy counts from a series of counting stations on the freeway main line. The stations are processed as a series of sections, with each section having an upstream and a downstream station. A number of parameter values are derived from the occupancy counts at the upstream and downstream stations. If these values lie outside some predetermined range, an incident is said to be detected. A separate option in FRED allows the operator to interactively select the algorithm to be used and its sensitivity.

Once an incident has been detected, a signal is sent to FRED that a possible incident exists between the upstream and downstream stations that triggered the detection. A message window is placed in the center of the screen to notify the TOC operator, who is then required to acknowledge the message (see Figure 6). On the central map display, the color of the freeway section in which the suspected incident is located changes from green to flashing red, and an arrow appears next to it. The suspected incident is only known to lie between the two counting stations that triggered the detection, and, until an on-site report is received, its location remains approximate.

The second method of detecting an incident is by way of outside reports, which are usually on-site reports of an incident from freeway motorist call-boxes, police officers, cellular car phones, aerial observers reporting on traffic conditions, or Caltrans maintenance personnel. It is assumed that all such reports are first received by a communications center manned by operators who enter the reports into a data base in a predetermined format. Information such as incident location, nature of incident, number of lanes blocked, and presence of



Current Time :- "29 Jul 90 11:51:08 a.m."

MESSAGE-BOARD

#215 11:51:06 a.m. Incident detected between stations WB5 and WB6

Incident detected between stations WB5 and WB6

Please confirm the above message has been received

Incident 1 detected at 11:51

FIGURE 6 Incident detection message.

injuries and fatalities is supplied. If the incident is a major one, such as an overturned truck, then more detailed information may be required, such as a description of the load, whether flammable or toxic materials are involved, and whether specialized assistance is required.

In FRED the communications center is simulated by an external program that accepts incident reports and checks to see if they match a particular protocol before entering them onto a report log. The report program examines the type of incident to determine whether extra information is required, and, if so, the communications center operator is prompted for the information. Each report is allocated a priority ranging from 1 to 5, depending on the nature of the incident and the source of information. For example, a report originating from a police officer at the scene of an incident would receive the highest priority of 1. Priorities become important when considering whether or not the report should be passed on to the freeway TOC. Only Priority 1 incidents are communicated to FRED; all other incident reports are written to a report log that can be consulted by the TOC operators when needed (see the following section).

For a number of reasons, it is important that FRED receives reports only after preliminary processing. First, the operators at the freeway TOC should not be overwhelmed with unreliable reports or multiple reports of the same incident. Second, reports should be received by FRED in a recognizable format to allow them to be easily understood by the operator and to enable the development of rules that operate on the reports. For example, the type of incident (such as overturned truck or two-car accident) must appear on a list of incident types recognized by FRED so that rules such as Rule 2 in Figure 2 can function properly.

FRED deals with outside incident reports in the same way as incident detection reports—a message requiring acknowledgment is placed in the center of the screen.

An important aspect of the incident detection phase is the prevalence of false incident reports, that is, reports of incidents that either do not exist or are not severe enough to warrant response. A role of any system that aims to reduce the cognitive load of its operators is to reduce this false alarm rate to manageable levels. The preliminary processing of outside reports is one means of reducing the level of false alarms from this source. However, reducing false alarms triggered by loop detector data is more difficult. More incident detection algorithms involve a trade-off between detection rate (percentage of true incidents detected) and false alarm rate (percentage of false alarms over a certain period of time). Increasing the detection rate by altering threshold parameters necessarily increases the false alarm rate. Currently, FRED leaves this trade-off problem to the external detection algorithm, but it is hoped that heuristics can be incorporated into FRED to act as a further filter of incident detection reports.

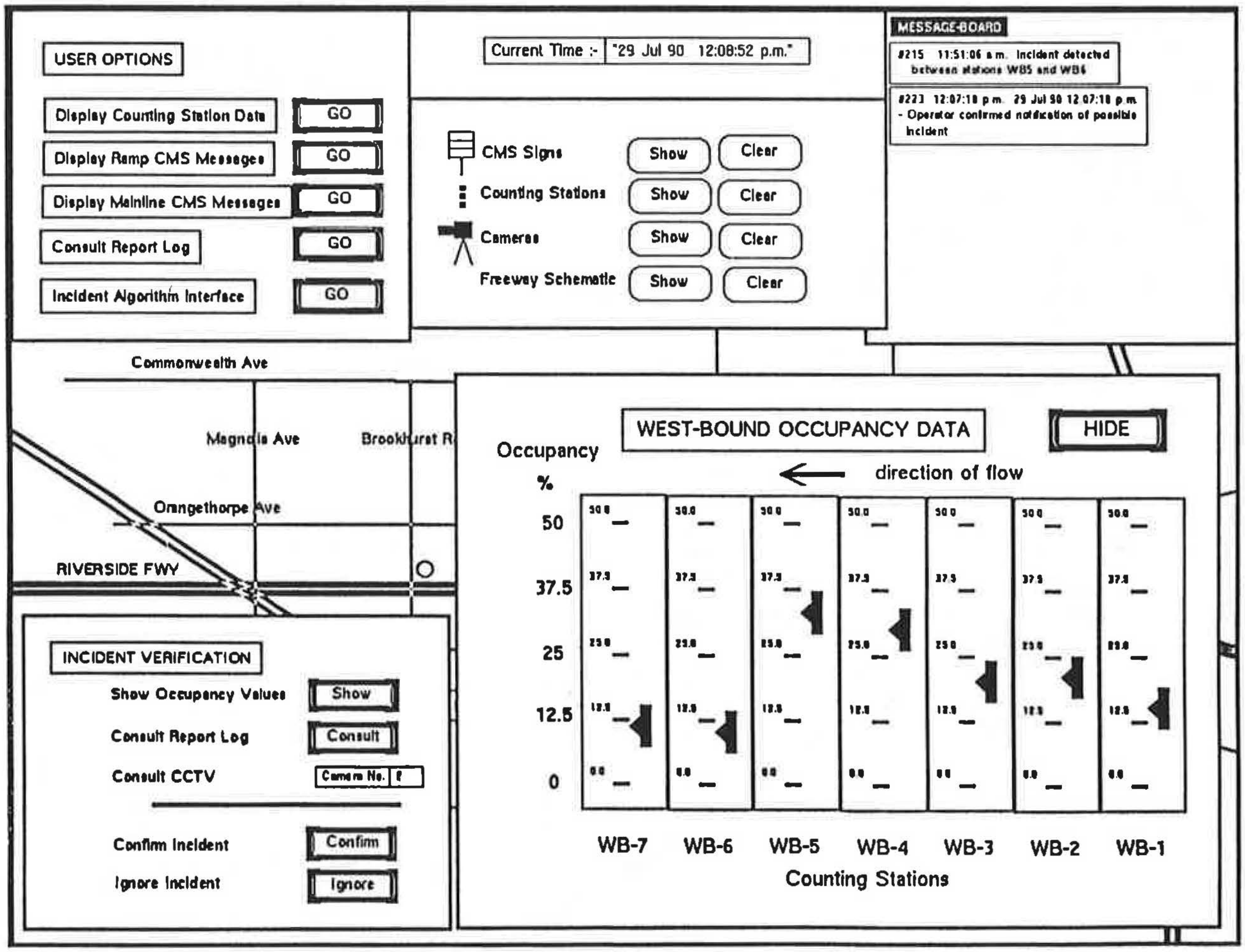
Incident Verification

Once an incident has been detected or reported to the freeway TOC operator, it is the operator's task to verify the incident before any incident responses are formulated. In the current FRED system, there are three verification methods available: (a) CCTV, (b) inspection of loop data, and (c) consultation of the report log. Figure 7 shows the incident verification window (in the bottom left corner of the screen) that is presented to the operator as a summary of the verification options.

The incident verification window in Figure 7 displays the identification number of the camera closest to the suspected incident location. Although it is technically feasible for FRED to directly control the positioning and zooming of the camera, this action is not simulated, partly because of the uncertainty surrounding the incident's location. If CCTV is available and visibility is sufficient, the use of a CCTV camera is the primary means of verifying an incident. However, under poor visibility conditions or in sections where CCTV is unavailable, the operator must rely on other methods.

The graphic display of current values of traffic parameters, such as occupancy and volume, may help operators recognize an incident. Figure 7 shows a display of occupancy counts for the seven counting stations located on the westbound section of the freeway (see Figure 5 for the location of counting stations). The sudden discontinuity of occupancy between Stations WB5 and WB6 suggests a major disruption of flow between these two stations. Stations downstream of WB6 show low occupancy values, suggesting more freely flowing traffic, whereas the upstream values are higher, indicating the onset of congestion. Experienced operators may be able to recognize certain patterns in traffic conditions that indicate the presence of nonrecurring congestion. Such expertise could be encoded into FRED as a separate knowledge base and used in filtering false alarms. Graphic displays of traffic conditions can also be useful during incident monitoring, particularly when deciding if an incident has terminated.

As mentioned in the section on outside reports, a log of all reports is maintained external to the FRED system, with only high-priority reports being sent directly to TOC operators. In the incident verification phase, the operator is able to interrogate the report log for reports relevant to the time and location of the incident being verified. A standard data-base query procedure is followed, with the operator providing values for certain fields (e.g., all reports with a time stamp be-



Current Time :- "29 Jul 90 12:08:52 p.m."

MESSAGE-BOARD

#215 11:51:06 a.m. Incident detected between stations WB5 and WB6

#223 12:07:18 p.m. 29 Jul 90 12:07:18 p.m. - Operator confirmed notification of possible incident

FIGURE 7 Incident verification.

tween 8:30 and 8:50 a.m. and a location between Magnolia and Euclid streets). Details of relevant reports, if any, are displayed on the screen. The operator may decide that an incident detection report combined with a low-priority outside report is sufficient to confirm the presence of an incident.

If an operator verifies an incident, an incident confirmation window is displayed, as shown in Figure 8. Examples of type-in boxes are shown in this figure. The operator is required to enter, among other incident attributes, the type of incident (determined perhaps from a CCTV camera). The string entered by the operator becomes an attribute of the current incident object. Thus, objects can receive values of attributes from external sources, internal inferences, or the operator. If the operator has visual contact with the scene of the incident via CCTV, details relating to the incident can be entered into FRED simply by typing the values into the appropriate boxes contained within the incident confirmation window. Additionally, the operator is allowed to move a marker denoting the approximate position of the incident to a more precise location. The position of the marker is recorded and used in the response stages.

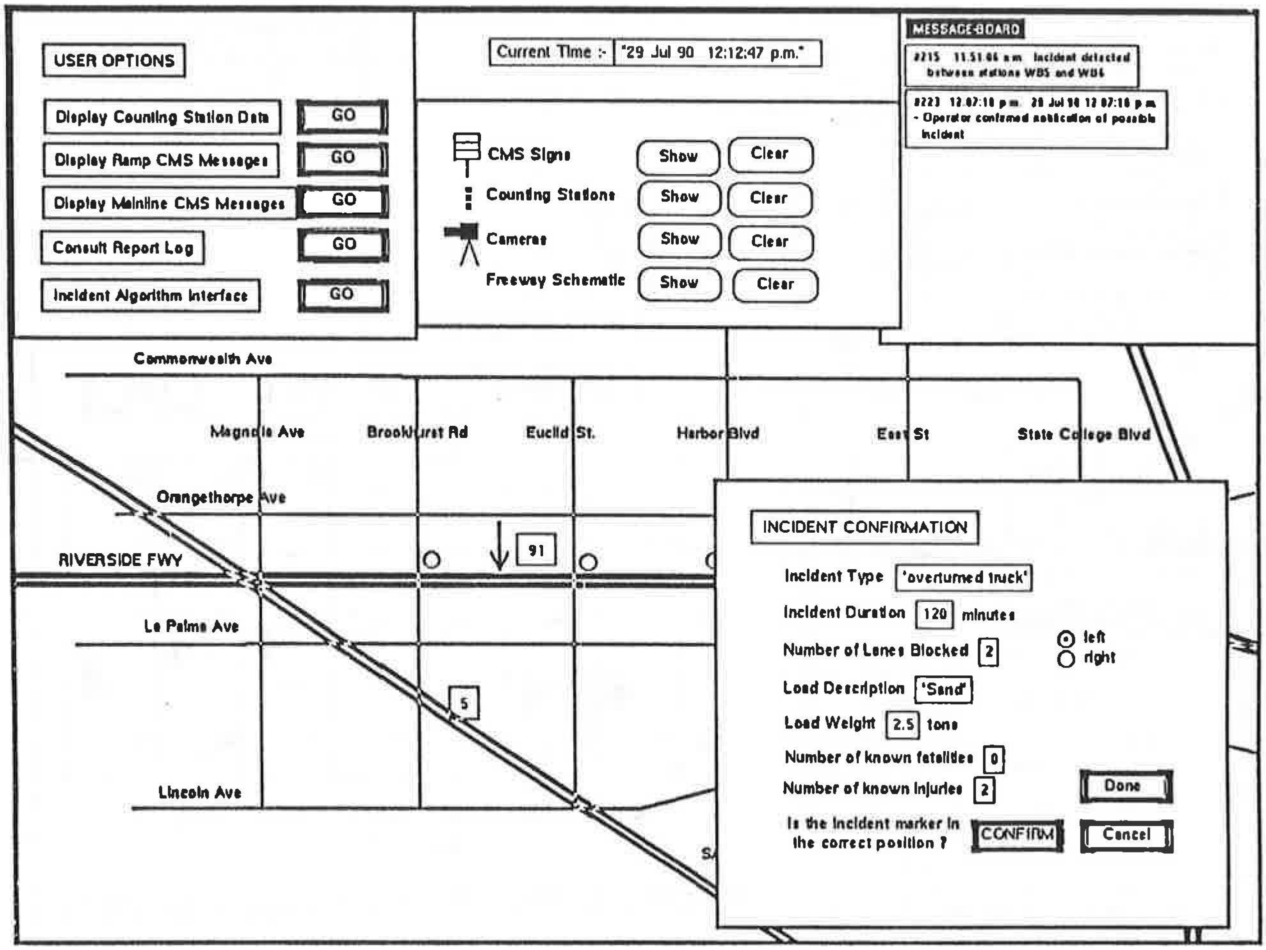
Incident Response

The formulation of incident response strategies is the major area in which a sophisticated expert system approach is warranted. In the FRED system, there are currently three main response elements: (a) the MITMT, (b) real-time ramp metering, and (c) CMSs.

MITMT Response

As explained previously, the Los Angeles district of Caltrans instituted the MITMT. The team's primary purpose is "to furnish, as rapidly as possible, equipment and manpower to aid in management of traffic at or near major traffic incidents on freeways." Currently, the team's equipment consists of 12 sedans and 11 mobile CMSs. Personnel include 5 primary and 18 standby members, and the team is available 24 hr a day.

The role of FRED in invoking the MITMT is to determine whether or not an incident is major and, if so, to provide necessary information about the nature of the incident to the response team. Both these tasks are achieved by a set of rules, of which Rule 2 in Figure 2 is an example. In effect, the rule states that, if a so-called "orange-alert" incident is confirmed, if the expected duration is 2 hr or more, and if the number of lanes blocked is two or more, the MITMT should be sent. This rule is drawn from the existing guidelines for operation of the MITMT. Incident types are arranged in lists according to their severity. Orange-alert incidents include spilled loads or jack-knifed trucks. Red-alert incidents include overturned trucks, bomb threats, and hazardous material spills. If the



Current Time :- '29 Jul 90 12:12:47 p.m.'

MESSAGE-BOARD

#215 11:51:46 a.m. Incident detected between stations WB5 and WB6

#223 12:07:18 p.m. 28 Jul 90 12:07:18 p.m. - Operator confirmed notification of possible incident

INCIDENT CONFIRMATION

Incident Type 'overturned truck'

Incident Duration 120 minutes

Number of Lanes Blocked 2 ● left ○ right

Load Description 'Sand'

Load Weight 2.5 tons

Number of known fatalities 0

Number of known injuries 2

Is the incident marker in the correct position ? CONFIRM

FIGURE 8 Incident confirmation.

operator confirms the response, an incident report is sent to the MITMT dispatch office providing such details as location, type, expected duration, lanes blocked, number of injuries and fatalities, and description of any spilled load. Further work on FRED should provide recommended diversion strategies to be implemented by the on-site team.

Ramp Metering and Closure

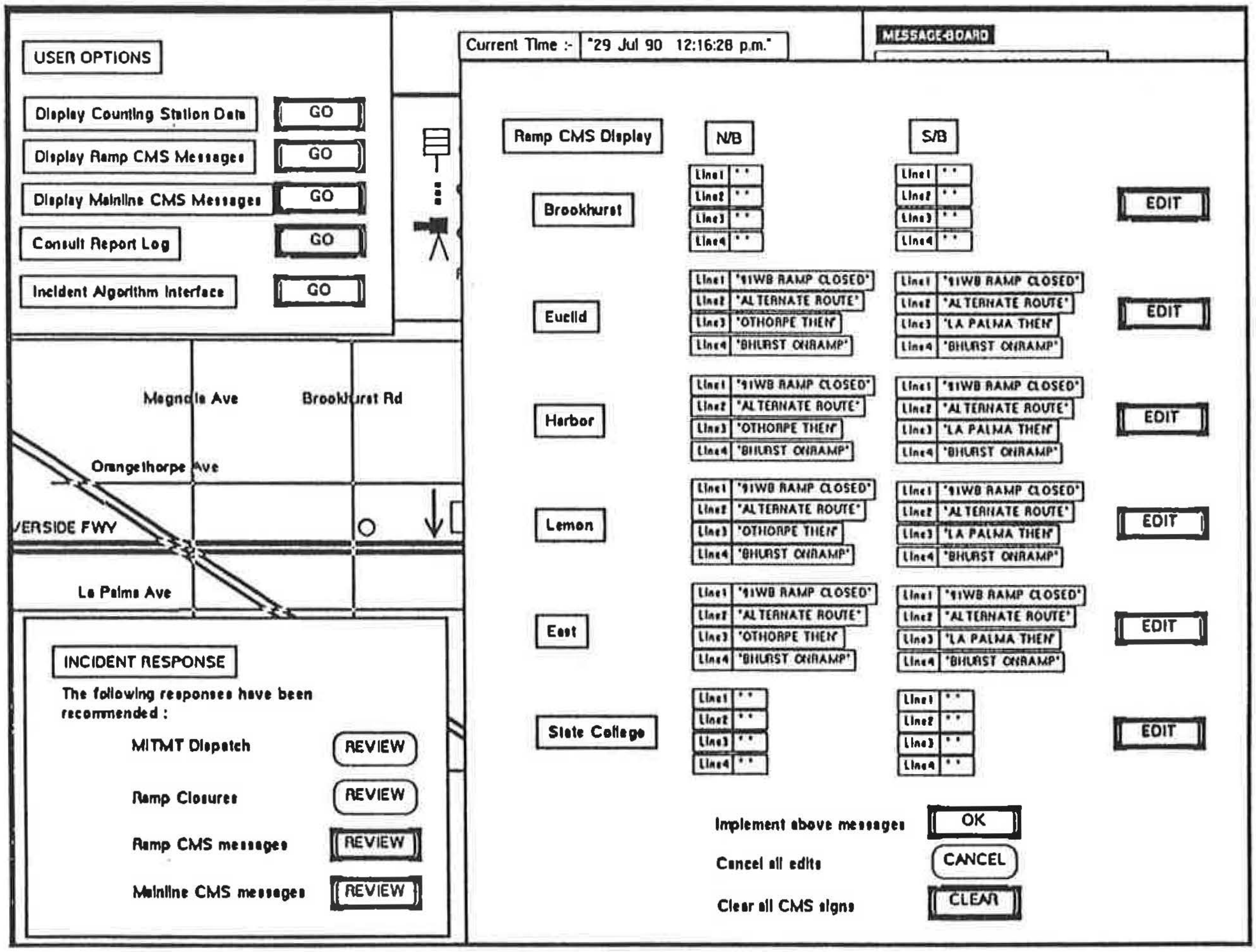
A simplified version of a real-time ramp metering algorithm developed for use in Seattle, Washington (*7*), is implemented as a module external to the FRED system. This algorithm computes optimal metering rates every 30 sec in an attempt to maintain a high level of service for traffic downstream of each ramp. It is assumed that all entrance ramps on the case study section are metered and that real-time control of these meters is possible.

The FRED system allows the ramp metering algorithm to operate independently and only intercedes when the capacity of a section of the freeway has been drastically reduced by an incident. To reduce demand at the incident site, ramps upstream may be recommended for closure. This task is performed in stages. In this initial version of FRED, an estimate is made of the capacity at the incident site using such information as number of lanes blocked. Flow conditions upstream of the incident are then examined to determine the expected demand on the freeway at the incident site. The severity of the incident is determined to be the extent to which expected demand exceeds capacity at the incident site. If the severity of the incident is above a specified threshold, all ramps within a certain distance upstream of the incident are recommended for closure. The threshold and upstream distance values can be modified by TOC operators.

Arterial Street CMS Information

If ramp closure is to be an acceptable response mechanism, advance information must be provided to motorists intending to use the entrance ramps. Such information should indicate which ramp is closed and, more importantly, provide an alternative route. The formulation of messages to be posted on CMSs on arterial streets near entrance ramps is the responsibility of another section of FRED. Each major arterial street with an interchange to SR-91 within the case study section was assumed to have a CMS positioned on the north and south approaches. Thus, each entrance ramp was provided with two CMSs that could provide information about closure.

The major task in formulating the messages is to determine an alternative route around the incident site using arterial streets. Two street names were provided: (a) the entrance ramp immediately downstream of the incident and (b) the arterial street parallel to the freeway leading to the entrance ramp. Figure 9 shows the set of messages recommended for the arterial street CMSs for an incident on SR-91 westbound



Current Time :- "29 Jul 90 12:16:28 p.m."

Ramp CMS Display

		N/B	S/B
Brookhurst	Line1	" "	" "
	Line2	" "	" "
	Line3	" "	" "
	Line4	" "	" "
Euclid	Line1	'91WB RAMP CLOSED'	'91WB RAMP CLOSED'
	Line2	'ALTERNATE ROUTE'	'ALTERNATE ROUTE'
	Line3	'OTHORPE THEN'	'LA PALMA THEN'
	Line4	'BHURST ONRAMP'	'BHURST ONRAMP'
Harbor	Line1	'91WB RAMP CLOSED'	'91WB RAMP CLOSED'
	Line2	'ALTERNATE ROUTE'	'ALTERNATE ROUTE'
	Line3	'OTHORPE THEN'	'LA PALMA THEN'
	Line4	'BHURST ONRAMP'	'BHURST ONRAMP'
Lemon	Line1	'91WB RAMP CLOSED'	'91WB RAMP CLOSED'
	Line2	'ALTERNATE ROUTE'	'ALTERNATE ROUTE'
	Line3	'OTHORPE THEN'	'LA PALMA THEN'
	Line4	'BHURST ONRAMP'	'BHURST ONRAMP'
East	Line1	'91WB RAMP CLOSED'	'91WB RAMP CLOSED'
	Line2	'ALTERNATE ROUTE'	'ALTERNATE ROUTE'
	Line3	'OTHORPE THEN'	'LA PALMA THEN'
	Line4	'BHURST ONRAMP'	'BHURST ONRAMP'
State College	Line1	" "	" "
	Line2	" "	" "
	Line3	" "	" "
	Line4	" "	" "

Implement above messages OK

Cancel all edits CANCEL

Clear all CMS signs CLEAR

INCIDENT RESPONSE

The following responses have been recommended :

MITMT Dispatch REVIEW

Ramp Closures REVIEW

Ramp CMS messages REVIEW

Mainline CMS messages REVIEW

FIGURE 9 Incident response—CMS messages.

between Brookhurst and Euclid. Four ramps have been closed—at the Euclid, Harbor, Lemon, and East interchanges. The entrance ramp immediately downstream of the incident site is at Brookhurst. The parallel arterial street for northbound traffic is Orangethorpe Avenue, and the one for southbound traffic is La Palma. The operator is allowed to edit any of the messages before implementation or to cancel them all.

This formulation of an alternative route via arterial streets will, in practice, require some assessment of the capacity of the alternative streets and entrance ramps. An appropriate approach is to coordinate the diversion of traffic from the freeway, as determined by FRED, with the arterial street system to improve the flow of traffic through the entire corridor.

Main-Line CMS Information

If nonrecurrent congestion is expected due to a detected incident, information should also be provided to motorists on the freeway approaching the incident site. In the case study, motorists are notified via CMSs located on SR-91 as well as on connecting freeways I-5 and SR-57. The location of the incident and the number of lanes blocked are given. This information is derived directly from the incident parameters provided in the verification stage, with some message composition processing.

For each incident response, the operator is presented with the recommended action, if any, and asked for confirmation. Figure 9 shows the incident response window displayed in the bottom left corner of the screen after the formulation of all responses. The operator is allowed to review each response before implementation, as illustrated in the discussion of arterial street CMS messages.

CONCLUSIONS

The response of traffic operations specialists in southern California to initial demonstrations of FRED has been most favorable. However, much research remains to be done to incorporate the proposed additional decision support functions in FRED. This work is ongoing.

ACKNOWLEDGMENTS

This research was supported by a U.S. Department of Transportation and University of California Transportation Center award for "Real-Time Decision Support for Freeway Surveillance and Control" and by a grant from the National Science Foundation.

REFERENCES

1. *Report to Congress on Intelligent Vehicle-Highway Systems.* Office of Economics and Office of the Assistant Secretary for Policy and International Affairs, U.S. Department of Transportation, 1990.
2. S. G. Ritchie. A Knowledge-Based Decision Support Architecture for Advanced Traffic Management. *Transportation Research*, Vol. 24A, No. 1, 1990.
3. *G2 User's Manual.* Gensym, Inc., Cambridge, Mass., 1988.
4. *Traffic Control Systems Handbook.* Publication LP-123. Institute of Transportation Engineers, Washington, D.C., 1985.
5. *Traffic Operations Center Operator's Manual.* District 7, California Department of Transportation, Los Angeles, 1989.
6. H. J. Payne and H. C. Knobel. *Development and Testing of Incident Detection Algorithms, Volume 3: User Guidelines.* Report FHWA-RD-76-21. Technology Service Corporation; FHWA, U.S. Department of Transportation, 1976.
7. L. N. Jacobson, K. C. Henry, and O. Mehyar. A Real-Time Metering Algorithm for Centralized Control. Presented at 68th Annual Meeting of the Transportation Research Board, Washington, D.C., 1989.

Publication of this paper sponsored by Committee on Freeway Operations.

Improved Data Screening Techniques for Freeway Traffic Management Systems

Don Cleghorn, Fred L. Hall, and David Garbuio

Proper functioning of computer-based freeway traffic control algorithms depends on good data being received from roadway detectors. At present, only elementary screening procedures for these data are in use on most freeway traffic management systems. There has recently been some work to improve such procedures. That work has been extended, and several new screening approaches are proposed. The first approach develops a theoretical upper bound for part of the flow-occupancy data and is useful for single-detector systems. The other approaches have been developed for paired-loop systems and involve comparisons of loop data as well as the development of boundaries defining acceptable combinations of speed, flow, and occupancy data.

The ability to monitor the conditions on a freeway is the fundamental feature of a freeway traffic management system (FTMS), on which all other functions depend. Most FTMSs measure flow and occupancy, and some also measure average speed. Using these data, the central computer of an FTMS can perform many complex functions, such as incident detection and travel time estimation. Unfortunately, the most sophisticated and elaborate computer algorithms inevitably fall short of their goals if the information they receive is of poor quality. Some new procedures for the identification and removal of bad FTMS data in an on-line computer environment are described. The resulting screened FTMS data is a more accurate representation of freeway conditions and thus should improve system performance.

The first section provides background, describing current data screening approaches. The second section describes the FTMSs from which data have been taken for analysis. Because the objective is to demonstrate the effectiveness of some proposed procedures, not to test hypotheses, the development of the procedures is explained as the results are presented, rather than discussing development separately from applications. Hence, the third section contains both the discussion of the procedures and a demonstration of the type of results that can be obtained from each of the proposed screening methods. The final section contains conclusions.

BACKGROUND

The effects of bad data on freeway monitoring and control algorithms have been documented in the past. Dudek et al. (*1*) noted that hardware problems resulted in an increase in the number of false alarms observed in the Gulf freeway warning system. Courage et al. (*2*) found that design and operational parameters can have significant effects on accuracy, even without hardware failures. For example, the detector scan rate can have a large effect, especially in a system in which the central computer monitors the detectors directly. Courage et al. determined that the minimum expected error in speed observations (for the smallest scanning interval examined) from such systems is 19 percent. Chen and May (*3*) also found that slight misadjustments in the detecting equipment (which are bound to occur from time to time) led to significant differences in output values.

Although the screening of data from vehicle detectors is not a new concept, little research has been done on the topic, and the testing methods that exist are still very simple (if used at all). In a 1984 survey of 32 freeway management projects in North America, only 19 had data checking procedures (*4*). Most of these 19 systems checked for stuck-on or stuck-off conditions and chattering in the detector amplifiers. Only 11 systems checked the plausibility of values received from the detector stations. Jacobson et al. (*5*) divided data screening tests into two categories: microscopic and macroscopic. Microscopic tests are defined as those tests done by the controller microprocessor itself as pulses are received from the detectors. Macroscopic tests are performed by the central computer after the data have been aggregated over time.

The macroscopic plausibility testing done most often (*4*) is comparison of a single variable against upper or lower threshold values. For example, the California Department of Transportation (Caltrans) algorithm checked averages of values over 5-min periods; the Chicago system checked data averaged over 1 min (*6*). The thresholds are usually constants input by system personnel on the basis of an examination of the reasonable range of the variables. For example, occupancy is only defined from 0 to 100 percent and both extremes can be observed in traffic, so its reasonable limits are clear. Some system operators choose to reduce this range, judging that the extremes are rare enough to be unimportant (*5*). Values of speed or flow cannot be less than 0 regardless of the units used. Their maxima are not easy to determine, although historical data may be used to produce estimates.

These threshold tests implicitly assume that the acceptable range for a variable is independent of the values of the other variables. Because combinations of variables are not tested, single-variable tests using such thresholds cannot catch completely impossible combinations, such as an occupancy of 3 percent together with a flow of 2,000 vehicles per hour (vph). This observation is clearly not reasonable, but the only published test for such combinations is that reported recently in Washington State (*5*). The Washington State approach in-

D. Cleghorn, UMA Engineering, 5080 Commerce Boulevard, Mississauga, Ontario, Canada L4W 4P2. F. L. Hall, Department of Civil Engineering, and D. Garbuio, Department of Geography, McMaster University, Hamilton, Ontario, Canada L8S 4L7.

volves setting limits for acceptable values of volumes for any given occupancy on the basis of plausible ratios between the two within specific occupancy ranges (see Figure 1). Although ingenious, their method provides wider ranges than are necessary, due to the stepped nature of the boundaries. Testing on the basis of combinations of variables is clearly on the right track, but a tighter limit should be possible.

On the whole, then, current FTMSs use data from detector stations with minimal examination of credibility. Because the purpose of an FTMS is to monitor a freeway system and thus allow its control, the quality of the data collected is of critical importance to the proper operation of the FTMS. The designs of several new data screening algorithms, which incorporate tests to detect previously overlooked flaws in FTMS data, are presented in the following sections.

SOURCES OF DATA

Data from two FTMSs are used in the analysis, both on the Queen Elizabeth Way (QEW) in Ontario, Canada. The first system is on the Burlington Skyway and its approaches. The Skyway is the portion of the QEW that crosses the entrance to Hamilton Harbour; therefore, it is high enough for ocean-going and Great Lakes shipping to pass under safely. This height is accomplished by a 3 percent grade for about 1.25 km. The second FTMS used is in Mississauga on the main western commuter approach to Toronto, a section that experiences recurrent congestion nearly every weekday.

These systems have been selected because they are representative of many others, not because they are unique. Indeed, the Skyway system has only recently (1986) been put into operation and so contains state-of-the-art components, both hardware and software, as of that time. The Mississauga system was initially installed in 1978 (*7*). It has had some equipment replacements during the past 2 years but still has some older equipment. Hence, the stations in this system offer a range of data error experience. The data collection systems and screening procedures for these systems are described in the following paragraphs.

Most of the data collection stations have two inductance loop detectors embedded in the roadway, 6 m apart, for each lane at a station. The detector loops are monitored by a Type 170 controller, which scans the loops at 60 Hz. The controller summarizes the information gathered from the loops over a 30-sec period for transmission to the central computer. If a value cannot be calculated for a loop during that 30-sec interval, the valuc is markcd as missing by the controller. Missing variables can arise for a variety of reasons. For example, speed cannot be calculated when there has been no traffic during the interval. Before sending data to the central computer, the station controller may flag one or more detector loops as bad because of a stuck-on or stuck-off condition. In the Skyway FTMS, a loop is considered stuck if it is in the same state (either occupied or not occupied) continuously for 2 hr.

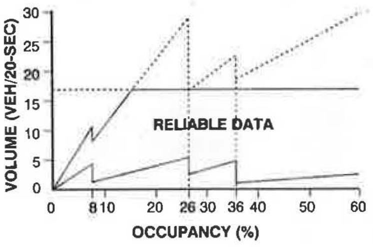


FIGURE 1 Boundaries on reliable data set in the Washington State study (*5*).

The data received by the central computer consist of 30-sec average values of speed, as well as values for volume and occupancy for each lane and station in the system. Basic data screening is done by the central software before the incident detection algorithms are run, and 5-min averages are tested to discover failed detectors (*8* and informal communication, Freeway Traffic Management Systems Section, Ontario Ministry of Transportation, April 1988). The simple range tests are performed every interval and can be summarized as follows:

1. Volume must not be negative and must not exceed a specified threshold, currently 60 vehicles per interval (roughly double the maximum observed to date).
2. Occupancy must not be negative and must not exceed 100 percent.
3. Speed must not be negative and must not exceed 150 km/hr. (Actual speeds in excess of 130 km/hr are occasionally observed, so this limit is only about a 10 percent margin).

If any one of these tests on 30-sec data fails, a flag is set to mark the invalid datum. Any incident detection algorithms that require data that have failed the test are disabled for that lane or station for the current 30-sec interval.

PROPOSED COMPARATIVE TECHNIQUES FOR SCREENING

Systems with paired loops in each lane at each station, such as those on the QEW, have more data-screening options available. Not only can they test the validity of flow-occupancy data pairs, but they can also compare the received speed-flow-occupancy points with a calibrated three-dimensional speed-flow-occupancy region and compare upstream and downstream loop values. The following discussion of improved screening techniques is divided into two main sections, the first covering an improved method for single-loop systems and the second identifying the additional screening that can be done with paired-loop systems.

Single-Loop Systems: Setting Boundaries on Flow-Occupancy Points

The Washington State approach—trying to identify the region of feasible flow-occupancy pairs—represents a greatly improved method for screening FTMS data from single-loop systems. It makes use of knowledge about the relationship between these variables that is ignored by single-variable range tests. However, those boundaries can be improved, given their stepped nature (see Figure 1). To improve those boundaries, emphasis has been placed on the upper bound because experience has shown that nearly all pairs of flow-occupancy values are possible below the normal flow-occupancy function, at the time congestion begins.

Two methods for obtaining a tighter upper bound suggest themselves. The first relies on historical data to set the limits. The second, and the one pursued here, is to try to develop a theoretical approach to an upper bound, thereby eliminating both the Washington State indirect method and the need for historical data. An additional benefit of such an approach would be that the boundary could be manipulated to suit a specific system using values of the variables acquired there.

The relationship between volume and occupancy depends on speed, vehicle length, and detector length. Following Athol (*9*), the relationship is

$$\text{Occupancy} = [\Sigma\ (x_i + d)/u_i]/T$$

where

x_i = length of Vehicle i,
u_i = speed of Vehicle i,
d = detector length, and
T = duration of the time interval.

The summation is taken over the vehicles passing in T.

This equation does not reduce to any simple form except under the assumptions of uniform vehicle length and speed. Fortunately, those assumptions are reasonable for the bulk of uncongested flow, as found by Hall and Persaud (*10*), so it may be possible to establish an upper limit for those conditions from the equation that results from those assumptions:

$$\text{Volume} = (\text{occupancy} * \text{speed} * T) \div (\text{vehicle length} + \text{detector length})$$

Although the calculations were done with the equation in this form, it is easiest to explain with occupancy calculated from volume. The upper bound of interest can be visualized as the minimum value of occupancy allowed for a given volume. If volume is constant, occupancy will decrease as speeds increase and as vehicle lengths decrease. Hence, the upper bound should be calculated using the highest feasible speeds and the shortest reasonable vehicle lengths. The maximum speed used in this analysis was 130 km/hr, but this limit can easily be changed for different systems. A minimum vehicle length of 4 m was used, which is the shortest that might be expected in the median lane. Motorcycles are shorter, but the average vehicle length during 30 sec is of interest. For the shoulder lane, a higher value might be reasonable. The detection zone length used in the calculations was 2 m.

The result of the calculations, up to an occupancy of 20 percent, is shown in Figure 2 along with the observed data for Station 17 on the Mississauga system. The slope of the line is the ratio of speeds to combined vehicle and detector lengths. If lower speeds or longer vehicles are used for the calculation, the slope decreases. Theoretically, volume is zero with an occupancy of zero. However, an observation of up to three volume counts at zero occupancy is not unusual because of truncation of the occupancy values in these systems. Thus, a vertical shift of the boundary upward by three volume counts, as shown in the figure, is needed to ensure that the boundary reflects the way this particular system operates on the data. Similar shifts no doubt are needed for other systems. Although the shifted upper bound would accept all of the data in Figure 2 (as it should), such a boundary would reject the type of data that the Washington State screening test found. Similar data were gathered during testing from one station on the Mississauga system. As shown in Figure 3, this boundary can reject such data.

To continue the upper bound beyond the critical occupancy (i.e., that at which highest volumes occur) poses some difficulty. Observed maximum speeds decline abruptly but follow no obvious pattern. In addition, where previously the highest speed was wanted, to find the lowest possible occupancy, now the task is to find the highest possible occupancy for a given volume, which requires the lowest possible speed. To choose a minimum speed for higher occupancies is difficult. However, the existence of a tight upper bound for uncongested flow may be sufficient. If the computer is programmed to disable a detector when it consistently fails a screening test (e.g., for 1 or 2 hr), then the detector fault would likely have been discovered before congested conditions began. Otherwise, the Washington State boundary may be used within the congested area.

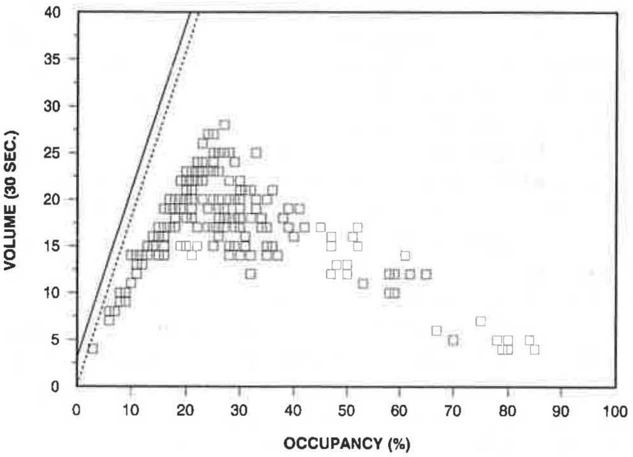


FIGURE 2 Upper bounds at Station 17, 6:00–11:00 a.m., May 15, 1990.

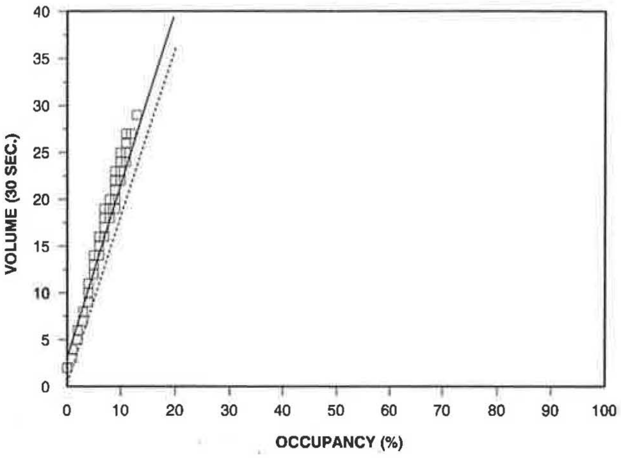


FIGURE 3 Upper bounds at Station 18, 6:00–9:00 a.m., May 15, 1990.

Screening Methods for Paired-Loop Systems

Three methods are discussed for paired-loop systems. The first is a comparison of the values obtained from the upstream loop of the pair with those from the downstream loop, for both volume and occupancy data. The second method identifies unusual combinations of speed, flow, and occupancy that may nonetheless contain valuable data. The final method involves setting feasible speed ranges for flow-occupancy pairs of data, on the basis of historical data.

Variation Between Loops at a Lane-Station

The two loops in a speed trap pair often have different volume and occupancy counts for the same 30-sec interval. Clearly, it is possible to have two volume counts that disagree but both of which pass the simple range test. The question, then, is how much of a difference to allow. A difference of two or less was selected. Because of the discreteness of the variables being measured on a 30-sec basis, it is expected that differences of one vehicle in the volume count will be encountered reasonably often. As well, there is the possibility of lane-changing maneuvers leading to a discrepancy of one or two vehicles, even between loops spaced only 6 m apart. Figure 4 shows that for properly functioning loops a difference of two covers nearly all of the data. There are other stations that regularly report larger differences, which probably indicates a loop pair in need of calibration.

A similar situation might be expected for occupancy. Inspection of the data, as shown in Figure 5, disproves this assumption. For uncongested flow (i.e., occupancies below about 25 percent in the figure), the difference in the two measurements of occupancy is small, but for congested flow the differences become much larger. The explanation is obvious, once the data have been seen. During uncongested travel, each vehicle is moving at roughly constant speed and occupies the two detectors in the pair for nearly the same amount of time. However, during congested flow, each vehicle is continually accelerating or decelerating and, thus, travels at somewhat different speeds as it crosses even these closely spaced detectors.

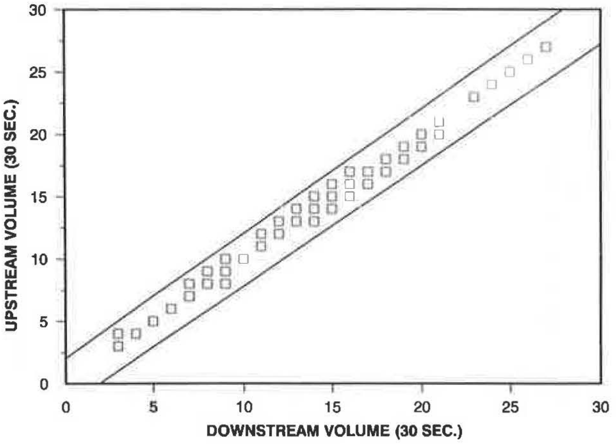


FIGURE 4 Upstream versus downstream volume at a double-loop station (Station 25).

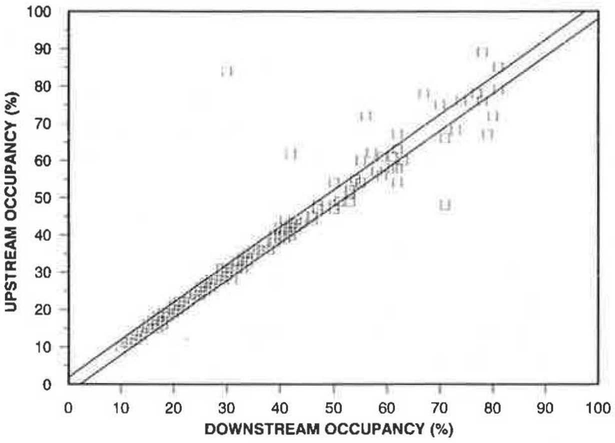


FIGURE 5 Upstream versus downstream occupancies at Station 14.

Unusual Values of Speed, Flow, and Occupancy Data

Not all apparently impossible combinations of the variables are actually bad data: the technicalities of data collection can result in some strange values. For example, observations with small positive occupancies (such as 1 or 2 percent) and zero volumes are occasionally found. These observations can be explained by the ways in which volume and occupancy are detected by a typical inductance loop detector and amplifier system.

Consider a signal generated by a vehicle passing over a detector loop. As long as the entire signal is contained within the same 30-sec interval, the output will be as expected (one volume count defined by the signal's leading edge, and a small positive occupancy count, the magnitude of which depends on the speed of the vehicle). If the 30-sec interval ends just after the leading edge of the signal, however, the subsequent interval has a small positive occupancy but no volume count (assuming no more vehicles pass over the detector in this interval).

Another seemingly nonsensical observation, that of zero occupancy and small positive volume, occurs several orders of magnitude more often than the zero volume observation. This observation can also be explained by the data collection process. Occupancy is reported as an integer value, so some information is lost in the truncation that produces the integer. This truncation means that a detector must be occupied for at least 0.30 sec in one interval to register an occupancy of 1 percent. A vehicle traveling at 100 km/hr must have length (as perceived by the detector) of at least 8.33 m to register 1 percent occupancy. For higher vehicle speeds, the perceived vehicle length over the 30-sec interval must be larger (e.g., 10.8 m at 130 km/hr). It is possible that two vehicles together will not exceed this perceived length threshold. The result is a positive volume accompanied by zero occupancy.

The final two types of data to be considered bring in the speed variable. In discussing them, it is important to recall that speed is measured across the two closely spaced loops in these systems and is not calculated from the flow and occupancy data. In the first case, speed is missing, but the other two variables are present. This observation is valid and should not call into question the other two variables. Speed is missing when the controller logic is unable to sort out vehicle arrivals at the two loops (e.g., lane changes or stop-and-go traffic), but this problem does not affect the other two variables. In the second case, speed is present and positive, but both volume and occupancy are zero. This condition caused numerous false alarms in an incident detection algorithm on the Skyway system before the error was discovered. Each variable on its own passed the threshold tests, so the additional test was needed to uncover the error.

Screening of Speed, Flow, and Occupancy Data on the Basis of Historical Data

Figure 6 shows 3 hr of FTMS data, representing both congested and uncongested operation. Several points in the speed-occupancy plot represent unrealistic conditions, such as 100 percent occupancy at 35 km/hr and 60 percent occupancy at 85 km/hr. There are, of course, many other possible combinations of the variables that are unreasonable. This type of error seems to occur at random intervals in the data set examined and has been observed in different lanes and stations. To catch the error, a screening mechanism that uses all three variables is needed.

The first step in designing such a screening method was to determine how a computer algorithm could describe the range of possible observations of the variables. In other words, the problem is to represent the region occupied in three dimensions by the set of all observable combinations of the traffic variables. A fitted mathematical model would be the ideal tool for representing the data in three-space because an equation is easy to evaluate and the decision on the validity of an observation would be trivial. Such a model would define a boundary about the region occupied by the data in three dimensions. Unfortunately, a simple mathematical model could not easily be applied to enclose such an irregular shape.

The next logical choice was a graphical model, that is, a method of defining the acceptable region of observations without using mathematical curve-fitting techniques. Because the shape in three dimensions was of no known polyhedron or other simple object, slicing the object into two-dimensional planes seemed to be a good way to simplify the problem. Volume was selected as the axis to divide the region because it is the only variable of the three that must be an integer by definition. It is also conveniently the variable with the smallest range, from 0 to perhaps 25 vehicles per 30-sec interval, whereas both occupancy and speed are observed over at least four times this range. Using volume results in the fewest slices to deal with later in the algorithm. The slices were made parallel to the speed versus occupancy projection so that each slice contained points representing various speed and occupancy pairs observed at a given flow rate.

The next task was to design a perimeter about the observations on each slice of the three-dimensional region. Several methods were tried. The best approach started with the further division of each speed-occupancy slice of the three-dimensional region into columns of constant occupancy. The choice of occupancy over speed was arbitrary, but it has the advantage of a constant range of observations (from 0 to 100 percent), which simplifies coding. The mean and standard deviation of the speed observations at each occupancy were calculated. The vertices of the perimeter were the mean plus or minus a number of standard deviations for each occupancy value (see Figure 7). This method provided a simple way to test points against the calibrated perimeter. New observations could be compared with the maximum and minimum of the speed range (derived from the mean and standard deviation) for the appropriate volume slice and occupancy column.

One simplification for the algorithm, reducing computation time and memory requirements, involves the assumption that the standard deviation of speed for a given flow-occupancy pair is constant over time. Only the means need to be updated as the algorithm runs operationally. To test this assumption, 3 days of FTMS data from 1989 at one station on the QEW-Mississauga system were examined. The days covered a range of weather and traffic conditions. For selected volume-occupancy combinations, the *F*-test was used to compare the variance of speed on each day with that on the other 2 days. The results do not strongly support the assumption of constant standard deviation, but they do show that the most consistent variances occur in the mid-range of volume-occupancy points. Hence, it is probably acceptable to retain the assumption.

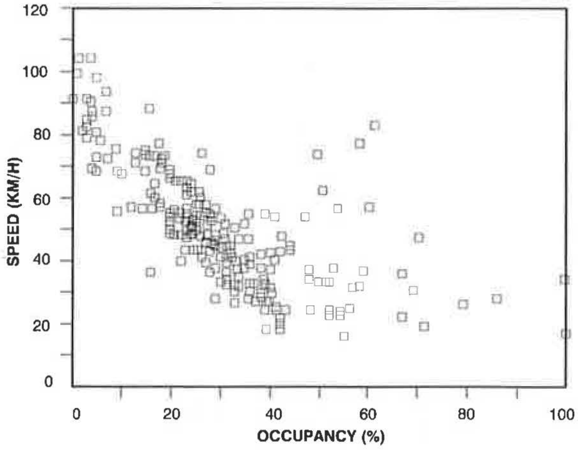


FIGURE 6 Three hours of FTMS data illustrating invalid combinations of speed and occupancy.

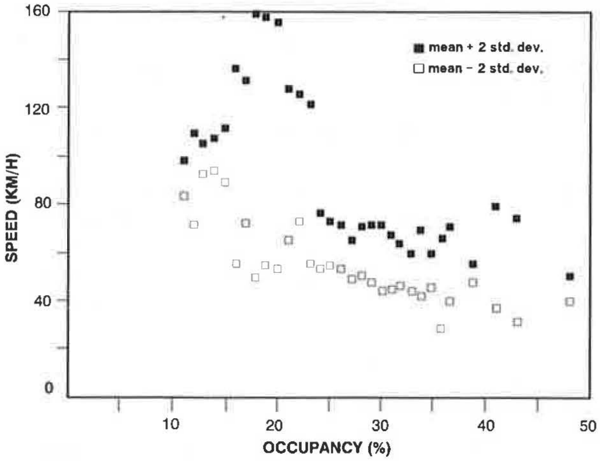


FIGURE 7 Calibrated speed ranges based on 7 days of data from QEW-Mississauga FTMS in 1989 (standard deviation calculated using 25 observations).

A final question about speed distribution remains: How many standard deviations should be used for the algorithm to accept good values and reject bad ones? It is known that a range of plus or minus three standard deviations will contain 99.7 percent of a normally distributed population. To check the assumption that speeds are distributed normally at each volume-occupancy pair, 4 days of weekday data from the median lane at three different stations on the QEW-Mississauga FTMS were examined. For each speed distribution, the coefficients of skewness and kurtosis were calculated. These statistics are qualitative indicators of the normality of a distribution. Skewness indicates whether or not one tail of a distribution is thicker than the other, with a zero value denoting a symmetric distribution. Kurtosis indicates the peakedness or flatness of a distribution relative to the normal distribution. A negative kurtosis represents a broad-peaked distribution, whereas a positive kurtosis indicates a sharp-peaked distribution. As with skewness, a zero value occurs when the peak of the distribution is normal.

As Table 1 indicates, roughly half of the distributions were skewed to each side, although the values were close to zero in almost all cases. This finding suggests that speed is distributed almost symmetrically about the mean for all volume-occupancy pairs. The kurtosis, however, indicates that 71 to 88 percent of the speed distributions had broader peaks than those of the normal distribution. Thus, the frequency of observations is relatively constant about the mean, instead of dropping off as does the normal distribution. For this reason, the choice of the number of standard deviations is made at the calibration stage, after the data have been inspected.

The final aspect of the algorithm design is the provision for on-line updating of the pattern to compensate for basic changes in freeway operations caused by such conditions as unfavorable weather (*11*). If the data are acceptable, they are incorporated into the mean for the volume-occupancy pair. Even though the standard deviation is held constant, the movement of the mean serves to move the range of acceptable points as freeway conditions change.

Calibration and testing of the three-dimensional screening algorithm were accomplished off-line on a microcomputer. The program calibrated each volume-occupancy pair according to a specified number of observations required. Once a volume-occupancy pair was calibrated, the program immediately started screening with the calibrated mean and standard deviation, using a user-supplied multiplier for the standard deviation.

Two parameters are needed before calibration: (a) the number of observations required to build the distribution of speed at each volume-occupancy pair and (b) the range about the mean speed that will be accepted as valid, expressed as a number of standard deviations. Limited analysis of QEW FTMS data has shown that, to obtain an accuracy of ±10 km/h at the 95 percent confidence level, the number of observations required is between roughly 15 and 40, assuming that the sample standard deviation is a valid approximation of the true standard deviation of the population. To obtain an estimate of the suitable number of standard deviations to use, screening trials were run using the 3 days of 1989 FTMS data from the QEW-Mississauga system. The 3 days of data were run as one contiguous piece of data, simulating 3 consecutive days.

TABLE 1 SUMMARY OF SKEWNESS AND KURTOSIS RESULTS

		Skewness		Kurtosis	
Station	Number of Distributions	Number Positive	Number Negative	Number Positive	Number Negative
17	146	75	71	18	128
21	325	155	170	62	263
25	209	111	98	61	148

Overall, performance was as expected when the number of observations and the standard deviation multiple were varied (see Table 2). Larger numbers of observations used for calibration result in fewer points calibrated with the same number of observations, so the time to full calibration is increased. Increasing the multiple of the standard deviation results in reduced rates of data rejection. The rejection rate drops dramatically when *M* changes from 1 to 2, but the rate is reduced only slightly as *M* changes from 2 to 3.

On more than one occasion, the acceptable range for speed given the volume and occupancy reported by one loop did not include the observed speed, but, when the volume and occupancy from the second loop were used, the observed speed fell within the acceptable range. For this reason, the secondary loop data should be used for a confirmation check if available, before an observation is rejected. In this way, speed data can be used to select between volume-occupancy data that are not consistent for the two loops at a station.

CONCLUSIONS

The quality of FTMS data is an issue critical to the operation of the entire system. Currently accepted screening methods

TABLE 2 SCREENING RESULTS FOR VARYING COMBINATIONS OF NUMBERS OF OBSERVATIONS USED FOR CALIBRATION AND NUMBER OF STANDARD DEVIATIONS AWAY FROM THE MEAN

	N = 20					N = 25				
M	# obs	ok	fast	slow	% fail	# obs	ok	fast	slow	% fail
1	2317	1630	430	257	29.7	2055	1495	329	231	27.3
2	2317	2226	6	85	3.9	2055	1989	0	66	3.2
3	2317	2285	0	32	1.4	2055	2030	0	25	1.2

	N = 35					N = 40				
M	# obs	ok	fast	slow	% fail	# obs	ok	fast	slow	% fail
1	1657	1218	254	185	26.5	1496	1095	235	166	26.8
2	1657	1605	0	52	3.1	1496	1443	0	53	3.5
3	1657	635	0	22	1.3	1496	1477	0	19	1.3

	N = 65				
M	# obs	ok	fast	slow	% fail
1	863	621	131	111	28.0
2	863	821	0	42	4.9
3	863	847	0	16	1.9

Notes: M is the multiple of standard deviations used in screening
N is the number of observations used for calibration
# obs is the number of observations screened
ok - observation passed screening
fast - speed was judged too high by screening
slow - speed was judged too low by screening

leave considerable room for improvement. Systems with paired loops in each lane at each station have more data screening options than those with single loops. Not only can the paired-loop systems compare the received speed-flow-occupancy points with a calibrated three-dimensional speed-flow-occupancy region, following on the Washington State idea, but they can also compare upstream and downstream loop values. A further advantage of tests based on two- or three-dimensional regions is that the single-variable threshold tests would be unnecessary and could be eliminated. Hence, the overall data screening could be much more critical of the data.

The two- and three-dimensional tests provide better data screening and, at the same time, monitor the system for defects more effectively. A detector that consistently produces unreasonable combinations should be investigated and recalibrated. The improved monitoring resulting from this increased sensitivity would result in a system that produces better data overall and thus more successful operation of algorithms for such purposes as incident detection.

ACKNOWLEDGMENTS

This work was supported by a contract from the Ontario Ministry of Transportation and by an operating grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). Funding was also provided through a NSERC undergraduate student research award. The assistance of the Freeway Traffic Management Section of the Ontario Ministry of Transportation in making data available is greatly appreciated. Within that section, special mention should go to Phil Masters, Paul Lim-Hing, and Mark Fox.

REFERENCES

1. C. L. Dudek, C. J. Messer, and A. K. Dutt. Study of Detector Reliability for a Motorist Information System on the Gulf Freeway. In *Transportation Research Record 495*, TRB, National Research Council, Washington, D.C., 1974, pp. 35–43.
2. K. G. Courage, C. S. Bauer, and D. W. Ross. Operating Parameters for Main-Line Sensors in Freeway Surveillance Systems. In *Transportation Research Record 601*, TRB, National Research Council, Washington, D.C., 1976, pp. 19–26.
3. L. Chen and A. D. May. Traffic Detector Errors and Diagnostics. In *Transportation Research Record 1132*, TRB, National Research Council, Washington, D.C., 1987, pp. 82–93.
4. C.-Y. Jeng and A. D. May. *Monitoring Traffic Detector Information and Incident Control Strategies: A Survey*. Working Paper UCB-ITS-WP-84-2. Institute of Transportation Studies, University of California, Berkeley, June 1984.
5. L. N. Jacobson, N. L. Nihan, and J. D. Bender. Detecting Erroneous Loop Detector Data in a Freeway Traffic Management System. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 151–166.
6. H. J. Payne, E. D. Helfenbein, and H. C. Knobel. *Development and Testing of Incident Detection Algorithms, Volume 2: Research Methodology and Detailed Results*. Report FHWA-RD-76-20. FHWA, U.S. Department of Transportation, April 1976.
7. 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.
8. *Burlington Skyway Freeway Traffic Management System Software Functional Description*. Report WP-85-81-02. IBI Group, Toronto, Ontario, Canada, and JHK & Associates, Atlanta, Ga., July 1983, rev. Feb. 1984.
9. P. Athol. Interdependence of Certain Operational Characteristics Within a Moving Traffic Stream. In *Highway Research Record 72*, HRB, National Research Council, Washington, D.C., 1965, pp. 58–87.
10. F. L. Hall and B. N. Persaud. An Evaluation of Speed Estimates Made with Single-Detector Data from Freeway Traffic Management Systems. In *Transportation Research Record 1232*, TRB, National Research Council, Washington, D.C., 1989, pp. 9–16.
11. F. L. Hall and D. Barrow. Effect of Weather on the Relationship Between Flow and Occupancy on Freeways. In *Transportation Research Record 1194*, TRB, National Research Council, Washington, D.C., 1988, pp. 55–63.

Publication of this paper sponsored by Committee on Freeway Operations.

Operational Evaluation of Truck Restrictions on I-20 in Texas

Michael C. Zavoina, Thomas Urbanik II, and Wanda Hinshaw

With the increased expansion of highways to six lanes, questions have arisen as to the proper operational strategy of those facilities. One strategy gaining support is to restrict large vehicles from one or more lanes. The effects of such restrictions, however, had not been extensively studied. The operational effects of a truck restriction on I-20 near Fort Worth, Texas, were analyzed. Vehicle distributions according to classification, vehicle speeds, and time gaps between vehicles were examined to evaluate the operational effectiveness of this left-lane truck restriction. No analysis or evaluation of accident rates or pavement wear was undertaken. Although the directional distribution of trucks changed significantly because of the restriction, no effects were found that could be attributed to the truck restriction in the directional distribution of cars, speeds of either cars or trucks, or time gaps between vehicles.

Recent emphasis of transportation engineering has shifted from the design of new facilities toward maintaining, enlarging, and improving the operation of existing ones. Computer traffic-monitoring systems, changeable message signs, and high-occupancy-vehicle lanes have all been used to improve the operational characteristics of a facility. Another area in which such changes are taking place is in the operational strategy of multilane highways, in which restrictions of trucks, essentially all large vehicles and vehicles pulling trailers, are being examined as tools to increase operational performance. Some engineers and highway users suggest that large trucks are impeding the free flow abilities of smaller vehicles. In the purest sense, because the average speed of trucks was less than the average speed of cars, trucks must be impeding cars, at least to some degree. It has been suggested that trucks should be restricted, leaving one or more lanes clear for non-truck traffic. Conversely, it has been suggested that increasing the concentration of truck traffic would induce increased pavement damage and otherwise restrict the movement of other vehicles. Restricted access of entering vehicles may decrease safety, whereas increased pavement wear may demand higher design standards for the roadway, at least for the outside lanes. The validity of these suggestions has not been fully evaluated by previous research. This lack of relevant research, therefore, indicates need for a well-designed, controlled experiment to study the effects of truck restrictions.

OBJECTIVES

The limitations of most research on this subject center on the lack of a proper study design in which the conditions both before and after a lane restriction are thoroughly examined. It is the primary objective of this study to perform such an experiment. Specifically, this study is concerned with the highway operations aspect of truck restrictions. Although the focus of the following paragraphs is on the left-lane truck restriction at Interstate 20, restrictions were implemented and studied at two other sites in Texas as part of a larger study sponsored by the Texas State Department of Highways and Public Transportation (*1*). All results and conclusions presented were supported by these other studies, except as noted.

BACKGROUND

FHWA surveyed the 50 states, the District of Columbia, and Puerto Rico in June 1986 to study the extent to which lane restrictions had been used (*2*). All 52 surveys were returned, 28 of which reported using no restrictions. The other states reported using restrictions, usually temporarily, for one or more of the following reasons:

1. To improve highway operations,
2. To reduce accidents,
3. To provide for more even pavement wear, and
4. To ensure better operation and safety through construction zones.

A similar study was performed by Sirisoponsilp and Schonfeld (*3*) in February 1988. This study also surveyed the 50 states, the District of Columbia, and Puerto Rico, but only 31 of those surveyed responded. Fourteen states reported having experience with truck restrictions, whereas 17 states reported no experience. Of the 14 states having truck restriction experience, 5 implemented statewide restrictions and 4 were currently studying their effectiveness.

Other studies (*4,5*) have analyzed the effects of truck restrictions on highway operations. However, because of a lack of before-and-after comparisons and an inadequacy of control, the results of these studies are questionable. Therefore, reliable conclusions cannot be drawn without the results first being replicated. Perhaps the most comprehensive research on this subject was performed by Hanscom (*6*) in 1989. A before-and-after study design with a control site was used to evaluate the effectiveness of three different truck restrictions. Hanscom reported voluntary compliance by a higher percentage of trucks at all three sites. To determine the impedance of cars by trucks, the average platoon lengths behind trucks during the before and after periods were computed. The average platoon length change between the before and

Texas Transportation Institute, Texas A&M University System, College Station, Tex. 77843-3135.

after periods for the test and control sites was then compared, and significant differences between the two were found. The report also found that there were no adverse speed effects resulting from the restrictions. Although the study design was very good, the lack of an appropriate control site makes conclusions based on comparisons between the test and control sites less meaningful. The control sites existed upstream of the test sites and were reported to differ both in composition of traffic and in total volume. The platoon length discussed previously was highly dependent on volume, and, because the volume of the test and control sites varied by as much as 30 percent, that measurement was questionable for determining impedance. Because volumes were determined by 5-min counts and manual methods were used to measure speeds and determine following distances, further doubt is cast on the validity of the results. The need for replication of these results, therefore, is evident.

PUBLIC SURVEYS

In order to obtain public comment and aid in the development and analysis of the restriction, public surveys were conducted both before and after implementation of the restriction. Two initial surveys were conducted before the restriction: (a) a trucker survey of large-vehicle operators and (b) a motorist survey of automobile operators. These surveys were structured to determine the opinions of drivers as well as to determine the most effective signing system to convey the intent of the restriction. In addition, three secondary surveys of similar format, two of motorists and one of truckers, were conducted after the restriction. These surveys were intended to poll drivers who had experienced the restriction, and the surveys were structured to determine the opinions of drivers about the effectiveness of the sign as well as the restriction in general.

Initial Surveys

A total of 124 motorist surveys and 140 trucker surveys were completed. These surveys were structured to perform two principal functions: (a) gain an understanding of the opinions of motorists and truckers toward a left-lane restriction of large vehicles and (b) identify which sign best relates the intended message both to motorists and to truckers.

To accomplish the first objective, a question was asked about the respondent's feelings toward the left-lane restriction of trucks. This question revealed that 60 percent of motorists favored the restriction of larger vehicles, whereas only 28 percent of truckers favored such a restriction. Of the truckers not favoring the restriction, 19 percent stated that the restriction would cause merging conflicts, 14 percent stated that the restriction would impede cars, and 13 percent stated that the restriction would cause undue congestion.

In order to accomplish the second objective, three different regulatory signs were introduced. For each sign, questions were asked as to the legality of driving certain types of vehicles in the left lane. The effectiveness of each sign, therefore, was determined by computing the percentages of correct responses to each sign option. The sign that exhibited the highest percentages of correct responses by both types of drivers was deemed the clearest in conveying its meaning. This process ultimately yielded a sign that read No Trucks Trailers in Left Lane.

Secondary Surveys

A total of 345 motorist surveys and 87 trucker surveys were completed. These surveys were designed to perform two principal functions: (a) assess the effectiveness of the sign in being noticed and in conveying its intended meaning and (b) assess the opinions of motorists and truckers toward the restriction and its impact on highway operations.

For the first objective, respondents were first asked if they noticed the sign while driving on the facility. Sixty-eight percent of the motorists answered positively to this question, as did 76 percent of the truckers. Although this result also means that 32 percent of the motorists and 24 percent of the truckers did not see the signs, those missing the signs might have been traveling in the right two lanes. Additionally, the respondents were asked to which types of vehicles they thought the sign applied. This question was similar to those in the initial trucker and motorist surveys, in which the effectiveness of the sign was analyzed by determining the percentages of correct responses to each vehicle type. The average percentage of correct responses was 88 and 73 percent for the motorists and truckers, respectively.

In order to accomplish the second objective, respondents were asked if they thought the restriction improved operations. Forty-five percent of the motorists thought the restriction had improved operations, whereas only 20 percent of the truckers believed that it had. Large percentages of respondents were unsure whether operations had improved (43 percent of motorists and 28 percent of truckers). Approximately half of each group of respondents provided comments. Twenty percent of all motorists surveyed offered that the restriction was a good idea; to the contrary, 28 percent of the truckers suggested that it was a poor idea.

STUDY METHODOLOGY

The three study sites mentioned previously were chosen primarily because they met the requirements of the study. That is, they were chosen because they were six-lane, rural Interstate highways with a speed limit of 65 mph (60-mph truck speed limit). Although a control section would be desirable, no such sections were used because the short lengths of the sites would undoubtedly induce end effects. Although one direction could have been used as the control section, the problems associated with reducing the effects of the intervening variables (volume, directional distribution, truck percentage, peak period, and geometry) would have made before-and-after comparisons weak at best. The site highlighted was located on I-20 between Fort Worth and Weatherford, Texas. The total length of the section was approximately 9 mi, and its 1988 average annual daily traffic (AADT) was 39,000 vehicles. There was an approximate 3 percent upgrade in the eastbound direction and a 3 percent downgrade in the westbound direction. Although flat grade would have been desirable across both directions, conclusions can still be drawn as long as the effect of the grade is considered in all analyses.

Data Collection

A system of tapeswitches was designed and implemented to collect all data. Two tapeswitches, placed 15 ft apart to permit the computation of vehicle speeds, were temporarily installed across all traffic lanes. A computer program detected and collected the electronic activations of all tapeswitches, while another computer program reduced the activations into useful operational data. To check the accuracy of the collection and reduction procedure, the highway segment was also videotaped during all data collection periods. To obtain the base conditions of the highway, 10 hr of data were collected during peak and nonpeak periods on August 17 and 18, 1988. One year later, in August 1989, signs were manufactured and installed. These signs were spaced evenly throughout the length of the site, with approximately 1 mi between signs. Two signs were placed back to back on one pole in the median, and a Begin and End message sign was placed atop the first and last signs, respectively. No attempts were made to enforce the restriction. After a 90-day adjustment period, 15 hr of data were collected on November 14, 15, and 16, 1989.

A peak period was defined to account for volume differences within the data. In this manner, speed and headway averages could be compared with some confidence that volume did not control the results. Peak periods were determined by examining 15-min volumes in both directions. The intervals in which the volumes were clearly greater than the average were considered to be in the peak period, with everything else constituting the nonpeak period. Using this analysis, the peak periods were defined as 4:00 to 6:00 p.m. in the westbound direction and 6:30 to 8:30 a.m. in the eastbound direction.

Operational Characteristics and Data Manipulations

Three operational characteristics of traffic flow were examined: (a) vehicle classification, (b) vehicle speed, and (c) time gap between vehicles. The time gap between vehicles is that between the passing of a fixed point by the rear axle of the leading vehicle and the front axle of the following vehicle. This parameter was deemed applicable (more than leading or lagging headway) because it did not incorporate vehicle length and therefore provided a more accurate description of how closely vehicles follow one another.

The vehicle classification system used was designed such that comparisons could be made between groups of vehicles, with each group composed of vehicles assumed to have similar operating characteristics. It is, therefore, important to understand the limitations of the classification system, as some vehicles may exhibit the characteristics of more than one classification. Two classifications were used: (a) vehicles with two axles and (b) those with more than two axles. This system permits assumptions to be made about the nature of the vehicles within them. Vehicles with two axles generally include passenger cars, pick-ups, vans, motorcycles, and some single-unit trucks; vehicles with more than two axles generally include some single-unit trucks, passenger cars and pick-ups pulling trailers, all tractor-trailer combinations, and buses. For simplification, vehicles with two axles are referred to as cars, and vehicles with three or more axles are called trucks.

In order to analyze the distribution of vehicles, the percentage of vehicles in each classification was examined as a percentage of (a) total vehicles in the lane, (b) total vehicles in the direction, and (c) total vehicles in both directions. In addition, the percentage of each classification in a lane as a percentage of the total number of vehicles of the same classification in a direction was examined. This type of analysis relates how each classification of vehicle is distributed across a direction. Changes in the distributions of trucks and cars across each direction were evaluated to determine the impacts of the restriction. A redistribution of trucks was certainly expected, but the magnitude of the change was not known. The redistribution of trucks may effect a corresponding redistribution of cars, and the combination of these changes may have either positive or negative consequences.

In order to analyze vehicle speeds, the arithmetic mean of the speeds of each vehicle classification in each lane was computed and examined. Because the sequence of vehicles was also known, the arithmetic means of the speeds of cars following cars and of cars following trucks were computed and examined (for all lanes). One important point to consider when forcing trucks to switch lanes is that those trucks may not have adjusted their speeds after switching lanes. Furthermore, which trucks were complying with the restriction, the faster or the slower ones? This question is important because those trucks previously in the left lane were exceeding the speed limit on average by as much as 10 mph. If the trucks that switched lanes did not adjust their speeds, an increase in the speed differential in the center or even the right lane might result, thereby increasing the potential for hazard within those lanes. In addition to examining the average speeds of groups of vehicles, the speed differential was computed for each lane. A decrease in the speed differential in a lane would likely increase the safety of the roadway. The absolute value of the difference between the speed of each vehicle and that of the preceding vehicle was calculated. The average speed differentials before and after the restriction were then compared to test for statistical significance. Finally, the cumulative distribution functions of the speed for each classification of vehicle were examined. The cumulative distribution function relates important information about the variability of the speeds within a lane. The manner in which the cumulative distribution functions were examined is explained at the end of this section.

In order to analyze the time gaps between vehicles, the arithmetic mean of the time gaps of four groups of vehicles was computed and examined in each lane. These four groups included the four combinations of the leading vehicle's classification and the following vehicle's classification. Although these average time gaps provide a general indication of how closely vehicles are following one another, they are meaningless unless all vehicles are evenly distributed throughout time. The use of peak and nonpeak periods reduces the chance of irregular distributions but does not guarantee a perfectly regular one. Therefore, conclusions based on average time gaps should be made knowing these limitations. The entire distribution of time gaps should be examined to fully differentiate changes between the before and after stages. Examination of the cumulative distribution functions allows comparisons to be made only at the smaller time gaps, in which differences result from interactions among vehicles. This distinction is important because, as the time gaps become larger, the influence the leading vehicle has on the following vehicle is minimal and, instead, depends heavily on traffic volume.

A hypothesis provided by some highway users in the surveys mentioned previously was that, because faster trucks would be forced into the slower traffic stream of the outside lanes, they would bear down on the cars in front of them. To test this hypothesis, vehicle pairs consisting of cars followed by trucks were identified and the time gaps between them examined. The proportion of pairs in each lane in which the time gaps were less than 2.0 sec were then compared before and after the restriction. These tests were repeated using a criterion of 1.5 sec for the time gap.

The cumulative distribution functions both of the time gaps between vehicles and of vehicle speeds were examined for the following four variables: (a) stage (before or after implementation of the truck restriction, (b) period (peak or nonpeak); (c) lane, and (d) group. The group variable included the four combinations of the leading vehicle's classification and the following vehicle's classification. Therefore, the four groups were (a) cars following cars, (b) cars following trucks, (c) trucks following cars, and (d) trucks following trucks. Manipulation of these four variables allowed many different types of comparisons to be made. Four types of comparisons were made by plotting and examining the cumulative distribution functions in the following manners:

1. All four groups on one graph, with stage, lane, and period variable (24 graphs);
2. Both stages on one graph, with lane, period, and group variable (48 graphs);
3. Comparable lanes (left, center, or right lanes of both directions) on one graph, with lane pair (1-4, 2-5, and 3-6), stage, period, and group variable (48 graphs); and
4. Comparable lanes and both stages on one graph, with lane pair, period, and group variable (24 graphs).

Statistical Methods

Different statistical tests were used to test the significance of observed changes in the operational parameters discussed previously. An alpha level of 0.05 was used for all tests. The chi-square statistic was used to test the significance of changes in the directional distribution of vehicles. Contingency tables were constructed to observe changes in the distributions of vehicles across lanes from before to after the restriction. Each direction of traffic was analyzed separately, as was each classification of vehicle. Collapsed two-by-two contingency tables were used to test the changes in the proportions of trucks in the left lane. The use of the chi-square statistic in this case is equivalent to a z-test of differences between two proportions.

Speeds of vehicles according to classification were compared in several ways using analysis of variance techniques. A general linear model was constructed and included main effects for the direction of travel (westbound or eastbound), lane (1, 2, or 3), period (before or after), and vehicle classification (car or truck). The analyses for peak and nonpeak periods were carried out separately. Single degree of freedom contrasts were constructed to test for individual lane differences, when overall F-tests supported their use. Comparisons of interest included differences between the speeds of cars and trucks, differences from before to after the restriction in the speeds of each classification, and the comparison of this latter speed difference between the two classifications. The interaction effect between period and classification addresses the last comparison, that is, whether the change in the speeds of trucks from before to after the restriction was the same as the change in the speeds of cars.

Tests involving the calculated speed differentials were constructed similarly. An overall analysis of variance that included effects for direction, lane, and period was performed for the peak and nonpeak periods individually. F-tests for single degree of freedom contrasts were appropriately designed, on the basis of the results of the overall analysis. Interaction effects involving direction and lane were interpreted to indicate the need for individual comparisons when significant at the 0.25 level.

The change in the proportion of trucks following cars with time gaps less than 2.0 and 1.5 sec was tested from before to after the restriction. The chi-square statistic calculated from a two-by-two contingency table was used to test the significance of the difference in each lane. Again, this test is equivalent to a z-test of differences between two proportions.

RESULTS

Compliance

The rate of compliance was analyzed by comparing the percentages of trucks in the left lane before and after implementation of the restriction and calculating the percentage changes. Table 1 presents the percent reduction of truck traffic in the left lane as a percentage of truck traffic in the direction. This table indicates that the percentages of trucks in the left lane decreased by between 62 and 76 percent. All four of these percentage reductions were found to be statistically significant.

Table 2 presents the left-lane truck percentages before and after the restriction. As can be seen in this table, the percentage of trucks in the left lane, as a percentage of truck traffic, was 11.7 percent or less before the restriction. In addition, because of the small ratio of trucks to cars, the overall percentage of trucks in the left lane (as a percentage of total traffic) before the restriction was 1.3 percent or less. After the restriction, the percentage of trucks in the left lane as a percentage of truck traffic was 4.4 percent or less, and the percentage of trucks in the left lane as a percentage of total traffic was 0.4 percent or less. These results exhibit the near nonexistence of trucks in the left lanes at the study site after the restriction.

TABLE 1 PERCENT REDUCTION OF TRUCK TRAFFIC IN LEFT LANES

Westbound		Eastbound	
Peak	Non-Peak	Peak	Non-Peak
64%	76%	62%	64%

TABLE 2 LEFT-LANE TRUCK PERCENTAGES BEFORE AND AFTER IMPLEMENTATION OF TRUCK RESTRICTION

	Westbound		Eastbound	
	% of Total	% of Trucks	% of Total	% of Trucks
	Before			
Peak Period	0.7	8.9	1.3	11.7
Non-Peak Period	0.8	5.3	1.1	7.9
	After			
Peak Period	0.2	3.2	0.3	4.4
Non-Peak Period	0.2	1.3	0.4	2.9

Vehicle Distributions

From the data, the percentage of trucks as a percentage of total traffic can be easily determined. Before the restriction, the percentage of trucks across all lanes of traffic was 8.4 and 14.2 percent during the peak and nonpeak periods, respectively. After the restriction, the percentage of trucks was 6.6 and 15.0 percent during the peak and nonpeak periods, respectively. The percentage of truck traffic, therefore, remained fairly constant from before to after the restriction. Higher percentages of trucks during the nonpeak periods are also evident.

Table 3 presents the percentages of trucks and cars before and after the restriction, as well as the percentage changes. The distribution of trucks changed significantly across both directions and during both periods. In addition, these distributions changed in a peculiar way. As expected, the percentage of trucks decreased in the left lane by between 62 and 76 percent. Unexpectedly, however, the percentage of trucks decreased in the middle lane as well (between 6 and 23 percent), with the percentage of trucks increasing only in the right lane (between 17 and 60 percent). This same pattern of change appeared in both the westbound and eastbound directions and during both the peak and nonpeak periods. Although the percentage of trucks in the right lane increased by as much as 60 percent (eastbound peak period), that increase was generally less than 25 percent. These results suggest that the trucks that moved from the left lane to the center lane caused a subsequent movement of trucks from the center lane to the right lane. Although not likely, it is possible that the trucks moved from the left lane to the right lane. Whatever the case, the concentration of trucks in the right lane was more pronounced than expected, whereas the concentration of trucks in the middle lane was actually less than expected. This result was not substantiated at the other two sites (*1*). At those sites, truck traffic generally increased in both the middle and right lanes.

TABLE 3 CHANGES IN DISTRIBUTION OF TRUCKS AND CARS

Lane	Trucks		Cars	
	Percentage (Before - After)	Change	Percentage (Before - After)	Change
	Peak Period			
Westbound				
Left	8.9 - 3.2	DOWN 5.7%	31.1 - 32.8	UP 1.7%
Center	41.9 - 39.1	DOWN 2.8%	42.8 - 41.5	DOWN 1.3%
Right	49.2 - 57.7	UP 8.5%	26.1 - 25.7	DOWN 0.4%
Eastbound				
Left	11.7 - 4.4	DOWN 7.3%	35.2 - 37.3	UP 2.1%
Center	55.5 - 43.0	DOWN 12.5%	41.6 - 40.6	DOWN 1.0%
Right	32.8 - 52.6	UP 19.8%	23.2 - 22.1	DOWN 1.1%
	Non-Peak Period			
Westbound				
Left	5.4 - 1.3	DOWN 4.1%	20.7 - 19.3	DOWN 1.4%
Center	40.6 - 31.1	DOWN 9.5%	47.9 - 50.3	UP 2.4%
Right	54.1 - 67.7	UP 13.6%	31.3 - 30.4	DOWN 0.9%
Eastbound				
Left	8.0 - 2.9	DOWN 5.1%	29.1 - 27.4	DOWN 1.7%
Center	51.1 - 48.0	DOWN 3.1%	46.6 - 48.1	UP 1.5%
Right	41.0 - 49.1	UP 8.1%	24.3 - 24.5	UP 0.2%

In both directions, the distribution of cars did not change significantly during the peak period but did change significantly during the nonpeak period. The sample sizes during the nonpeak periods of both directions were so large, however, that any variation between the before and after periods would be found statistically significant. The actual differences were so small in all lanes (usually less than 2 percentage points) that they were of no practical importance. Therefore, there was no change of practical significance in the distribution of cars across both directions and during both periods. Table 3 shows that all of the changes are 7 percent or less in magnitude and exhibit no consistent pattern in their direction of change.

Vehicle Speeds

Table 4 presents the changes in the speeds of trucks and cars. Changes in the speeds of cars were examined to verify that any changes in the speeds of trucks were classification dependent. Table 4 indicates that the speeds of trucks consistently decreased in the westbound direction and, during the

TABLE 4 CHANGES IN SPEEDS OF TRUCKS AND CARS

Lane	Trucks		Cars	
	Average Speed (Before - After)	Change (mph)	Average Speed (Before - After)	Change (mph)
	Peak Period			
Westbound				
Left	69.7 - 66.1	DOWN 3.6 *	70.8 - 69.5	DOWN 1.3 *
Center	64.4 - 61.9	DOWN 2.5 *	66.4 - 65.4	DOWN 1.0 *
Right	60.7 - 58.4	DOWN 2.3 *	62.9 - 61.9	DOWN 1.0 *
Eastbound				
Left	66.6 - 64.4	DOWN 2.2	70.1 - 70.1	NONE
Center	61.2 - 61.1	DOWN 0.1	67.6 - 68.3	UP 0.7
Right	56.0 - 56.5	UP 0.5	65.0 - 65.4	UP 0.4
	Non-Peak Period			
Westbound				
Left	70.0 - 67.0	DOWN 3.0 *	70.6 - 69.3	DOWN 1.3 *
Center	64.6 - 61.6	DOWN 3.0 *	66.4 - 65.2	DOWN 1.2 *
Right	60.7 - 57.7	DOWN 3.0 *	63.5 - 61.6	DOWN 1.9 *
Eastbound				
Left	64.6 - 68.4	UP 3.8 *	68.8 - 69.8	UP 1.0 *
Center	61.2 - 61.5	UP 0.3	66.0 - 67.0	UP 1.0 *
Right	57.0 - 57.9	UP 0.9 *	62.8 - 64.1	UP 1.3 *

* Denotes statistically significant change

nonpeak period, increased in the eastbound direction. The speeds of cars exhibited the same pattern of change, although the changes were generally smaller in magnitude. The changes were not found to be statistically significant in the eastbound direction during the peak period, when a significant increase in volume occurred. Because these changes were different for direction and period and were consistent for cars and trucks, it cannot reasonably be concluded that the changes were in response to the truck restriction. Therefore, there seems to be no positive or negative impact, as far as speeds are concerned, associated with the redistribution of trucks.

The average speed differential in each lane, as defined previously, was compared from before to after the restriction. Table 5 presents these average speed differentials, as well as the results of the tests for statistical significance of the difference from before to after the restriction. As can be seen in this table, 2 of the 12 tests revealed statistically different speed differentials. These decreases in the speed differential occurred in the eastbound direction during the peak period. (Although the speed differentials generally did not change after the restriction, it is of interest that they consistently increased across a direction from the median to the outside lane both before and after the restriction.)

Testing of the interaction between the average speeds of trucks and cars from before to after the restriction yielded inconsistent results. Table 6 presents the difference between the speeds of cars and trucks (average speed of cars minus average speed of trucks) in each lane before and after the restriction, as well as the results of the tests for statistical significance. Table 6 indicates that the speed difference between cars and trucks generally increased after the restriction. This result is consistent with the finding that truck speeds generally changed from before to after the restriction by a larger amount than car speeds. The increases are most pronounced in the westbound direction, which has a 3 percent downgrade.

TABLE 5 COMPARISON OF AVERAGE SPEED DIFFERENTIALS FROM BEFORE TO AFTER THE RESTRICTION WITH TEST RESULTS

Lane	Speed Differential (Before - After)	Difference Between Before and After	Statistical Significance
Peak Period			
Westbound			
Left	3.3 - 3.1	-0.2	NO
Center	3.7 - 3.6	-0.1	NO
Right	4.8 - 4.7	-0.1	NO
Eastbound			
Left	4.0 - 3.3	-0.7	YES
Center	4.7 - 4.7	0.0	NO
Right	6.8 - 6.3	-0.5	YES
Non-Peak Period			
Westbound			
Left	4.4 - 4.4	0.0	NO
Center	4.3 - 4.4	0.1	NO
Right	5.5 - 5.5	0.0	NO
Eastbound			
Left	4.1 - 4.1	0.0	NO
Center	4.9 - 4.9	0.0	NO
Right	6.5 - 6.5	0.0	NO

TABLE 6 COMPARISON OF DIFFERENCES BETWEEN SPEEDS FROM BEFORE TO AFTER THE RESTRICTION WITH TEST RESULTS

Lane	Difference between Car and Truck Speeds (Before - After)	Difference between Before and After	Statistical Significance
Peak Period			
Westbound			
Left	1.1 - 3.4	2.3	YES
Center	2.0 - 3.5	1.5	YES
Right	2.2 - 3.5	1.3	YES
Eastbound			
Left	3.5 - 5.7	2.2	NO
Center	6.4 - 7.2	0.8	NO
Right	9.0 - 8.9	-0.1	NO
Non-Peak Period			
Westbound			
Left	0.6 - 2.3	1.7	YES
Center	1.8 - 3.6	1.8	YES
Right	2.8 - 3.9	1.1	YES
Eastbound			
Left	4.2 - 1.4	-2.8	YES
Center	4.8 - 5.5	0.7	YES
Right	5.8 - 6.2	0.4	NO

In analyzing the cumulative distribution functions of the vehicle speeds, the methods of comparison discussed previously were used, and the 144 graphs were prepared and examined. When examining the cumulative distribution functions of only one stage (before or after) on one graph, the effect of grade on the speeds of trucks became quite clear. The cumulative distribution functions of trucks in the lanes of the eastbound direction were shifted to the left (lower speed) relative to the cumulative distribution functions of cars. As stated, there is a 3 percent downgrade in the westbound direction and a 3 percent upgrade in the eastbound direction. In Table 4, the average speeds of trucks in comparable lanes, during both periods and during both stages, were also always less in the eastbound direction. This finding suggests that, especially with steeper grades, trucks in the left lane may impede the free flow ability of cars. In trying to discover the differences between the data collected before and after implementation of the restrictions, the graphs were examined in an attempt to detect meaningful differences. However, no consistent differences could be found, even taking all variables into consideration. Therefore, under the current conditions, these results suggest that a redistribution of trucks does not effect any meaningful changes in the speeds of either cars or trucks.

Time Gaps Between Vehicles

In testing whether or not trucks bore down on cars more after the restriction, the chi-square tests discussed previously were performed, with both a 2.0- and a 1.5-sec time gap criterion. Table 7 presents the proportions of gaps less than 1.5 and 2.0 sec before and after the restriction, as well as the results of the tests for statistical significance. On the basis of a 2.0-sec time gap, only 1 of the 12 tests was found to be statistically significant. The tests based on a 1.5-sec time gap produced

TABLE 7 TIME GAP COMPARISON RESULTS

Lane	Proportion of Gaps < 1.5 Seconds	Proportion of Gaps < 2.0 Seconds
	Peak Period	
Westbound		
Left	25.0 - 22.2	25.0 - 27.8
Center	17.8 - 19.3	31.4 - 29.2
Right	5.3 - 7.7	12.8 - 12.4
Eastbound		
Left	8.3 - 20.0	8.3 - 20.0
Center	8.3 - 6.3	11.7 - 19.0
Right	0.0 - 6.7	0.0 - 13.5 *
	Non-Peak Period	
Westbound		
Left	6.0 - 15.0	10.0 - 20.0
Center	8.5 - 7.5	15.0 - 11.0
Right	2.9 - 2.5	7.2 - 5.7
Eastbound		
Left	5.3 - 8.0	10.6 - 10.0
Center	8.5 - 7.7	13.4 - 11.0
Right	1.8 - 5.4 *	4.2 - 5.4

* Denotes statistically significant change

similar results, with a different single comparison being found significant. Therefore, according to these results, the hypothesis that the restriction would cause trucks to bear down on cars more frequently was not substantiated.

The cumulative distribution functions of the time gaps between vehicles were analyzed to determine the changes, if any, that occurred between the before and after periods, or if meaningful observations could be made by considering only one stage. When analyzing the time gaps from one stage, the only consistent observation was that the time gaps of trucks following trucks were less than those of trucks following cars. Furthermore, the time gaps of trucks following trucks were usually also less than those of cars following cars and of cars following trucks. This observation generally held true during both stages, across all lanes, and during both periods.

When examining the cumulative distribution functions for gaps of both stages (before and after) on one graph, it was first established that the volume had not changed significantly since the restriction was implemented. If the volume changed coincident with the implementation of the truck restriction, the headways likewise changed because of the interdependence of headway and volume, thereby making headway comparisons between stages meaningless. To determine if the volume changed, the 15-min flow rates examined previously for determining the peak and nonpeak period definitions were again examined. Averages of the 15-min flow rates for each stage, lane, and period were taken but no significant volume changes were found except during the peak period in the eastbound direction. In that direction, volumes increased by between 31 and 58 percent from the before to the after stage. Therefore, conclusions based on observations of the cumulative distribution functions in the lanes of that direction should be made with knowledge of this change in volume.

To discover differences between the data collected before and after implementation of the restrictions, all of the graphs described previously were prepared and examined. Again, just as with vehicle speeds, no consistent differences could be found, even taking all variables into consideration. However, the site is rural and has a low volume-to-capacity ratio. With such large average headways, even during the peak period (usually greater than 5 sec), vehicles were not greatly affected by the vehicle in front of them. Therefore, it was difficult to detect meaningful differences in the time gaps between vehicles under these conditions. Nevertheless, the results suggest that a redistribution of trucks does not effect any discernible changes in the time gaps between vehicles.

CONCLUSIONS

This study, which examined the effects of a left-lane truck restriction on a rural six-lane Interstate in Texas, produced no startling results. Speeds, time gaps, and classifications of vehicles were recorded before and after implementation of the restriction, allowing the examination of several parameters in the investigation of the effects. The restriction succeeded in its purpose, with compliance rates between 62 and 76 percent achieved without enforcement. Few trucks were present in the left lanes of the roadway before the restriction. After the restriction, only 3 percent of all trucks remained in the left lanes. At the reported study site, the percentages of trucks significantly increased in the right lane (only) of each direction. At two other sites in Texas that were examined and reported on elsewhere (*1*), the percentages of trucks significantly increased in both the center and the right lanes. This redistribution of trucks across the three lanes of each direction did not cause any change of practical significance in the distribution of cars.

The examination of speeds resulted in statistically significant changes in the speeds of trucks relative to cars from before to after the restriction; however, the changes could not be attributed to the truck restriction. In the westbound direction only, the average speeds of trucks decreased more than those of cars after the restriction. The presence of a 3 percent downgrade in this direction cannot be ignored in interpreting this result. Furthermore, the possibility that enforcement activity was the cause of the larger shift in truck speeds should be considered, as trucks in the left lane were exceeding the speed limit by as much as 10 mph on average. Comparative shifts in the speeds of vehicles in the eastbound direction (3 percent upgrade) were only sporadically significant and, when significant, were not consistent in their direction of change.

Speed changes from before to after the restriction were also investigated by comparing the speed differentials in each lane. This measure was defined as the absolute value of the difference in the vehicle speeds within each pair of consecutive vehicles. Two significant reductions in average speed differentials were found. These changes occurred during the peak period in the eastbound direction. A coincident increase in volume, which also occurred only during the peak period in the eastbound direction, is the most probable cause of this result.

The final test, which explored the likelihood of trucks bearing down on cars, examined the proportion of instances in which trucks followed cars with small time gaps (1.5 and 2.0 sec). Only 2 of 24 tests were statistically significant, indicating that the greater concentration of trucks in the right lanes did not result in an increase in this event. This conclusion is based on the inconsistency of the results associated with the two time-gap values and the possibility of spurious results associated with multiple testing.

This study attempted to quantify measures that could be interpreted as indicative of operational performance. A left-lane truck restriction was then implemented and evaluated on the basis of examination of these measures. No definitive results were found that could be attributed to the implementation of the truck restriction. Although the study design was sound, the lack of an appropriate control site limited the interpretation of many of the results. Furthermore, the presence of a horizontal grade and significant volume changes over the course of the study also complicated the interpretation. Finally, on the basis of the characteristics of the reported study site, the application of these results should be limited to low-volume facilities.

Implementation of truck restrictions, such as the lane restrictions studied, theoretically has the potential to improve the capacity and safety of the roadways. The lack of evidence from the current research to support strong conclusions in these areas has two important implications. The first is the requirement that the results reported here be applied only to similar roadways, that is, rural, low-volume facilities with relatively little truck traffic. The second is the need for further research to investigate whether these results can be extended to higher volume roadways or roadways with larger truck percentages.

RECOMMENDATIONS

Because no discernible negative effects on highway operations could be attributed to the truck restriction, the restriction should be left in place. After a 2-year period, the feasibility of performing an accident analysis should be examined. Such a study would investigate whether the restriction caused an increase in accidents. In addition, more research should be performed on the differential design of pavements on six-lane highways.

ACKNOWLEDGMENT

This study was funded by the Texas State Department of Highways and Public Transportation.

REFERENCES

1. M. C. Zavoina, T. Urbanik II, and W. Hinshaw. *An Operational Evaluation of Truck Restrictions on Six-Lane Rural Interstates in Texas*. Research Report 1152-1F. Texas Transportation Institute, Texas A&M University System, College Station, 1990.
2. *Effects of Lane Restrictions for Trucks*. FHWA, U.S. Department of Transportation, June 1986.
3. S. Sirisoponsilp and P. Schonfeld. *Impacts and Effectiveness of Truck Lane Restrictions*. Maryland Department of Transportation, Baltimore, Feb. 1988.
4. *An Evaluation of the I-95 Truck Restriction in Broward County: Executive Summary*. Traffic Operations Department, Florida Department of Transportation, Tallahassee, Nov. 1982.
5. N. J. Garber and R. Gadiraju. *The Effect of Truck Traffic Control Strategies on Traffic Flow and Safety on Multilane Highways*. School of Engineering and Applied Science, University of Virginia, Charlottesville, Sept. 1989.
6. F. R. Hanscom. *Operational Effectiveness of Three Truck Lane Restrictions*. 1989.

The contents of this report reflect the views of the authors, who are responsible for the opinions, findings, and conclusions presented herein. The contents do not necessarily reflect the official views or policies of the Texas State Department of Highways and Public Transportation. This report does not constitute a standard, specification, or regulation.

Publication of this paper sponsored by Committee on Freeway Operations.

Development of Appropriate Design-Hour Volumes for Urban Freeways in Large Texas Cities

Carol H. Walters and Christopher M. Poe

In the planning and design process, 24-hr volumes in the design year are converted to design hour volumes (DHVs) by using *K* factors [ratio of 30th-highest-hour two-way volume to two-way average annual daily traffic (ADT)], *D* factors (directional split during the 30th-highest hour), and *T* factors (percentage heavy vehicles in the DHV). Lack of precision in selection of these parameters can make a ±50 percent difference in the design of urban freeways. The procedures used to select these factors in Texas urban areas have been examined, and refinements to the process are proposed. Using data from automatic traffic recorder (ATR) stations for 1973 to 1988 in the Texas cities of Houston, Dallas, Fort Worth, Austin, and San Antonio, along with manual classification data for 1984 to 1988, *K, D*, and *T* factors have been studied under increasingly urban conditions. Emphasis has been placed on the *K* and *D* factor research. Variables identified and tested for significance in predicting DHVs included facility type (radial or circumferential freeway), weekday ADT per lane, degree of capacity utilization during the peak period, employment density near the ATR station, length of peak period, and distance from the central business district. Although the initial data sets have yielded an unstable model over time using a multivariable regression analysis, ranges of K_D (*K* factor by direction) for ranges of each variable are developed for use in determining reasonableness of preselected *K* and *D* values.

Determining the number of lanes required for an urban freeway (new or expanded) involves estimating how much traffic will use the facility during the design hour. Currently, the methodology involves estimating the 24-hr volume that is expected to use the facility in the design year and multiplying that volume by planning parameters that will calculate the amount of traffic expected by direction in the design hour. The design-hour volume (DHV) influences the size of the facility, the amount of right-of-way needed (which may or may not be feasible to acquire), and the overall cost of the project. In rural areas where traffic volumes are low and right-of-way costs are also relatively low, the precision of the planning process is not as critical; however, in urban areas where traffic volumes are high and costs for right-of-way along freeway corridors continue to escalate, any imprecision in the planning process can be very costly. In extreme cases, a high estimation on required size of a facility may introduce such insurmountably high costs that it may preclude approval of any improvement at all, if a benefit-cost approach is used to determine priorities among projects. In short, decisions concerning which facilities can and cannot be justified as cost-effective, as well as the general mobility of a region, hinge on estimation of the DHV.

Extensive research has been conducted on the development of travel demand forecast models, as a means of estimating 24-hr volume, and on design, construction, maintenance, and even operation of the nation's highways. However, little research has focused on the development of planning parameters, such as the *K* factor [ratio of two-way 30th-highest hour to two-way annual average daily traffic (ADT)], directional split in the design hour (*D*), and percentage of trucks (*T*) present in the DHV. Moreover, the research that has previously been conducted on *K* factors has concentrated on less urbanized roadways with low ADT. This research is unique in that it focuses on urban freeways with ADT as high as 240,000 vehicles per day (vpd).

The objective has been to investigate whether improvements can be made in the estimation of planning parameters for urban areas that can be incorporated into the planning process. Specifically, this research has identified issues and concerns in the process as it is generally employed and has drawn on a large data base to test whether the various parameters can be reliably predicted if certain future conditions have been forecast. Statistical relationships and trend line analyses for *K* factors, directional splits, and peak-hour truck percentages have been studied in relation to facility type, ADT, degree of congestion, employment density, and location within the urban area. Emphasis is placed on the results of the research on *K* and *D*.

ISSUES AND CONCERNS

The following are some general concerns about the current planning procedure that merit discussion. These concerns were developed in concert with planners and designers from the Texas State Department of Highways and Public Transportation (SDHPT).

Design Sensitivity

The magnitude of potential over- or underdesign caused by multiple estimations of the planning parameters is an issue of concern. In the current planning process, an estimation is made of each of the planning parameters. These parameters are used in an equation of the following general form to estimate the DHV and, then, the required number of lanes (*N*):

Texas Transportation Institute, Texas A&M University System, 1600 E. Lamar Boulevard, Suite 120, Arlington, Tex. 76011.

$$N \geqslant \text{ADT} * [K * D * (1 + T)]/[\text{service volume}] \qquad (1)$$

Service volume assumes level terrain (≤2 percent grade) for simplicity. There is a range of plausible values that can be chosen (on the basis of existing conditions on similar facilities) for each of these parameters and, thus, a potential error associated with each estimation. When these parameters are used together in the equation for estimating the number of lanes, the corresponding range of potential error is also multiplied. Table 1 presents each of the planning parameters, the usual accepted ranges for each of these values, and the variation from the mean. Typically, as an area develops from rural to urbanized, *K, D*, and *T* factors would all be expected to decrease, whereas the expected service volume would probably increase. Thus, all the errors would compound, rather than cancel one another, if the urbanization trend is under- or overestimated. Hence, even if the ADT projections were exact, the magnitude of the error in design could be more than 50 percent.

A simple, if exaggerated, example may clarify the point: a freeway expected to carry an ADT of 120,000 vpd in the design year is estimated to have a *K* factor of 0.12 and a *D* of 70 percent, on the basis of existing data on a nearby freeway; the nearest truck count exhibits an 8 percent truck factor. These factors are then furnished by the planner to the designer for use in establishing the DHV. On the basis of a desired level of service (LOS) of *D* or better, a maximum service flow rate of 1,700 passenger cars/hr/lane (pcphpl) is selected by the designer. The freeway would need to be 7 lanes in each direction, or 14 lanes. Another planner might see an urbanization trend in the population and employment forecast and select a *K* factor of 0.09, a *D* factor of 60 percent, and a *T* factor of 3 percent. A service volume of 1,900 pcphpl might be selected by the designer (traffic counts on freeways in urban areas in excess of 2,200 vehicles per hour per lane are becoming increasingly common). This calculation would mean that a freeway of four lanes in each direction, or an eight-lane freeway, would probably be sufficient. In many cases, those selecting the parameters for use in calculating the DHV are not involved in design and are not fully aware of the tremendous impact these factors have on facility design.

Need for Site-Specific Values

A single *K* and *D* value is normally used for planning and design of a freeway, which may be many miles in length with highly variable usage patterns. The design of ramps and merging, weaving, and diverging areas are all dependent on DHVs computed using these factors. However, data obtained on existing facilities do not support the conclusion that main lanes and their adjacent ramps exhibit similar peaking patterns. Figures 1 and 2 show peaking characteristics on the main lanes and adjacent ramp pairs for a radial freeway through an area of high employment in Dallas. A ramp pair is defined as two ramps that are expected to have balanced 24-hr volumes because trips made on one ramp in one direction have a return trip on the other ramp in the return direction. Ratios of peak-hour volume to daily volume on the ramp pairs vary ±35 percent from the *K* factor on the main lanes, averaging 10 percent higher, and the directional splits on ramp pairs in the corridor vastly exceed that of the freeway main lanes.

TABLE 1 EFFECT ON DESIGN BECAUSE OF VARIATION IN PARAMETERS

Parameter	Range of Values	Variation from Mean	Effect on Facility Size
K	.08 - .12	± 20 %	± 20 %
D	.50 - .70	±17 %	± 17 %
T	.02 - .10	±67 %	± 4 %
Service Volume	1700 - 2000 pcphpl	± 8 %	± 8 %
N, Number of Lanes			± 58 %

Need for Separate Peak Periods

Each element of a freeway is designed for its own DHV; however, these design hours may not occur at the same time. For example, the main lanes of a freeway, an entrance ramp, and an exit ramp all must be designed for their DHV. However, although the freeway and one of the ramps may peak during one peak period, the other ramp may peak during the other peak period. Designing each element for the same peak period may cause overdesign of some of the elements, particularly for weaving sections. The need for local data on peak-hour travel patterns and behavior of each of the freeways is apparent.

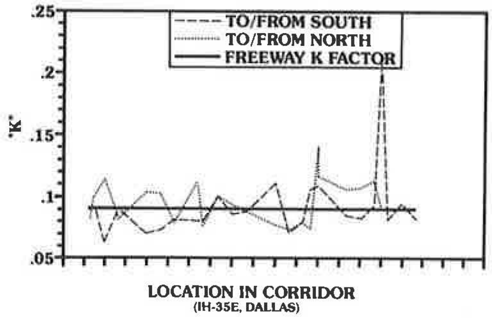


FIGURE 1 Comparison of freeway *K* factor with *K* factor on adjacent ramp pairs.

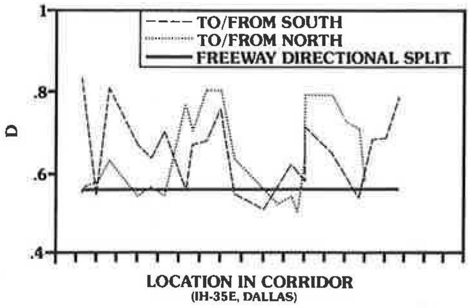


FIGURE 2 Comparison of freeway *D* factor with *D* factor on adjacent ramp pairs.

System Effects

System effects—freeway sections constrained from capacity improvement—are virtually neglected in this planning process. Use of a high K factor for a facility integral with others operating at a low K factor is unrealistic, unless there is an identifiable way to load or unload the extra peak-hour capacity thus provided on the new or improved facility. Peak-hour volumes will likely be either metered or constrained by upstream or downstream congestion. These volumes are capped by system constraints, whereas 24-hr volumes have ample capacity to grow, causing a lower K factor coupled with unused capacity even during the peak.

Weekday Traffic Versus Daily Traffic

There is a need for consistency in the definition and use of planning parameters. K factors are normally estimated from automatic traffic recorders (ATRs), which allow continuous data collection over a year. ADT, the sum of a year's worth of data divided by the number of days in that year, includes holidays and weekends and is thus lower than an average weekday calculation (AWDT). K factors are calculated as the ratio between the 30th-highest hour during the year and the ADT, both for two-way traffic. The problem comes in the application of these K factors to arrive at the DHV. Typically, design year ADT has been estimated in urban areas through use of a travel demand model, which simulates typical weekday travel. Use of the standard K factor with these ADT values introduces a potential overdesign in the vicinity of 10 percent.

Directional Design Hour

The calculation of DHV on the basis of two-way volume multiplied by the directional split that happens to be associated with the 30th-highest hour may not accurately predict the directional 30th-highest hour. In some instances, the one-way volume during the 30th-highest hour is lower than that occurring during the 200th hour. Use of a directional K factor (K_D, the ratio of the 30th-highest one-way hour to the one-way ADT) might offer a simple solution to this problem. It would also eliminate one planning parameter and, thus, one more opportunity for accumulated error.

Effects of Time

K and D factors both tend to decline over time. Use of current data to predict future K factors for design may fail to incorporate the predictable effects of urbanization, which can include significantly increased volumes of off-peak trips and off-peak direction trips generated by new commercial land uses. Low K factors generally are associated with high congestion; however, some of the data exhibited low K factors because of multiple land uses with trips spread out throughout the entire day and relatively low congestion levels. The selection process for these parameters should involve consideration of these trends.

Peak-Hour Truck Percentages

The use of truck percentages to adjust DHV to equivalent passenger-car volumes for design can have a major effect on the sizing and design of facilities, but appropriate peak-hour, peak-direction vehicle classification data are scarce. Typically, 24-hr data from a limited number of manual classification sites in mostly rural or urban fringe areas are used, applying standardized adjustment factors (that are based on facility type and originally derived from these sites) to adjust for peak-hour conditions. Given the relatively small number of truly urban sites, the potential exists for the percentage of trucks in the design hour to be grossly overestimated for routes with typical urban commuting traffic. On routes with high volumes of commuting traffic, the volume of trucks tends to remain fairly constant throughout the day. However, automobiles exhibit heavy peaking characteristics, overwhelming the trucks during the peak hour in the peak direction, which naturally lowers the percentage of trucks.

METHODOLOGY FOR ADDRESSING CONCERNS

Specific recommendations were developed to minimize the impacts of each factor of uncertainty in the planning process. The following paragraphs describe the methodology for addressing the eight concerns identified with the current process for developing DHV:

1. Design sensitivity,
2. Need for site-specific values,
3. Need for separate peak periods,
4. System effects,
5. Weekday traffic versus daily traffic,
6. Directional design hour,
7. Effects of time, and
8. Peak-hour truck percentages.

The first four issues indicate a critical need for planning parameters to be selected with a high degree of attention given to local travel patterns. Although this emphasis introduces additional time and cost into a planning process that is generally underfunded and overcommitted, the potential savings are great, both in terms of avoiding overdesign (which results in unusable capacity) and avoiding underdesign (which results in congestion that could have been avoided). It is recommended that the planning staff, following the development of ADT values for design purposes, work closely with those involved in design to ensure that realistic parameters are assigned to each section of roadway, including special attention to one-way flows on ramps and direct connections. Collection of representative peak-period data is imperative when there are multiple constraints to pursuing the ultimate design.

The fifth issue is relatively easy to address. A K_w factor (based on weekdays only, excluding holidays) can be easily calculated from ATR data. It is this factor that should be applied to the 24-hr volumes resulting from travel demand models. Calculations of ADT for other purposes, such as pavement design and environmental impact, should proceed on the basis of appropriately reduced volumes from the travel demand models.

The sixth issue, like the fifth, is easily addressed through ATR data. As mentioned, the data reduction program could be modified to seek the 30th-highest volume in each direction, and a directional K factor could be calculated as the ratio between that hour and the AWDT in that direction. This calculation would yield a K_D factor (which would also include the K_w factor), that could be applied directly to the one-way volumes from travel demand models, as adjusted by the transportation planning department, to calculate the DHV.

The seventh and eighth issues involve careful study of the elements involved in the effect of urbanization on the planning parameters. The remaining paragraphs deal with the analysis of ATR data over time on various types of facilities under differing levels of urbanization in the major cities of Texas. In the interest of brevity, the subject of trucks is treated elsewhere. However, results indicated that, in general and except for special identifiable cases, a peak-hour directional truck percentage on radial freeways in highly urbanized areas in Texas rarely exceeds 4 percent, with that on circumferential freeways being slightly higher. For example, the average in the Dallas–Fort Worth metroplex is 2.7 percent for radial freeways and 3.1 percent for circumferential freeways.

ANALYSIS OF *K* AND *D* FACTORS OVER TIME

The methodology for estimating planning parameters for urban freeways used several years of data from ATR stations to examine K and D factors. In order to simplify the analysis and gain more precision, a combined K and D factor was used that also incorporated AWDT as its basis. As discussed, this factor offers a potential benefit to the planning process in that it reduces the number of estimations needed to calculate the DHV and makes the resulting K factor usable with travel demand models. Therefore, K_D is defined as the 30th-highest directional hour divided by the AWDT for that direction, and it is this parameter with which the remaining analysis deals. Variables that influence the value of K_D, along with general trends, were identified, and development was begun using regression analysis on a statistical model to predict K_D under the influence of these variables. Finally, reference tables were developed to aid the transportation engineer in choosing appropriate values of K_D for planning freeway improvements in urban areas.

ATR Data for Urban Stations

The Texas SDHPT has an extensive ATR system throughout the state. The system consists of induction loop detectors placed in the lanes of the highways that continually collect traffic volumes for every hour of the year. The Texas Transportation Institute (TTI) obtained tapes of raw data from the ATR stations for the past 15 years and compiled an initial list of 50 stations for study in Austin, Dallas, Fort Worth, Houston, and San Antonio. Some problems with this data base quickly became evident:

- ATR stations on radial freeways were often congested, obscuring K factors as a measure of demand because of truncated peaks, as shown in Figure 3.

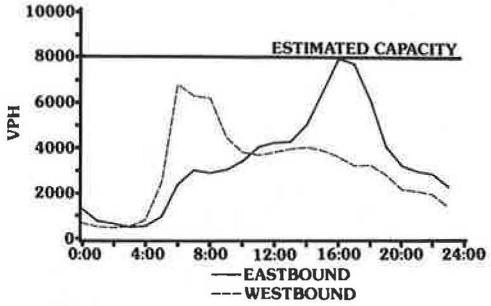


FIGURE 3 Truncated peaks not reflective of demand: IH-30 east, Dallas.

- Clock-hour data may obscure the true peaks, causing peaks to appear truncated when they are not.
- Many ATR stations are between heavy exit and entrance ramps, resulting in apparently below-capacity flow rates, even on facilities known to be operating at or near capacity.
- Many of the 50 stations in the urban areas carried less than 5,000 ADT per lane and were therefore discarded as not reflecting urban traffic characteristics.

Recognizing these limitations, the ATR data were further analyzed to identify likely descriptors of urbanization acting over time to influence the values of K_D. The variables identified and their individual relationships with K_D are described in the following subsection.

Variables Used in Estimating K_D

Many variables were evaluated on a preliminary basis for their influence on K_D. Variables were selected for evaluation on the basis of their potential as descriptors of the differences in K_D under varying urban conditions. It was essential to select variables for which data readily exist and for which data can be estimated for the future. Preliminary analyses of each variable were conducted, and the variable list was refined to include six that showed promise: facility type (radial or circumferential), AWDT per lane, distance from the central business district (CBD) of each city, utilization index (measure of peak-hour capacity usage), peak-period ratio (measure of peak-hour volume to 3-hr peak period), and employment density in the zones adjacent to the ATR station.

Facility Type

Facility type refers to radial or circumferential freeways. The ranges of K_D for radial and circumferential freeways were different enough to suggest that the data should be separated to examine K_D by facility type. In general, circumferential facilities have lower K_D values than radial facilities that have similar characteristics. This difference can be attributed to a diversity of trips throughout the day and two-way peaking during each peak period on many circumferential facilities. Radial facilities, unless constrained or serving multiple types of land use, often serve high one-way commuting patterns.

AWDT per Lane

The use of AWDT alone to estimate K_D was not very significant. However, AWDT per lane as a measure of use on a 24-hr basis had significant correspondence to K_D, as indicated in Figures 4 and 5. As AWDT per lane increases, K_D decreases, partly because the peak hour may reach capacity and cause some peak spreading to occur and partly because high-activity areas remain busy all day, diluting the effect of purely commuting traffic. Another factor may be peak-hour volumes that are constrained or metered by connections with other facilities. A number of low K_D stations with low AWDT per lane were found as well in areas where commuting trips are simply not a dominant pattern. This variable is not the most desirable for predicting K_D because an estimate must be made of the number of lanes planned for a facility before an estimation of K_D can be made. This estimation is apt to be a self-fulfilling prophecy, because K_D controls the number of lanes designed for a given per-lane capacity.

Distance from CBD

The distance from the CBD was another variable with significance in estimating K_D. Even though the major urban areas vary in size and development intensity, the majority of the ATR stations are within a 5-mi radius of the CBD. The correspondence between distance from the CBD and K_D is shown in Figures 6 and 7. K_D increases the farther the stations are from the CBD. The relationship between distance and K_D is, as expected, less significant for circumferential facilities than for radial facilities. This difference is due to the dependence on downtown commuting patterns that is prevalent on radial facilities and missing on circumferential facilities. Further, circumferential routes serve a cross-town pattern of trips that are relatively independent of the distance from the CBD. To account for the variety in sizes of the cities studied, the distance variable for each city was normalized by dividing it by the square root of the area of the city.

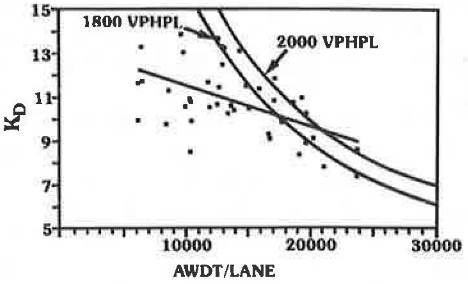


FIGURE 4 K_D factor versus AWDT per lane on radial freeways in urban areas.

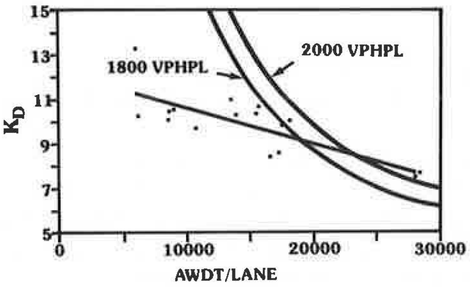


FIGURE 5 K_D factor versus AWDT per lane on circumferential freeways in urban areas.

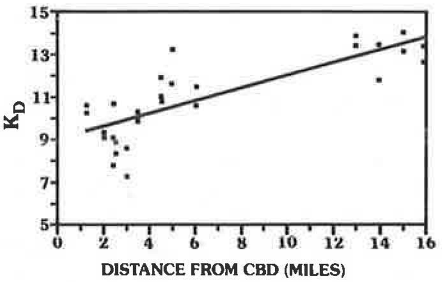


FIGURE 6 K_D factor versus distance from CBD on radial freeways in urban areas.

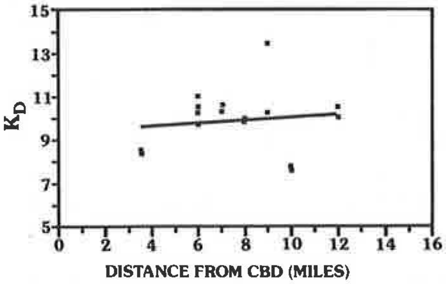


FIGURE 7 K_D factor versus distance from CBD on circumferential freeways in urban areas.

Utilization Index

The utilization index is a measure of the peak-hour usage or a rough volume-to-capacity ratio in the peak hour. For comparison, a capacity of 1,800 vehicles per hour per lane (vphpl) was assumed, and the volume was the 30th-highest hour of the year for a station by direction. A high utilization index suggests a congested peak hour; therefore, as the utilization index increases, K_D should decrease due to trips being made outside the peak hour. The relationship of K_D and utilization index for radial facilities is shown in Figure 8. As can be seen, there is little correlation between utilization index and K_D; the expected negative slope is barely discernible, and the data are highly scattered. Examination of the data reveals a number of stations with high peak-hour usage and thus a high utilization index (probably by commuter traffic), but there is little use any other time of the day, resulting in high values of K_D. At the same time, there are some congested stations that have constrained volumes in the peak hour, and thus a low utilization index, and high volumes for many other hours of the day, resulting in low values of K_D. Figure 9 shows the data for circumferential facilities, as well as a slightly better correlation.

As for using utilization index for future planning, knowing a value for the utilization index would be essentially the same

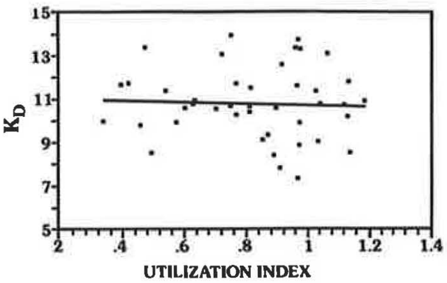


FIGURE 8 K_D **factor versus utilization index on radial freeways in urban areas.**

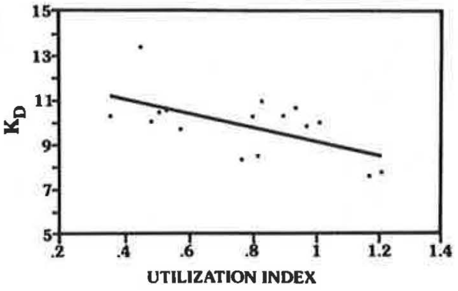


FIGURE 9 K_D **factor versus utilization index on circumferential freeways in urban areas.**

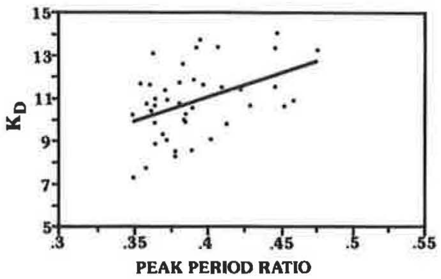


FIGURE 10 K_D **factor versus peak-period ratio on radial freeways in urban areas.**

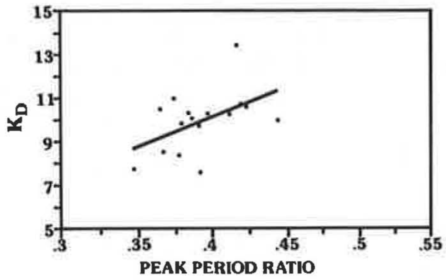


FIGURE 11 K_D **factor versus peak-period ratio on circumferential freeways in urban areas.**

as knowing the value of K_D; however, for planning purposes, a desired level of usage during the peak hour could be chosen and used as an input in estimating K_D. For example, choosing a utilization index of 0.9 might mean accepting an LOS of D in the peak hour. Overall, however, this variable is highly unpredictable and therefore not particularly useful, at least for radial freeways.

Peak-Period Ratio

The peak-period ratio, a rough measure of the length of the peak period, is calculated by dividing the highest hourly volume in the 3-hr peak period by the sum of the total volume in that peak period. A congested station with even use during all 3 hr of the peak period could have a peak-period ratio of 0.333; conversely, a station with traffic only during the peak hour and little traffic in the shoulder hours of the peak period would have a peak-period ratio approaching 1.0.

As the peak-period ratio decreases, the value of K_D also decreases, reflecting congestion and consequent travel increases outside the peak. Figures 10 and 11 show the correlation for radial and circumferential facilities, respectively. Although the data scatter is high for radial facilities, a pattern is discernible; for circumferential facilities, the correlation is weak or nonexistent.

Employment Density

Employment density is the measure of employment around the permanent count stations. The employment assumptions used in the traffic assignment models by the Texas SDHPT were incorporated into a variable. A block 2 mi by 1 mi, with the permanent count station at the center, was identified for each station, and the employment figures for all serial zones within the block were divided by the actual total area of the serial zones. Types of employment were not disaggregated.

As employment density increases, the value of K_D decreases. The increase in employment generates more trips in off-peak hours, which in turn results in lower values of K_D. The relationship of employment density and K_D is shown in Figures 12 and 13. Greater correspondence was found for circumferential facilities than for radial facilities.

Results of Statistical Analysis

In general, the results from the statistical analysis proved disappointing. Although some fairly strong correlations were found on an individual basis, multivariable regression analysis

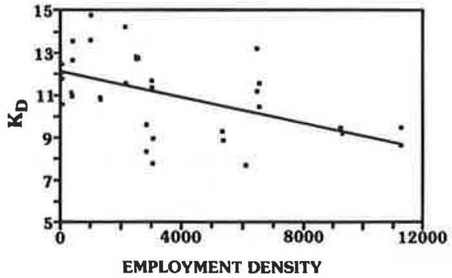


FIGURE 12 K_D **factor versus employment density on radial freeways in urban areas.**

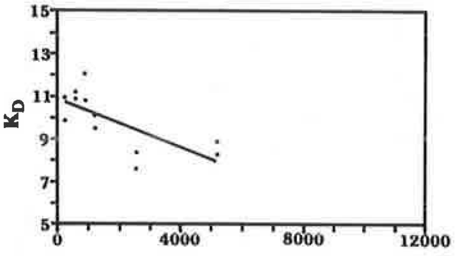


FIGURE 13 K_D factor versus employment density on circumferential freeways in urban areas.

failed to achieve a better correlation than $R^2 = 0.70$, unless AWDT per lane and the utilization index were used together, which violates statistical procedure because together they define K_D. Also, the models proved unstable from year to year.

The problem lies in the data base, which is extensive but problematical. After eliminating stations not reflective of demand (because of congestion) or not fully urban, too few stations remained to form a significant enough data set to establish a reliable predictive relationship for the variables tested. However, the methodology developed for analysis exhibits promise if a larger, usable data base were available. In recent years, several urban districts in Texas have been installing permanent traffic recording equipment for future surveillance systems. Therefore, advantage may be taken of the increase in permanent counting capabilities throughout urban areas to expand the data base, and it is hoped this type of regression analysis may eventually allow an estimate of K_D to be calculated directly.

Development of Reference Tables

The purpose of this research has been to develop a useful tool to aid in the selection of planning parameters for urban freeways. Because the models developed from the data base did not lead to estimations of K_D that were statistically reliable, a reference table was developed for use as a guide in determining reasonable ranges of K_D under various types of conditions. Before establishing this table, the data were further qualified by eliminating data constrained by peak-hour bottlenecks, stations with less than 10,000 vehicles per day per lane, stations farther than 15 mi from the CBD, and stations with less than 70 percent capacity utilization during the peak hour. This step eliminated the extreme variability found in stations not experiencing conditions reflective of appropriate design hours in urban areas. In effect, constrained stations fail to reflect demand, and underutilized facilities have simply not matured.

The ranges of K_D found for each of the variables are presented in Table 2, separated into radial and circumferential facilities. In general, K_D is lower for circumferential than for radial facilities for equal values of the variables. K_D falls as AWDT per lane increases, falls as the peak-period ratio rises (indicating a lengthening of the peak), and falls again as em-

TABLE 2 RANGES OF K_D FACTOR FOR RADIAL AND CIRCUMFERENTIAL FACILITIES

RADIAL			CIRCUMFERENTIAL		
AWDT/LN Range	Radial K_D	Correlation	AWDT/LN Range	Circumferential K_D	Correlation
10,000-15,000	10.0-14.0		10,000-15,000	10.0-11.0	
15,000-20,000	9.0-12.0	High	15,000-20,000	8.0-10.0	High
20,000 +	7.0-9.0		20,000 +	7.0- 8.0	
Employment Density	Radial K_D	Correlation	Employment Density	Circumferential K_D	Correlation
0-5,000	Highly Variable		0 - 5,000	9.0-11.0	
5,000-10,000	9.0-11.0	Medium	5,000-10,000	7.0- 9.0	High
10,000 +	7.0-9.0		10,000 +	No Data	
Distance	Radial K_D	Correlation	Utilization Index	Circumferential K_D	Correlation
0-3 miles	7.0-11.0		0.7-0.8	10.0-12.0	
3-5 miles	10.0-12.0	Medium	0.8-1.0	9.0-11.0	Medium
5 + miles	12.0-14.0		1.0 +	7.0- 9.0	
Peak Period Ratio	Radial K_D	Correlation	Peak Period Ratio	Circumferential K_D	Correlation
.33-.42	7.0-12.0	Medium	.33-.42	7.0-11.0	Medium
.42-.48	11.0-14.0		.42-.48	11.0-12.0	
Utilization Index	Radial K_D	Correlation	Distance	Circumferential K_D	Correlation
.7-0.8	Highly Variable	Low	0-3 miles	Highly Variable	Low
0.8-1.0			3-5 miles		
1. +			5 + miles		

ployment density increases. For radial facilities, K_D rises as the distance from the CBD increases; for circumferential facilities, it falls as the peak-hour utilization rises.

Table 2 allows a check on the reasonableness of K and D factors developed using current procedures. First, K_D must be calculated. Ideally, the data collected at the ATR stations could be reduced to provide K_D directly—the 30th-highest directional volume divided by the AWDT in the corresponding direction. However, if more precise data are not available, it can be approximated by use of the following formula:

$$K_D = 1.84 * K * D \quad (2)$$

This formula can be derived as follows:

$$\begin{aligned} K_D = {} & \text{30th-highest-hour volume (one-way)} \\ & \div \text{AWDT (one-way)} \end{aligned} \quad (3)$$

which can be approximated as

$$\begin{aligned} K_D = {} & [\text{30th-highest-hour volume (two-way)} * D] \\ & \div [\text{AWDT (two-way)/2}] \end{aligned} \quad (4)$$

$$\begin{aligned} K_D &= \text{ADT} * K * D/(\text{AWDT}/2) \\ &= 2 * K * D * (\text{ADT/AWDT}) \end{aligned} \quad (5)$$

The ratio ADT/AWDT has been found to vary by ATR station, but a reasonable value from the data is 0.92. Thus,

$$K_D = 2 * K * D * 0.92 = 1.84 * K * D \quad (6)$$

Table 2 can be used to check whether the calculated K_D falls into the ranges for the various variables observed from this data set. The ranges of K_D are divided into radial and circumferential facilities; for each facility type, the variables are listed in order of their correlation with K_D. Each variable is described by whether there is high, medium, or low correlation between that variable and K_D. If a calculated K_D falls outside the range of a highly correlated variable, there is evidence from the data set that the K_D should be adjusted. If a calculated K_D falls outside the range of a medium- or low-correlation variable, there may still be reason to consider a change in the calculated K_D.

CONCLUSIONS AND RECOMMENDATIONS

This research has focused on the process of selection of planning parameters used in the calculation of DHV and on determining the predictability of these parameters under measurable urbanization trends. The importance of precision has been established, in that design may be significantly affected by the assumptions implicit in the choice of the parameters. The major recommendation made from these findings is that those responsible for the selection of planning parameters should work closely with those responsible for design. To the greatest extent possible, local data on peak-hour traffic patterns should be provided to the planners to be incorporated into the selection of these parameters, and the variability along the corridor should also be accounted for, especially where ramps and direct connections are concerned.

The parameters K and D have been reduced to one parameter, K_D, to provide greater precision in the analyses, with K_D representing the 30th-highest directional volume as a percent of the average weekday directional volume (AWDT_D). It is recommended that this single parameter be provided for design purposes because 24-hr volumes from travel demand assignment models reflect weekday forecasts by direction, and thus the conversion to DHV can be made directly. Annual summaries of data from the ATR stations could also be revised to include the calculation of two new parameters: (a) an adjusted K factor that incorporates AWDT for weekdays only and (b) a K_D factor, defined as the 30th-highest directional hour divided by the AWDT for that direction.

The variables tested for significance included facility type (radial or circumferential), AWDT per lane, the degree of peak-hour utilization of the facility, the relative number of hours in the peak period, the relative distance from the CBD, and the adjacent employment density. Unfortunately, too little data were usable, which prevented development of a statistically reliable predictive model; however, as more permanent count stations come on line, this potential should be reevaluated. Increasing efforts toward developing freeway surveillance, communication, and control systems will result in many more usable data locations where 24-hr as well as peak-hour freeway volumes can be monitored by 15-min increments. These data will result in a wealth of opportunity to further address the predictability of K_D under varying circumstances.

Because reliable predictive models are not yet available, a set of ranges has been developed for K_D under a range of conditions for each variable, and these ranges are recommended for use as guidelines in checking the reasonableness of estimated K_D factors.

ACKNOWLEDGMENT

This research was sponsored by the Texas State Department of Highways and Public Transportation, in cooperation with FHWA, U.S. Department of Transportation.

Publication of this paper sponsored by Committee on Freeway Operations.

Real-Time Knowledge-Based Integration of Freeway Surveillance Data

NEIL A. PROSSER AND STEPHEN G. RITCHIE

An advanced processing capability based on the use of real-time knowledge-based expert system (KBES) technology to integrate diverse types of traffic surveillance data for freeway monitoring and control purposes, particularly as part of future "smart roads" projects, is described. One of the major functions of the prototype system is the acquisition and processing of input data drawn from sensors and processes in the real world. The real-time nature of these processes, and the associated need for decision-making information and recommendations, places particular importance on the efficient handling of data to avoid unnecessary overloading of the expert system. The relevant types of data include traffic occupancies and volumes from loop detectors in the pavement, information on traffic conditions from closed-circuit television cameras, field reports from police officers and other official personnel, and cellular and emergency telephone calls from motorists. Emphasis is placed on the way in which the data are acquired, processed, and integrated within a prototype KBES framework to achieve the objectives of the incident management tasks, specifically those of incident detection and verification. Examples are given of the implementation of these features.

As part of an ongoing research effort to investigate application of artificial intelligence techniques for advanced traffic management, a prototype real-time knowledge-based expert system (KBES) has been developed for managing nonrecurring congestion on urban freeways. The objective of this system, called "freeway real-time expert system demonstration", (FRED) is to provide decision support in a future traffic operations center (TOC) to traffic control room operators responsible for the monitoring and control of so-called "smart roads" or intelligent vehicle-highway systems (*1*). The FRED system is described in a companion paper by Ritchie and Prosser in this Record. The only comparable expert system currently known is one being developed by INRETS, the French transportation agency, for monitoring and control of arterial street intersections (*2*). The G2 expert system shell used in the INRETS project is the same as the one used in the FRED prototype system.

More specifically, FRED provides decision support to TOC operators in the field of incident management. Incidents are events that lead to nonrecurrent congestion. Examples of such events are accidents, spilled loads, stalled or disabled vehicles, temporary maintenance and construction activities, signal and detector malfunctions, and special events. Incident management involves detecting and responding to incidents in order to alleviate the resultant delay to motorists. In future smart road TOCs, it will become increasingly difficult, if not impossible, for human operators to function effectively without automated assistance because of the increased amount of incoming TOC data and the complexity both of the networks and of incident management and response functions.

The monitoring aspect of FRED involves synthesizing input data from a number of sources. The freeway system is monitored, in the first instance, to detect when and where an incident has occurred. Once an incident has been detected, more information is required about the state of the system to formulate appropriate responses. One of the major functions of the system is the acquisition and processing of data drawn from sensors and processes in the real world. The real-time nature of these processes, and the associated need for decision-making information and recommendations, place particular importance on the efficient handling of data to avoid unnecessary overloading of the expert system.

Focusing on the initial incident detection and verification phases, which represent the foundation of incident management, a description is provided of an advanced processing capability based on the use of real-time KBES technology to integrate diverse types of traffic surveillance data for freeway monitoring and control purposes. The relevant types of data include traffic occupancies and volumes from loop detectors in the pavement, information on traffic conditions from closed-circuit television (CCTV) cameras, field reports from police officers and other official personnel, and in-vehicle cellular and roadside call-box telephone calls from motorists. Emphasis is placed on the way in which the data are acquired, processed, and integrated to achieve the objectives of the incident management tasks, specifically those of incident detection and verification. Examples are given of the implementation of these features in FRED.

SYSTEM OVERVIEW

Figure 1 shows the different stages of incident management (*3*). In the companion paper in this Record, the development of FRED—a component prototype real-time expert system for managing nonrecurring congestion on urban freeways—is discussed, and the application of FRED to a section of the Riverside Freeway in southern California is presented as a case study to illustrate the current capabilities of the system. As mentioned, the focus here is on the way in which data are acquired and processed to achieve the objectives of the incident management tasks. The incident detection stage relies heavily on data acquisition and processing and incorporates most of the features of interest in the following sections.

Figure 2 shows the overall FRED system layout, which in terms of incident detection and verification is consistent with

Institute of Transportation Studies and Department of Civil Engineering, University of California, Irvine, Calif. 92717.

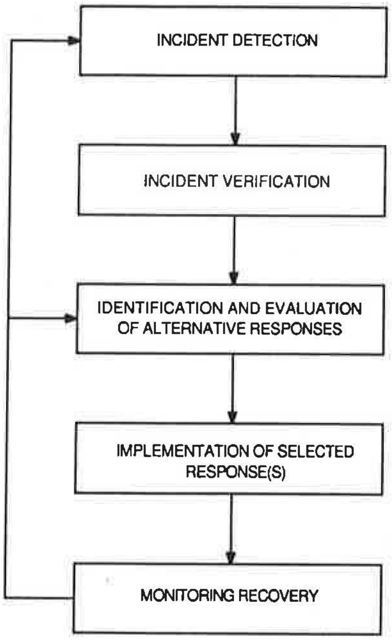

INCIDENT DETECTION
INCIDENT VERIFICATION
IDENTIFICATION AND EVALUATION OF ALTERNATIVE RESPONSES
IMPLEMENTATION OF SELECTED RESPONSE(S)
MONITORING RECOVERY

FIGURE 1 Incident management tasks.

the current capabilities of the Los Angeles freeway TOC (operated by the California Department of Transportation) and also with many other TOCs around the country. Of particular note are the external sources of data in the detection phase. Two streams of data, representing the detection and reporting of an incident, respectively, can be recognized. First, freeway loop detectors transfer 30-sec occupancy counts to an incident detection algorithm on a central computer, which in turn notifies FRED of the detection of an incident. Second, a number of agencies and sources may provide on-site accounts of an incident, and a communications center receives and filters their reports before sending high-priority ones to FRED. Thus, the two primary methods of incident detection are (a) an automated incident detection algorithm, and (b) outside reports.

The incident detection algorithm (*4*) operates on 30-sec occupancy counts (the percentage of time a loop is occupied by vehicles) obtained from sensors placed on the freeway main line at approximately 1-mi spacing. Each sensor site has a series of loop detectors, one per lane, arranged to form a counting station. Local controllers are responsible for compiling the counts for the detectors in each lane and then averaging them over the counting station. In the FRED prototype system, simulated occupancy counts are read from a data file. The incident detection algorithm examines the average occupancy counts for each station and, by comparing them to those for the station immediately downstream, can detect the presence of a significant disruption to traffic flow between the two stations. The sensitivity of the algorithm can be controlled by varying threshold parameters. Once an incident has been detected by the algorithm, a message is sent to FRED indicating the location of the counting station immediately upstream of the detected disruption.

The FRED system has been developed using a real-time expert system shell known as G2 (*5*), running under the Unix operating system and X-windows on a Sun SPARCstation 1 workstation. The so-called "California" incident detection algorithms (*4*) and the report processing algorithm were written as separate C programs capable of running on a different processor. Communication between the FRED expert system and the external C programs is via data files.

In the freeway control environment, an incident management system must operate in real time if it is to detect, verify, and formulate responses to incidents in a sufficiently short period of time for responses to be implemented effectively. For example, one type of response strategy is the provision of advance information to motorists approaching the site of an incident via changeable message signs or in-vehicle navigation systems. Such information, to be of benefit, must be provided soon after the occurrence of an incident. In this context, "real time" probably implies a maximum interval of

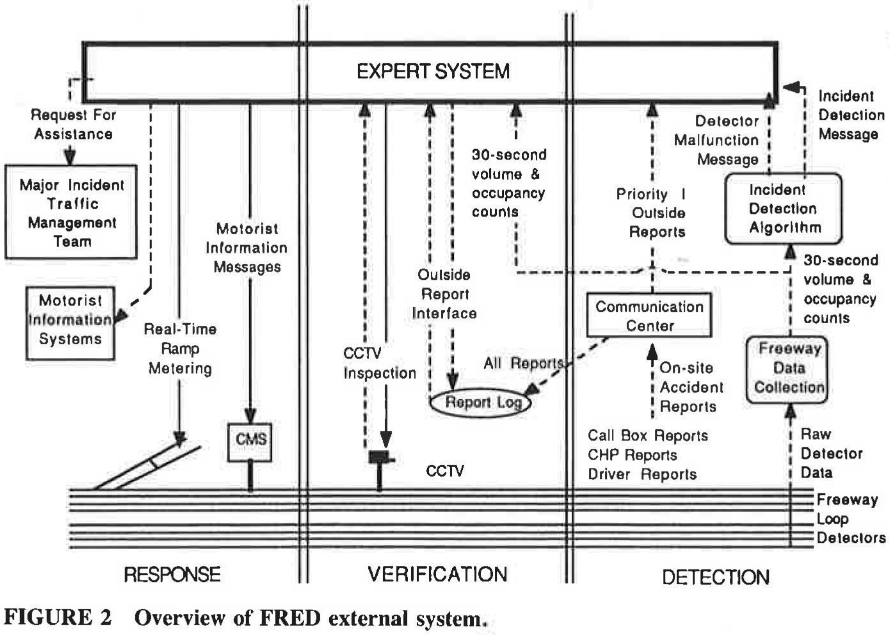


FIGURE 2 Overview of FRED external system.

several minutes between incident occurrence and response implementation. The real-time operating constraint introduces some important issues related to concurrent processes and data transfer, which are discussed in the next section.

DISTRIBUTION OF PROCESSING AND DATA TRANSFER

The workload on the TOC at any one time is directly proportional to the number of current incidents. Effective allocation of the expert system's resources is essential to maintain the speed of the overall system. The expert system should be devoted to those tasks for which it is best suited: the processing of high-level knowledge rather than simple algorithmic tasks. For FRED, the major high-level task is the formulation of responses to incidents. The detection of further incidents should proceed concurrent and external to the expert system. When another incident is detected, the expert system can be interrupted to include the new incident in its set of current incidents. Currently, FRED only handles one incident at a time, but the system structure was developed to facilitate the future treatment of multiple incidents.

Given that different tasks are allocated to separate and concurrent processes, the transfer of data between the processes becomes a major issue. An advantage of a knowledge-based approach is that the expert system has the ability to decide what data it requires at any particular stage and is able to send out requests for it. In this way, the system can select the information it needs for a specific task from the vast amount of data available. The decision of whether to transfer data from an external routine to the expert system can also be made by the external routine itself. In this case, the expert system is waiting to be interrupted by the occurrence of an event.

These aspects of the operation of a real-time expert system are discussed in terms of the implementation of FRED. Before doing so, the manner in which data and knowledge are represented in the system is outlined.

DATA AND KNOWLEDGE REPRESENTATION IN FRED

As explained, the FRED system was developed using a real-time expert system shell called G2 (*5*), which combines a knowledge base with sophisticated data interfacing and screen management capabilities. G2, like most sophisticated expert systems, is hybrid in nature, in that knowledge and data are present separately. The manner in which the knowledge is applied to the data determines the behavior of the overall system.

Data Representation

All data in FRED are represented as a series of objects. Each object contains a number of predetermined attributes that represent the data relevant to that object. Objects are all instances of a particular class, and the class definition specifies the attributes of all objects within the class.

The attributes of objects can receive values from a number of sources. The data can be inferred from rules by the inference engine, determined by an external routine, or simulated by the G2 simulator incorporated into the G2 system. Additionally, the system operator can enter values for data interactively.

The two most important object classes in the incident detection and verification phases are the counting stations and incident classes. Tables 1 and 2 present the attributes of each of these object classes along with their type and source. Variable attributes are either symbolic or quantitative. The counting station attributes either are constant or, for volumes and occupancies, receive their values from an external routine. Incident attributes either are derived from rules and formulas as part of the inference process in the expert system or are explicitly set by the operator or an outside report. Other attributes within these objects, of less importance, are not shown.

An important feature of G2 is the ability to create and delete objects during the running of the system. Such transient objects are particularly suitable for incident objects. When an incident is detected, an instance of the incident object class is created and its attributes set to the parameters of that incident. Once the incident has terminated, the object can be deleted. Thus, at any one time, the number of incident objects equals the number of current incidents. Counting stations, on the other hand, are examples of permanent objects because they represent a fixed element of the freeway system.

Knowledge Representation

Knowledge is encoded into the FRED system as a set of rules forming a knowledge base. These rules consist of antecedent and consequent parts. The antecedents specify conditions that

TABLE 1 INCIDENT OBJECT ATTRIBUTES

ATTRIBUTE NAME	TYPE	SOURCE
Status	Symbol	Inference Process
Upstream_station	Symbol	Inference Process
Downstream_station	Symbol	Inference Process
Milepost	Quantity	Operator/Report
Direction	Symbol	Inference Process
Type	Symbol	Operator/Report
Duration	Quantity	Operator/Report
Lanes_blocked	Quantity	Operator/Report
Side_blocked	Symbol	Operator/Report
No_injuries	Quantity	Operator/Report
No_fatalities	Quantity	Operator/Report
Load_description	Symbol	Operator/Report
Load_weight	Quantity	Operator/Report
Total_lanes	Quantity	Inference Process
Capacity	Quantity	Inference Process

TABLE 2 COUNTING STATION OBJECT ATTRIBUTES

ATTRIBUTE NAME	TYPE	SOURCE
Incident_status	Quantity	External
Occupancy	Quantity	External
Volume	Quantity	External
Downstream_station	Symbol	Constant
Milepost	Quantity	Constant
Direction	Symbol	Constant

must be satisfied before actions in the consequents can be executed. When this occurs, the rule is said to have fired. Essentially, it is the execution of the actions within the consequents of the rules that drives the system.

The most important action command is the conclusion action because it drives the inference process. Conclusions about the state of the system are drawn, or inferred, from the conditions in the antecedent of the rule, and linking conclusions to antecedents achieves a chaining of individual rules via the inference process. The state of the system is embodied solely within the objects and their attributes. Rules make reference to the attributes of objects in a number of ways. For example, they can refer to the attribute of a particular instance of an object. The following rule refers to the occupancy attribute of the object WB5, which is an instance of the counting station class:

if the Occupancy of WB5 > 30
then. . .

It is more likely, however, that rules will be generic and will refer to object classes rather than instances of that class. The following rule applies to any instance of the counting station class and is the rule that initiates the incident management process (Incident_status refers to the attribute of the counting station class that contains a flag indicating whether or not an incident has been detected):

if the Incident_status of any count_station = 1
then activate the subworkspace of inc-detection

Such generic rules are well suited to the encoding of knowledge.

The chaining of rules can be done in two ways: forward and backward. Forward chaining attempts to match facts about the current state of the system to conditions embodied in the antecedents of all rules in the knowledge base. Rules that are fired may then provide further information, via conclusions, about the state of the system, which can fire more rules, and so on.

Backward chaining operates in reverse. The system attempts to infer a hypothesis about the state of the system by firing only the rules that are relevant to that hypothesis. An attempt to fire the relevant rules is made by examining the conditions contained in the antecedent. To satisfy these conditions, more rules may become relevant.

Thus, forward chaining is a data-driven process, whereas backward chaining is a goal-driven process. FRED is primarily data driven but uses backward chaining when necessary. For example, the major event that drives the incident management process is the detection of an incident. To respond to a particular incident, rules may require further information. Backward chaining will occur to furnish this information.

DATA ACQUISITION AND INTEGRATION IN FRED

This section examines the way in which data are transferred from external routines and sensors to the FRED expert system. In the current implementation of FRED, there are no sensors in the physical sense but, rather, a series of external routines that simulate these sensors. Separate programs model the provision of counting station data to the incident detection algorithm and the processing of outside reports. Even in real applications, the expert system would receive data from controller routines that would operate in a manner similar to that used in FRED.

Figure 3 shows the system components from the data interfacing perspective. FRED, and the external routines interfacing with it, are all processes running concurrently and sharing the resources of the central processing unit. Communication between these processes is via data files. The external routines write any data that need to be passed to FRED to the data files, which are accessed by a separate G2 standard interface (GSI). Data read from these files are passed to objects within FRED that have been designated as having external data sources. Application-specific bridge routines must be written within the GSI module to perform the interrogation of data files and returning of data to FRED.

The G2 expert system and the GSI module communicate by object indexes. Each object that has an external routine as its data source is assigned a unique object index by GSI when the system is started. In addition to the system-assigned index, each interface object has attributes, specified by the system developer, that can be used to identify the type of data. In FRED each variable with an external source has a data type attribute that is used to determine from which external routine the data originate. For example, the volume variable for each counting station has the same value for the data type attribute. In this way, the interface bridge routine knows when to access the data file containing the current volume values.

Data can be transferred from the external routines to FRED in two ways: as solicited input or unsolicited input.

Solicited Input

As part of the process of backward chaining, the inference engine may require the value of an attribute that has a data source external to the expert system. From the expert system's perspective, the external entity is polled for the current value of that variable; once it returns, the value processing continues. At a lower level, the inference engine passes a request for data to the GSI, specifying the index of the object for which a value is sought and passing the value of the data type

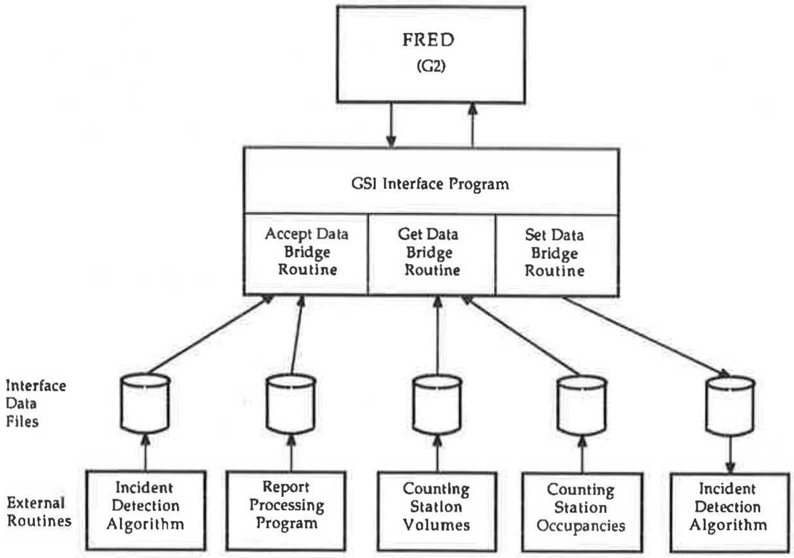


FIGURE 3 FRED data interface structure.

attribute. The "get data" bridge routine determines which data file to interrogate, opens it, and obtains the required value.

For example, when formulating responses to incidents, 30-sec volume counts are often required from the freeway mainline and ramp counting stations. These values are written to a data file every 30 sec by a separate simulation routine. If volume counts are required, a request is sent to GSI, and the "get data" bridge routine is called, passing the object index and the data type attribute. The bridge routine recognizes the data type attribute as being the one for volume counts. It opens the file containing the current volume counts, selects the appropriate value, and returns them to FRED.

In order to ensure an efficient interface, requests for data are usually grouped so that the interface bridge routine is passed an array of object indexes and data type attributes representing requests for a series of objects. In the preceding example, requests for volumes from a series of counting stations are grouped and sent to GSI at the same time. This procedure ensures that the volume data file is opened only once.

Unsolicited Input

Sensors or programs external to the expert system can send input without it being directly requested by G2. In this case, the external routine, not the expert system, is making the decision to transfer data. Such data transfers can act as interrupts to the current operation of the system. An "accept data" bridge routine is called by the interface program every second, and this routine interrogates data files that may contain input from external routines. If the information is found, it is returned along with the appropriate object index.

In FRED, the detection of an incident by the incident detection algorithm leads to unsolicited data transfer from the algorithm to FRED by the GSI interface. The Incident_status attribute of each counting station (see Table 1) has an external source, namely, the incident detection algorithm. When this algorithm detects an incident, the identification number of the counting station immediately upstream of the incident is written to a data file. This file is examined by the "accept data" bridge routine in GSI every second. If new data are found, the bridge routine reads the identification number of the upstream counting station from the file and returns 1 as the value for the Incident_status attribute of that counting station.

Passing Data from FRED to External Routines

The behavior of the external routines can be controlled by FRED. Object attributes with external sources can be set by rules in the knowledge base. This procedure is the reverse of data retrieval because GSI is passed a value along with an object index. The "set data" bridge routine can then use that value to set some aspect of an external routine. For example, the calibration of the incident detection algorithm can be altered by setting the algorithm number and threshold parameters. Updated values for these parameters are sent to the "set data" bridge routine and then written to a data file by the bridge routine. The incident detection algorithm interrogates this data file at regular intervals; if new values are found, they are used to recalibrate the model.

Validity of Data

As mentioned, data representing the state of the external system are continually changing, and the inference engine must be able to change its conclusions on the basis of new data. This process is termed nonmonotonic reasoning. Conclusions inferred at one stage of the system may become invalid later. Thus, data and conclusions have validity intervals associated with them. For example, counting station volumes are valid for only 30 sec. Any conclusions inferred from these variables are also valid for only 30 sec.

Variables with external data sources have prescribed validity intervals. Variables whose values are inferred from rules have their validity intervals supplied by the inference engine as the minimum validity interval of the facts used in the inference. For example, in the following rule, if facts *A* and *B* expire in 5 and 10 sec, respectively, then fact *C* will have an expiration time of 5 sec from the time at which the rule fired:

if *A* is true and *B* is true
then conclude that *C* is true

System Robustness

Any system that interfaces with a large number of external processes must be tolerant of malfunctions and breakdowns. A feature built into the inference engine causes a failed data request to be retried after a specified interval, typically 5 sec. If such attempts repeatedly fail, the variable causing the request is said to have timed out. A special rule handles such a situation. For example, the following rule fires when the volume attribute of any counting station times out:

whenever the volume of any count_station c1 fails to receive a value
then inform the operator that "Request for volume value failed"

Additionally, rules can be written to check for erroneous data, such as negative or excessively high volume counts.

In G2 and FRED, the interface program can send error messages and statuses back to the expert system in the case of erroneous communication with an external routine. A separate body of rules within the expert system can be invoked to handle such situations.

Selective Knowledge Processing

In order to maintain a real-time expert system, consideration must be given to speed of rule processing. In FRED there is a need for control over incident management tasks, particularly when considering multiple incidents. The tasks associated with managing each incident must be organized to use the resources of the inference engine and the TOC operator efficiently.

The FRED system was built in such a manner that the rules associated with each incident management task could be invoked when required. Responsibility for invoking these subknowledge bases resides with a set of management rules. Rules in FRED are activated and deactivated during the running of the system, thus reducing the load on the inference engine at any one time. For example, none of the incident management rules are active when no incidents have been detected. Once an incident is detected, the incident detection rules are activated, followed by the incident verification rules, and so on.

Currently, FRED only works on a single incident at a time, but later expansion will incorporate multiple incidents, with some method of ranking the incident management tasks for several incidents. The formulation of incident responses is complicated when considering multiple incidents on the same section of freeway.

System Performance

The case study corridor used for developing the FRED prototype system is relatively small when considering the amount of surveillance data received by the system. Future research should involve on-line testing in a larger freeway environment to evaluate the actual real-time performance of the system. However, the FRED system approach will undoubtedly meet the processing requirements of incident management. The G2 shell used in the system has recently been successfully employed in a number of industrial and aerospace applications, including the U.S. space shuttle, with much more demanding performance criteria. The nature of freeway traffic control is such that the typical response times are several minutes rather than seconds. The most likely performance limitation in FRED will be its ability to handle several simultaneous incidents. Ongoing research is addressing this necessary capability.

DATA INTEGRATION IN INCIDENT DETECTION AND VERIFICATION

In order to illustrate the manner in which data arriving from different sources are integrated within FRED, the basic tasks in incident detection and verification are considered. To simplify the example, it is assumed that there are two ways to establish the existence of an incident: (a) automated incident detection with verification by operator-controlled CCTV and (b) on-site incident reports.

Automated Incident Detection and CCTV Verification

The initial event in this process is the detection of an incident by the incident detection algorithm. The detection message is sent as unsolicited input to FRED by setting the Incident_status attribute of the counting station that triggered the detection to 1. This counting station, call it *C*, is always the counting station immediately upstream of the incident site. A rule is fired whenever the Incident_status attribute of any counting station is set to 1 and leads to the creation of an incident object. Referring to Table 1, the following attributes of the new incident are set at this stage:

- *Status.* Set to the symbol "possible" at this stage.
- *Upstream_station.* Set to *C*—the name of the counting station that triggered the original detection.
- *Downstream_station.* Set to the name of the counting station object immediately downstream of the incident. Obtained from the value of the Downstream_station attribute of *C*.
- *Milepost.* The approximate milepost position of the incident set at this stage to the average of the Milepost constants of the upstream and downstream counting stations.
- *Direction.* The direction of the carriageway on which the incident was detected, for example, eastbound or westbound, copied from the Direction attribute of upstream counting station *C*.

From the operator's perspective, the next step is to verify the incident (in this simplified example, by consulting a CCTV camera). From visual inspection of the incident site, the operator is required to enter information for the following incident attributes:

- *Type.* A selection from a prescribed list of incident types, such as overturned truck, three-car accident, and stalled vehicle.
- *Duration.* An estimate of the duration of the incident in hours.
- *Lanes_blocked.* The number of lanes of the carriageway that are blocked by the incident.
- *Side_blocked.* The side of the carriageway, left or right, blocked by the incident.
- *No_injuries.* The number of confirmed injuries resulting from the incident.
- *No_fatalities.* The number of confirmed fatalities resulting from the incident.

For an overturned truck or spilled load, the following attributes need to be set as well:

- *Load_description.* A text string of free format describing the spilled load (e.g., diesel fuel).
- *Load_weight.* An estimate of the weight of the spilled load in tons.

Additionally, at the incident confirmation stage, the operator is able to specify the location of the incident more precisely. This step is done in FRED by moving a screen marker on a freeway schematic, drawn approximately to scale, to the position of the incident as seen from the CCTV camera. Once the location is confirmed by the operator, the Milepost attribute of the incident is updated using the coordinates of the incident marker. Finally, the Status attribute of the incident is set to "confirmed" to denote a confirmed incident.

Certain attributes of an incident have their values inferred from other attribute values by formulas and rules:

- *Total_lanes.* The total number of lanes at the incident site, set by examining a table that lists the number of lanes for different milepost locations.
- *Capacity.* The unrestricted capacity of the freeway at the incident site, again obtained from a table ordered on milepost values.

Outside Reports

The incident attributes outlined previously are set in a different manner when an incident is detected only from outside reports.

A major part of the acquisition of data from outside reports is performed by the report processing program external to FRED, which allows a communications center operator to enter an outside report. The user is prompted for all necessary details. The information is then transferred to FRED as unsolicited input and, as in the previous case, an incident object is created. The following incident attributes, described previously, are derived from the report:

- Milepost,
- Direction,
- Type,
- Duration,
- Lanes_blocked,
- Side_blocked,
- No_injuries,
- No_fatalities,
- Load_description, and
- Load_weight.

The Milepost attribute is used to infer the names of the upstream and downstream counting stations. The remaining attributes are inferred as in the previous case.

Once an incident has been confirmed, the data contained within the attributes are used to formulate responses. For example, information describing the nature of the incident is used to compute the expected capacity of the freeway at the incident site, and FRED then recommends whether to close entrance ramps upstream of the incident.

CONCLUSIONS

The operating requirements of a real-time expert system to aid traffic control operators in incident management place particular emphasis on the efficient acquisition and integration of data from a number of external sensors. The manner in which the FRED system achieves this objective was outlined, with particular reference to the data interfacing between several concurrent processes in the incident detection and verification phases of incident management. The system structure developed serves as a good foundation for the development of a sophisticated and robust real-time system for management of multiple incidents that commonly occur on large freeway systems. In this respect, the refinement, expansion, and evaluation of the FRED system is ongoing.

ACKNOWLEDGMENTS

This research was supported by a U.S. Department of Transportation and University of California Transportation Center award for "Real-Time Decision Support for Freeway Surveillance and Control" and by a grant from the National Science Foundation.

REFERENCES

1. *Report to Congress on Intelligent Vehicle-Highway Systems.* Office of Economics and Office of the Assistant Secretary for Policy and International Affairs, U.S. Department of Transportation, 1990.
2. S. Sellam, A. Boulmakoul, and J. C. Pierrelee. Intelligent Intersection: A Monitoring and Control Integrated System. Presented at OECD Workshop on Knowledge-Based Expert Systems in Transportation, Espoo, Finland, June 1990.
3. S. G. Ritchie. A Knowledge-Based Decision Support Architecture for Advanced Traffic Management. *Transportation Research,* Vol. 24A, No. 1, 1990.
4. H. J. Payne and H. C. Knobel. *Development and Testing of Incident Detection Algorithms, Volume 3: User Guidelines.* Report FHWA-RD-76-21. Technology Service Corporation, FHWA, U.S. Department of Transportation, 1976.
5. *G2 User's Manual.* Gensym, Inc., Cambridge, Mass., 1988.

Publication of this paper sponsored by Committee on Applications of Emerging Technology.

Less-Than-Truckload Trucking in Los Angeles: Congestion Relief Through Terminal Siting

RANDOLPH W. HALL AND WEI HUA LIN

Traffic congestion in the Los Angeles, San Francisco, and San Diego regions may cost individuals and industry as much as $2 billion a year in lost time and lost productivity. Much of this cost is borne by the trucking industry and its customers. A major study was conducted at the University of California to develop and evaluate strategies for reducing congestion delays incurred by less-than-truckload motor carriers in Los Angeles. It was found that surplus capacity exists in the Los Angeles freeway system throughout much of the day and that trucks naturally avoid congested highways. Though existing terminals tend to be well located, some areas—notably, west Los Angeles and the Los Angeles Airport–El Segundo area—are not well served. Further, a shortage of vacant land in the south-central industrial core may force carriers to build terminals in remote locations. It is proposed that motor carriers be encouraged to establish terminals in these areas and to consolidate and merge their pickup and delivery operations, which would facilitate service from more sites.

The state of California recently commissioned the Urban Freeway Gridlock Study (*1*) "to investigate the impact of large trucks on peak-period freeway congestion." Among the study findings was that "congestion in Los Angeles, San Francisco, and San Diego may cost as much as $2 billion per year."

Although the motivation for this and other studies (*2,3*) was to measure and reduce congestion caused by trucks, it is clear that a large portion of the congestion costs is borne by trucks. According to the Urban Freeway Gridlock Study, each hour that a truck spends on the road costs $44 in wages, maintenance, fuel, and overhead—a number considerably larger than the value of an automobile's time. Delays impose additional costs on shippers and receivers, which depend on carriers for timely service.

LESS-THAN-TRUCKLOAD (LTL) TERMINAL OPERATIONS

LTL motor carriers transport medium-size shipments—shipments that are too large for parcel post or UPS, but too small for truckload service. According to the Urban Freeway Gridlock Study (*1*), approximately 45 percent of the truck miles in the Los Angeles region are by LTL carrier, of which approximately half are by LTL common carrier (e.g., Consolidated Freightways or Yellow Freight System) and half are by LTL private carrier.

A typical LTL carrier transports shipments in three phases: local pickup, line haul, and local delivery. The pickup phase ordinarily occurs in the afternoon, after most of the days' orders have been received. After visiting multiple stops, the pickup and delivery truck deposits its load at an end-of-line terminal (usually the carrier's closest terminal to the shipment origins). From there, shipments are transported in larger line-haul vehicles, either to another end-of-line terminal or, in some cases, to a breakbulk terminal for further sorting. After the line-haul phase, shipments are delivered in the morning from an end-of-line terminal (usually the closest to the shipment destinations) in pickup and delivery vehicles. In terms of congestion relief, the important characteristics of the system are that (a) deliveries occur in the morning, (b) line haul occurs overnight, and (c) pickups occur in the afternoon.

Because line haul occurs overnight, it is not greatly affected by road congestion (nor does it contribute greatly to congestion). Pickups and deliveries, on the other hand, must occur during the day when businesses are open, often during the morning and afternoon travel peaks. These trucking routes are affected by congestion, especially if pickup and delivery vehicles travel in the same direction as commuters. That is, congestion has the biggest impact on delivery routes heading toward work centers in the morning and pickup routes heading away from work centers in the evening.

One way to provide congestion relief would be to schedule trucks so that they travel in off-peak periods. However, for LTL common carriers, such a schedule would require major changes in carrier, shipper, and receiver operations. Pickups and deliveries would either have to occur at midday, after deliveries are needed and before pickups are available, or they would have to occur overnight. However, overnight pickups and deliveries would disrupt line-haul operations and could be costly to shippers and receivers that do not currently have nighttime staffing.

Alternatively, trucking delay could be reduced by selecting better locations for LTL terminals. The strategic location of more satellite terminals would have the following beneficial effects:

- The lengths of pickup and delivery routes would be reduced, and
- Travel across congested road segments would be reduced.

In order to achieve the second objective, pickup and delivery trucks can be routed in the same manner as reverse commuters. For instance, if an LTL terminal were located

Department of Industrial Engineering and Operations Research, University of California Transportation Center, Berkeley, Calif. 94720.

near a work center, trucks would travel in the opposite direction of commuters in the morning, as they leave to make deliveries, and in the opposite direction of commuters in the evening, as they return with their pickups.

OBJECTIVE

The objective of the following paragraphs is to summarize the findings of a research study conducted at the University of California that assessed the potential for reducing trucking delays caused by urban road congestion in the Los Angeles region. The emphasis is on improved selection of terminal sites for LTL common carriers, especially on the feasibility of placing terminals at locations that exploit surplus freeway capacity (e.g., reverse commuting).

The next four sections address the following issues in order: (a) Where and when is there excess capacity on the Los Angeles freeway system? (b) To what extent are carriers currently exposed to congestion? (c) Where are trucking terminals currently located? and (d) Where are the opportunities for improved terminal sites?

Although the focus is on reducing delays incurred by trucks, any strategy that reduces trucking delay would benefit motorists through reduced traffic volume on congested roadways.

SURPLUS CAPACITY IN THE LOS ANGELES FREEWAY SYSTEM

It has been said that Los Angeles is a city without a center, where work and residences are spread amorphously throughout a massive region. Indeed, compared with other major cities, Los Angeles is decentralized. But it is not true that it has no center. In recent years, the downtown of Los Angeles has experienced tremendous job growth, including the construction of highrise buildings that top the 1,000-ft mark. This large concentration of employment has come to have a significant impact on traffic patterns throughout the region. Like all major cities, roadways leading toward the downtown are congested in the morning and those leading away are congested in the evening. The reverse commute directions tend not to be as congested, and roadways located far from the center tend to be less congested.

The dominance of the traditional downtown is evident in the city's ring-radial freeway network, shown in Figure 1, especially in the vicinity of the downtown. In the newer parts of the region—San Fernando Valley and Orange, Riverside, and San Bernadino counties—downtown Los Angeles is less dominant, and freeways follow more of a grid structure.

An attempt was made to determine when and where surplus capacity exists on the Los Angeles freeway system through the analysis of traffic flow data. Through examination of highway congestion maps generated by the California Department of Transportation (Caltrans), it was initially observed that few freeway segments are congested in both directions at the same time [major exceptions include Ventura Freeway (101); San Diego Freeway (I-405), from I-10 to I-110; Santa Monica Freeway (I-10); and several road segments downtown and in the vicinity of Anaheim]. These observations were supported by an analysis of average daily traffic (ADT) and peak-hour counts (*4–6*). Nevertheless, neither of these data sources provides a detailed picture of how traffic levels vary by time and direction.

To follow up, an analysis of traffic data generated by the Caltrans automated freeway control system was conducted. This system is one of the most sophisticated in the world for monitoring real-time traffic flows. Flow detectors have been installed at over 900 sites, covering most of the major freeways in Los Angeles and Orange counties. Twenty-two sites were selected, spaced roughly at 5-mi intervals along the routes radiating from downtown Los Angeles. Figure 2 shows the total traffic count among all 22 sites by direction and time of day, and Table 1 presents traffic counts for individual sites.

The following observations were made:

- From 6:00 to 9:00 a.m., traffic leading toward the downtown exceeds traffic heading away by 28 percent.

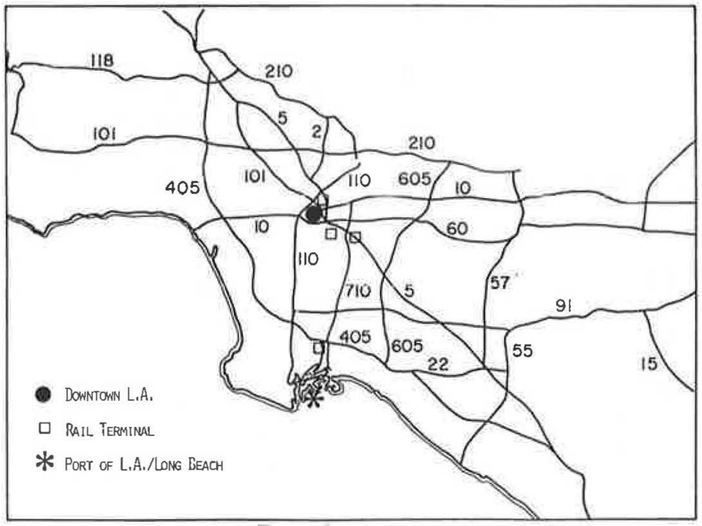


FIGURE 1 Major transportation facilities in Los Angeles region.

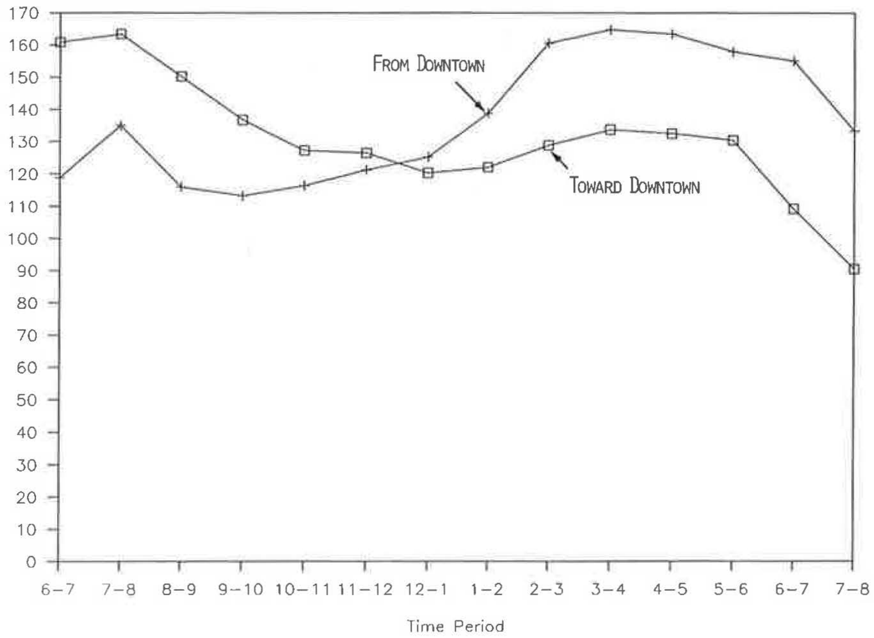


FIGURE 2 Traffic flow on radial routes by time and direction.

TABLE 1 TRAFFIC COUNTS BY TIME AND DIRECTION

			Average Hourly Traffic		
Freeway	Location	Direction	6-9	9-2	2-6
5	Burbank	To	7700	4800	5800
5	Burbank	From	5300	4200	6600
5	Tuxford	To	7100	3900	4000
5	Tuxford	From	3600	3700	6100
5	S. of 710	To	7900	6000	6000
5	S. of 710	From	5700	5500	6100
5	Los Feliz	To	7400	5400	6900
5	Los Feliz	From	6100	5100	7200
5	Rosencrans	To	5300	4500	4900
5	Rosencrans	From	4500	4400	4900
5	Orangethorpe	To	5800	5100	5700
5	Orangethorpe	From	10500	9100	11500
10	Crenshaw	To	6800	7300	7900
10	Crenshaw	From	5200	7500	7900
10	Westwood	To	7900	8000	8200
10	Westwood	From	8400	8000	8500
10	E. of 710	To	6000	5000	4000
10	E. of 710	From	3300	4400	5600
10	Atlantic	To	6700	5800	5400
10	Atlantic	From	4100	5100	6800
10	Rosemead	To	7100	6100	5900
10	Rosemead	From	4100	5300	7400
10	Citrus	To	7500	4800	5100
10	Citrus	From	4200	4300	7300
60	Atlantic	To	9700	6300	5400
60	Atlantic	From	4500	4900	7700
60	Rosemead	To	9000	6000	5800
60	Rosemead	From	4400	4600	7200
60	Turnbull	To	7400	5700	6500
60	Turnbull	From	5200	5700	8700
101	Western	To	6800	5700	5100
101	Western	From	5300	5400	7000
101	Vineland	To	8500	6800	5900
101	Vineland	From	5600	5600	8400
101	Van Nuys	To	6800	7200	7400
101	Van Nuys	From	9000	9900	9400
110	Slauson	To	6600	6500	5000
110	Slauson	From	6500	5900	7600
710	Imperial	To	6700	4800	6200
710	Imperial	From	5900	4200	6500
710	Del Amol	To	6000	4600	6400
710	Del Amol	From	6200	4300	5600
TOTAL		TO	158000	127000	131000
TOTAL		FROM	123000	123000	162000

- From 2:00 to 6:00 p.m., traffic leading away from the downtown exceeds traffic heading toward the downtown by 23 percent.
- Between 9:00 a.m. and 2:00 p.m., traffic levels leading toward the downtown average 80 percent of the peak-period count, and those heading away from the downtown average 76 percent of the peak-period count.
- Heading toward the downtown, the p.m. count is nearly the same as the midday count. Heading away from the downtown, the a.m. count is nearly the same as the midday count.

It appears that most radial freeways operate below capacity for most of the day. Most radial roads leading toward the downtown only operate at peak capacity from about 6:00 to 9:00 a.m. Most roads heading away from the downtown only operate at peak capacity from about 2:00 to 6:00 p.m.

EXPOSURE OF TRUCKS TO CONGESTION

Ordinary two-axle automobiles and light trucks compose the vast majority of vehicles on the Los Angeles freeway system, and general traffic patterns are dominated by personal trips. During peak hours, traffic is further dominated by commute trips, which are reflective of where people live and work.

On most Los Angeles freeways, less than 5 percent of the vehicles are large trucks (three axles or more), whose traffic patterns differ substantially from personal vehicles. Most of their trips begin and end at manufacturers, transportation terminals, or warehouses. Consequently, truck travel patterns reflect the locations of industrial facilities.

In the Los Angeles region, the largest concentration of manufacturers and warehouses is in south central Los Angeles along the Santa Ana Freeway (I-5) corridor. The largest transportation terminals are the ports of Los Angeles and Long Beach; Southern Pacific's Long Beach intermodal terminal, 20 mi south of downtown; and the Santa Fe, Southern Pacific, and Union Pacific intermodal terminals in the vicinity of the downtown (see Figure 1). In addition, unlike automobile traffic, which is dominated by commuting, truck trips are spread throughout their workday. These factors combine to cause truck travel patterns to differ substantially from general traffic patterns, in terms of both where they are on the road and when they are on the road.

The Urban Freeway Gridlock Study (1) found that "a substantial number of freeway sites had a smaller percentage of large trucks than is suggested by the statistical average." This issue was examined in depth through statistical analysis of truck and automobile traffic in the Los Angeles region. The data were collected by Cambridge Systematics Inc. as part of the Gridlock study using a video camera at 40 sites. For each site, truck (three or more axle) and automobile counts were obtained for six 15-min periods, two in the a.m. peak, two at midday, and two in the p.m. peak. In all, recordings were made for 221 15-min periods (19 data points were missing).

Figures 3a, 4a, and 5a show the percentage of trucks on the road by time of day. Similar graphs, for off-peak and p.m. peak, were presented elsewhere by Hall and Lin (4). The relationship between truck traffic and total traffic was ap-

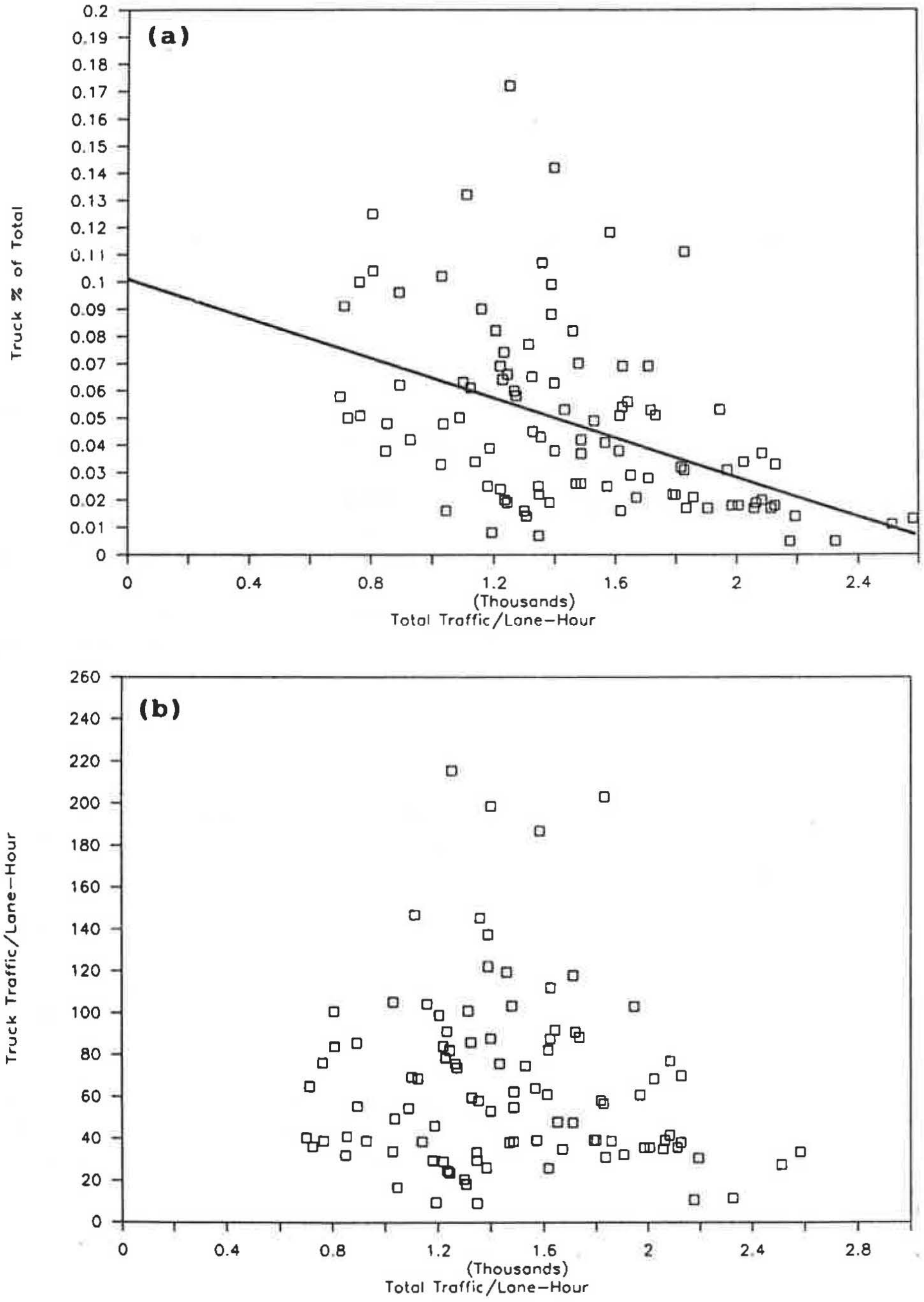


FIGURE 3 Truck traffic versus total traffic per lane-hour: Los Angeles a.m. traffic.

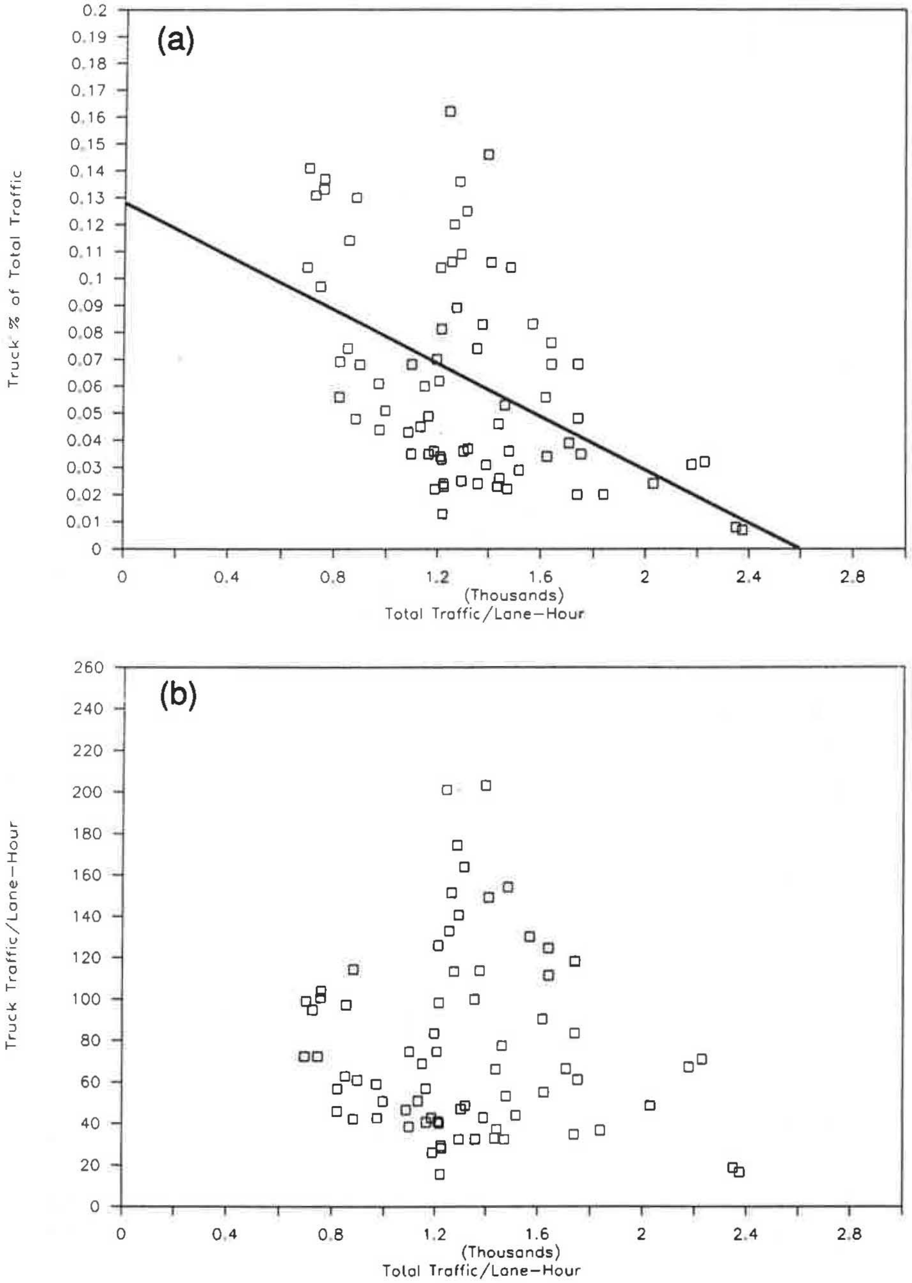


FIGURE 4 Truck traffic versus total traffic per lane-hour: Los Angeles off-peak traffic.

proximated by ordinary least-squares linear regression [negative correlation is weak (R_2 = 0.23) but significant, with t = 5.1].

It was observed that there is a natural tendency for trucks to avoid the most congested roadways:

- As a percentage of total traffic, truck traffic declines as traffic volume increases in the a.m., midday, and p.m. periods.
- As a percentage of total traffic, truck traffic is largest at midday, when roadways are the least congested, and smallest in the p.m. peak, when roadways are the most congested.

Figures 3b, 4b, and 5b convert the percentages into trucks per lane-hour. The figures demonstrate the following additional points:

- Truck traffic is nearly the same during the a.m. peak and midday but considerably smaller during the p.m. peak.
- Total traffic is largest during the p.m. peak, slightly smaller during the a.m. peak, and considerably smaller during the off-peak period.
- The largest truck volumes occur on freeways with medium traffic volumes (approximately 1,400 veh/lane-hr). Truck traffic is smaller on roads that have either a very large or a very small traffic level.

Roughly the same number of trucks is on the road during the a.m. peak and midday, probably because driver workshifts run from early morning through the afternoon. The higher percentage of trucks at midday is due to fewer automobiles

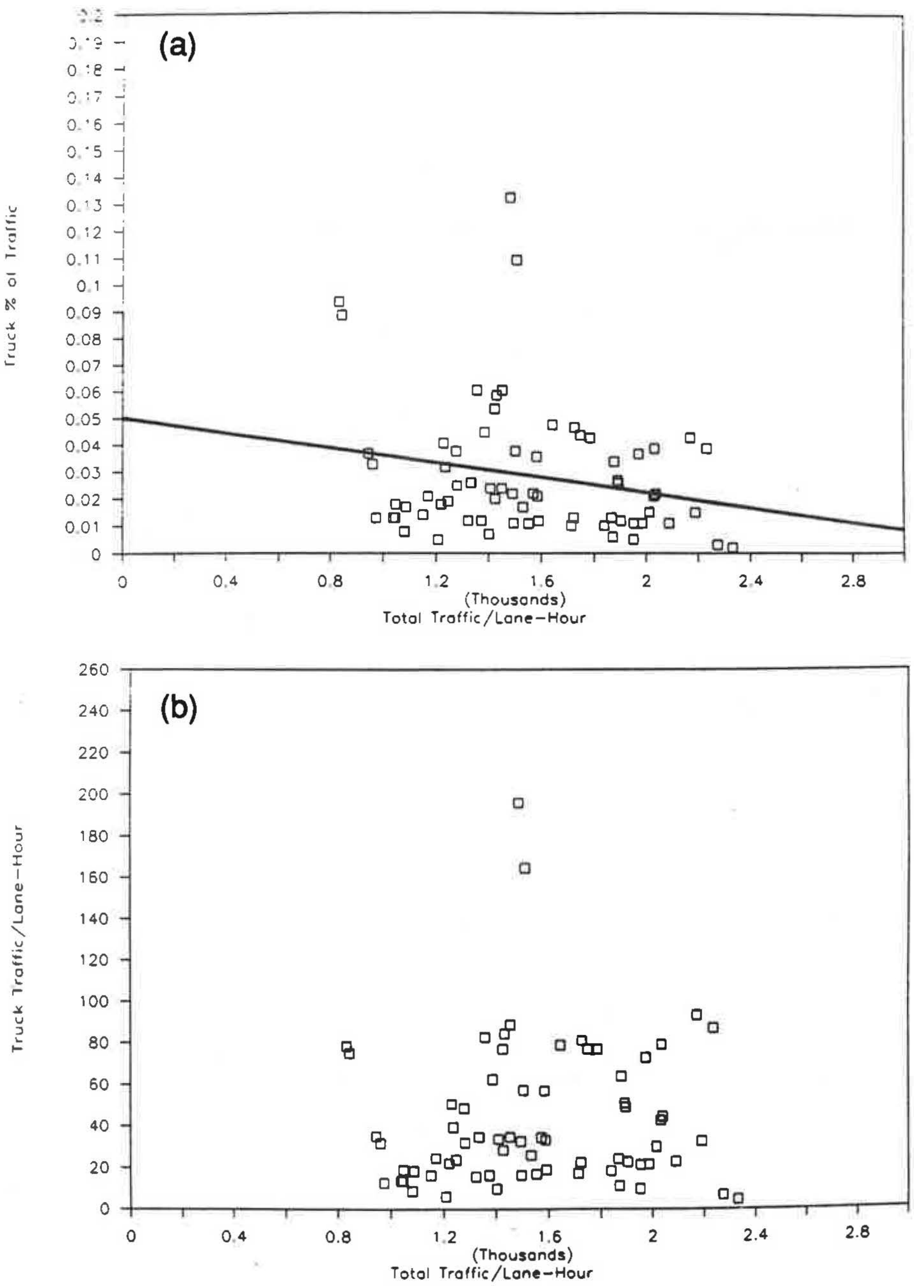


FIGURE 5 Truck traffic versus total traffic per lane-hour: Los Angeles p.m. traffic.

being on the road, not more trucks. During the p.m. peak, there are fewer trucks on the road because many carriers have terminated operations for the day. In all periods, there is a natural tendency for trucks to avoid congested routes.

The heaviest a.m. truck volumes occur on roadways that are relatively uncongested and are located near the ports of Los Angeles and Long Beach and near the south central industrial corridor. The smallest truck volumes occur heading toward the downtown or other employment centers, over congested roads. The absolutely lowest truck percentages occur on the Santa Monica Freeway, the most densely developed commercial corridor in the region.

In conclusion, trucks in the Los Angeles region seem to naturally avoid the most congested roadways. Nevertheless, it is impossible for trucks to avoid congested freeways completely (especially the Santa Ana and San Diego freeways).

LOCATIONS OF EXISTING TRUCKING TERMINALS

The travel pattern for trucks in the Los Angeles region depends on the locations of the major trip generators—manufacturers, warehouses, the ports, and rail yards—as well as on locations of trucking terminals. By locating terminals closer to trip generators, the total number of truck-miles can be reduced. By strategically locating terminals to exploit excess freeway capacity, the number of truck-miles over congested roadways can be reduced.

The existing locations of trucking terminals in the region are discussed in the following paragraphs. The analysis is based on two data sources: (a) the California Trucking Association (CTA) file of Los Angeles and Orange county terminals (San Bernadino and Riverside counties were excluded) and (b) service directories for seven major LTL carriers. Neither data source is comprehensive. Nevertheless, each gives a representative sample of existing terminal locations and provides insight into truck travel patterns in the Los Angeles region.

General Terminal Locations

The CTA data file contains the addresses of 600 terminals covering a wide variety of trucking types—LTL and truckload, private and common, and specialized carriers. The CTA data were coded by zip code, and the zip codes were ranked from largest to smallest according to number of terminals. From this ranking, it was found that the greatest concentrations of terminals are in south central Los Angeles—in the Santa Ana corridor stretching from downtown to Santa Fe Springs—and in the vicinity of the ports of Los Angeles and Long Beach, stretching up to Gardena. The first area is the manufacturing core of the region, and the second is the distribution core. Clearly, large numbers of motor carriers have selected sites to serve these customers.

Comparison with Manufacturing Employment

If terminals are to be located effectively, they should be placed close to the customers. Because manufacturers are the single largest source of customers, the terminal location pattern was compared with the pattern of manufacturing employment in the region (*7*). Through statistical regression, the relationship between number of terminals and manufacturing employment, by zip code, was approximated by a linear equation.

Figure 6 shows zip codes in which the number of terminals differs appreciably from that expected, given manufacturing employment. The Santa Ana Freeway corridor and the port area have especially large concentrations of terminals. On the other hand, several zip codes located in the suburbs, such as Chatsworth, El Segundo, and Redondo Beach, have far fewer terminals than expected.

Explanation of Terminal Clusters

Manufacturing employment is just one determinant of the number of terminals in a zip code. Other factors include manufacturing employment in nearby zip codes and location of major transportation facilities. The latter factor surely explains the large concentration of terminals in the port area. For carriers that serve the ports to locate elsewhere would be inefficient.

The concentration of terminals along the Santa Ana corridor is probably due to its central location for those carriers that serve the entire Los Angeles region from a single terminal. To test this reasoning, alternative terminal sites were analyzed according to the average straight-line distance to manufacturing job sites in Los Angeles, Orange, Riverside, and San Bernadino counties. The most central location was found to be in Bell, west of I-710 and south of I-5, in the center of a terminal cluster. Locations were further analyzed to see which sites were nearly optimal for minimizing average straight-line distance. The middle and outer rings of Figure 7, representing average distance within 5 and 10 percent of optimal, almost exactly match the terminal pattern of Figure 6.

LTL Terminals

Terminals were evaluated for seven LTL common carriers: the three major carriers (Consolidated Freightways, Roadway, and Yellow), three smaller nationwide carriers (ABF, ANR, and P.I.E.), and a large regional carrier (Viking). Terminal locations and boundaries of terminal service regions were obtained from published service directories (*8–14*).

Topography and, to a lesser extent, governmental boundaries play a key role in dictating terminal locations and service

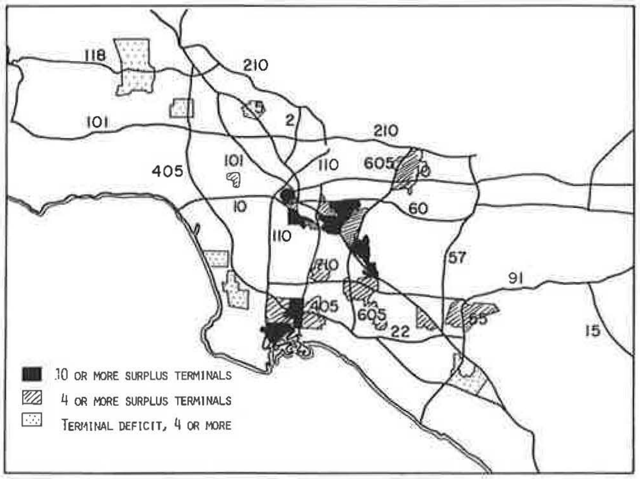

118
210
101
5
2
210
405
101
110
605
10
10
60
110
710
57
91
405
605
22
55
15
■ 10 OR MORE SURPLUS TERMINALS
▨ 4 OR MORE SURPLUS TERMINALS
⬚ TERMINAL DEFICIT, 4 OR MORE

FIGURE 6 Zip codes containing many, or few, trucking terminals.

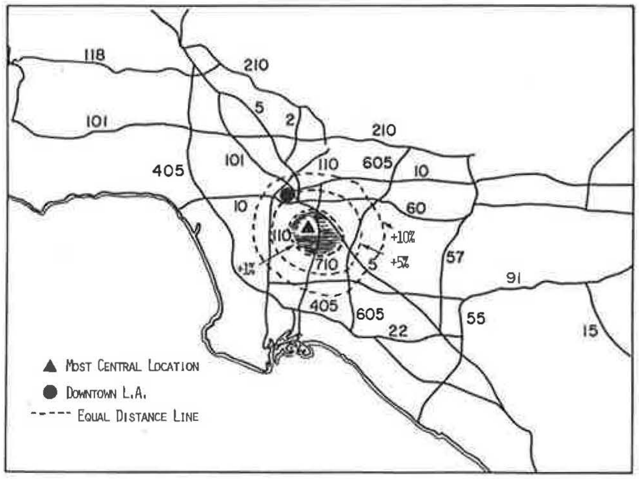


FIGURE 7 Most central location relative to manufacturing employment.

regions. All carriers appear to have divided Los Angeles into four distinct regions, as shown in the following table:

Region	*Boundaries*
San Fernando	Santa Monica and San Gabriel mountain ranges
Eastern	San Gabriel and Santa Ana ranges
Orange	Santa Ana range and Orange County line
Central	Santa Monica and Santa Ana ranges and Orange County line

To illustrate the region's topography, Figure 8 shows the positioning of 1,000-ft contour lines. Areas located above the 1,000-ft level tend to be sparsely populated and difficult to traverse. Consequently, service regions have been drawn to minimize pickup and delivery truck travel across the mountains.

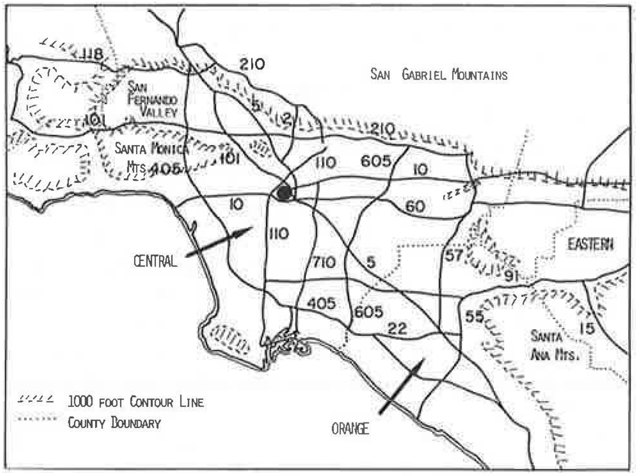


FIGURE 8 Los Angeles region topography.

TABLE 2 MAJOR MANUFACTURING CENTERS

		Large LTL Carriers			
Consolidated		Roadway		Yellow	
San Fernando/Ventura					
Oxnard	92613	Ventura	93001	Ventura	93003
Pacoima	91331	Pacoima	91331	Sun Valley	91352
Simi	93065				
Central: Santa Ana					
L.A.	90021	L.A.	90058	L.A.	90021
Montebello	90640			Pico Rivera	90660
Santa Fe S.	90670	Santa Fe S.	90670	Santa Fe S.	90670
Central: San Diego					
Carson	90745	Wilmington	90744	Carson	90810
Compton	90220	Gardena	90248	Gardena	90247
Inglewood	90304				
Central: East County					
Industry	91789	Industry	91746		
Irwindale	91706				
Orange County					
Irvine	92718	Irvine	92718	Irvine	92718
Los Alamit.	90720				
Orange	92665	Orange	92667	Orange	92668
Eastern					
Fontana	92335	Ontario	91743	Fontana	92335
San Bern.	92324	San Bern.	92316	Riverside	92507

			Medium LTL Carriers				
ABF		ANR		PIE		Viking	
San Fernando/Ventura							
Oxnard	93001	Oxnard+	93030	Ventura	93003	Oxnard	93030
Pacoima	91331	Pacoima	91331	Pacoima	91331	Sun Valley	91352
Central							
Carson	90810	Long Beach	90810	Carson	90746	Gardena	90248
Pico River	90660	L.A.	90023	L.A.	90058	Whittier	90601
				Montebello	90023		
				Industry	91769		
Orange							
Orange	92667	Orange	92665	Orange	92665	Anaheim	92806
Eastern							
Pomona	91767	San Bern.	92408	Fontana	92234	Pomona	91767
Colton	92324			San Bern.	92402	San Bern.	92408

• Recently closed

Table 2 presents the terminals by region, grouping terminals with similar locations. For instance, the first line gives the terminal serving locations to the west of San Fernando Valley. In all seven cases, that terminal is located in either Ventura or Oxnard. In Figure 9, locations of all seven carriers have been plotted, and Figure 10 shows terminal locations and service region boundaries for an example carrier, Roadway. Other maps are provided by Hall and Lin (*4*).

A minimum of six terminals is needed to serve the region competitively, one each for Oxnard, San Fernando Valley, Long Beach, central Los Angeles, Orange County, and San Bernadino-Riverside. The larger carriers have divided Los Angeles into smaller service regions and established more terminals. Roadway, for instance, serves central Los Angeles from five terminals: Gardena, Industry, Long Beach, south central Los Angeles, and Santa Fe Springs. Roadway has also established two terminals in Orange County (Irvine and Orange) and two terminals in the eastern region (Ontario and San Bernadino). Nevertheless, there is no precise relationship between carrier size and number of terminals. Yellow has 75

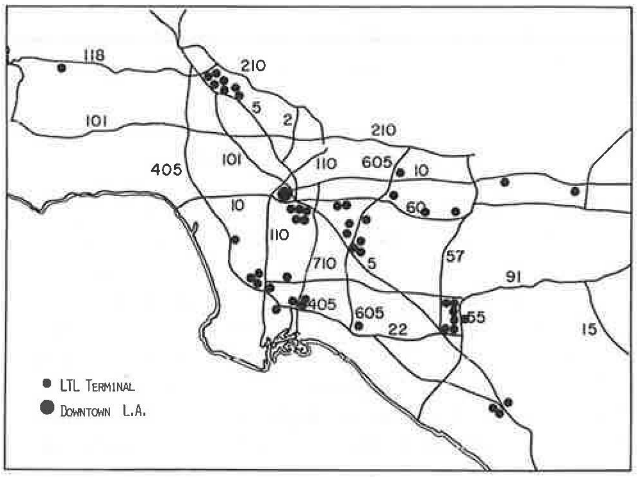


FIGURE 9 Locations of LTL terminals.

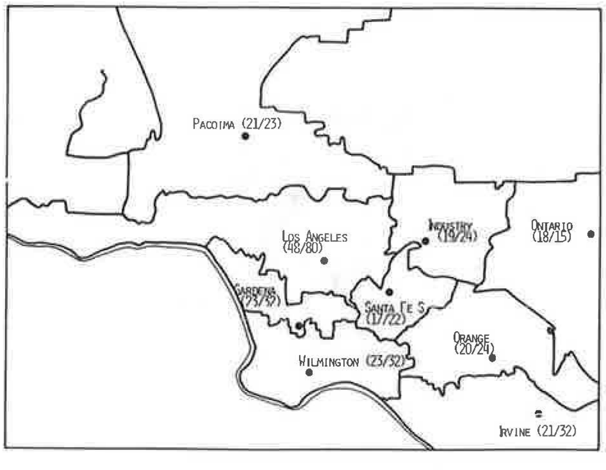


FIGURE 10 Roadway terminals and service regions (parentheses indicate tractors/doors).

more tractors and 73 more docks than Consolidated, but five fewer terminals. Viking, with six terminals, has a comparable number of docks as Roadway, which has 11.

The large carriers hold a natural advantage over the small in minimizing their pickup and delivery costs. With more terminals, large carriers can generally serve their customers from closer terminals. Consequently, they are not as exposed to road congestion delays. Consolidated, for instance, can serve the manufacturing center in El Segundo from nearby Inglewood. All other carriers would have to dispatch a truck from the Long Beach–Gardena area or from south central Los Angeles.

The distance and route from terminal to customer dictate the extent to which trucks are exposed to highway congestion. Some manufacturing centers enjoy much closer service than others, as indicated in Table 3. Accounting for existing terminal locations, freeway capacities, and prevailing traffic patterns, several industrial areas are especially difficult to serve:

- Carriers serving the entire region from south central Los Angeles:
  - —San Fernando Valley,
  - —West Los Angeles–Santa Monica, and
  - —Los Angeles Airport–El Segundo; and
- LTL carriers serving the region from multiple terminals (other than Consolidated):
  - —Glendale-Pasadena,
  - —Huntington Beach,
  - —Irwindale-Azusa,
  - —Los Angeles/Culver City, and
  - —Los Angeles Airport–El Segundo.

The San Diego corridor, from west Los Angeles to Redondo Beach, presents the biggest challenge. The San Diego Freeway is highly congested, often in both directions, and there is a general shortage of trucking terminals.

TABLE 3 LTL TERMINALS FOR SEVEN CARRIERS

Center/zip		Ave Dist to Terminal+	CF Dist to Terminal@	Access Route*
San Fernando				
Burbank	91504	5	6	5 south
Chatsworth	91311	11	10	5 north/118 west
San Fernan.	91342	5	5	5 north
Sepulveda	91406	5	5	surface,5 north/ 405 south
Central: I-5				
Commerce	90040	3	2	5 south
Santa Fe Sp	90670	4	0	surface
Vernon	90058	3	2	surface
Central: I-405				
Compton	90221	4	2	91 east,710 north
Culver Cty	90230	10	5	10 west
El Segundo	90245	8	2	405 north(CF:405 south)
Gardena	90248	2	3	surface
Redondo Bch	90278	6	3	405 north/surface
San Pedro	90731	6	5	110 south,surface
Santa Mon.	90404	13	8	10 west
Central: East County				
Azusa	91702	10	3	605 north
El Monte	91733	5	4	605 north
La Puente	91746	6	4	605 north/60 east
Orange				
Fullerton	92631	4	4	55 north
Hunt. Bch	92649	11	5	22 west
Irvine	92714	8	6	405 north
East				
Pomona	91766	8	6	10 west
Rancho Cuc.	91730	8	7	10 west,surface
Riverside	92507	7	6	215 south

+ Average straight-line distance from zip code to zip code.
@ Straight-line distance for Consolidated Freightways, zip to zip. Zero value indicates that terminal is located in manufacturing zip code.
* Most common access route among seven LTL carriers

LOCATIONS FOR NEW TERMINALS

Strategic terminal siting provides an opportunity for LTL carriers to reduce their congestion delays. In this section, positions for new terminals are examined in relationship to the existing manufacturing base and in relationship to freeway congestion. Because all carriers have made substantial investments in their existing facilities, and because additional terminals would require bigger investments, neither change could come quickly. Nevertheless, the cost of constructing a terminal is far less than the cost of constructing other industrial buildings, such as factories or even warehouses. So, establishing new terminals would not be impossible.

Servicing Entire Region from a Single Terminal

South central Los Angeles is the most central location from which to serve the entire Los Angeles region. It offers the minimum average travel distance to the region's manufacturing employment, and it offers good freeway accessibility to all but San Fernando Valley and west Los Angeles–El Segundo. It effectively exploits surplus freeway capacity heading south, to the ports and to the Santa Ana corridor.

An analysis of existing terminals indicates that a large number of motor carriers have wisely selected south central Los Angeles for their terminals. The only concern is that future carriers will avoid the area because of high land costs and the lack of vacant industrial land. In its 1989 analysis of the Los Angeles industrial real estate market, Grubb and Ellis Realty (*15*) stated the following:

> The scarcity of available land and the area's high prices have combined to keep construction activity low. The small amount of development activity that does occur consists primarily of teardowns of old structures, particularly multi-story industrial warehouse space, which is considered unsafe by earthquake or fire standards.

One alternative to south central Los Angeles is the Santa Fe Springs-La Mirada area, to the southeast, where there is some vacant industrial land. This area offers improved access to the fastest growing areas in the region: Orange, Riverside, and San Bernadino counties. San Fernando Valley, on the other hand, can only be reached by the circuitous I-605–I-210 route or through the congested downtown via the Santa Ana Freeway.

A second alternative is the Carson-Dominguez area to the south. It offers excellent access to the ports and the San Diego corridor, and good access to south central Los Angeles, via the Long Beach Freeway (I-710). Vacant industrial property is available, and prices are relatively low. The biggest drawbacks would be difficult access to San Fernando Valley and increased distance from the growing Riverside and San Bernadino counties.

In the future, lower land prices and vacant property may draw carriers even farther from the center, to East Los Angeles County or San Bernadino and Riverside counties. This shift would certainly create congestion; average travel distances would increase and trucks would be forced to travel in the same direction as commuters.

With the lack of vacant land in south central Los Angeles and the absence of a direct route to San Fernando Valley, carriers may find it impossible to serve the entire region from a single location in the future. Independent of freeway congestion, the size of the Los Angeles region presents a challenge. The urbanized area encompasses a region larger than the state of Delaware and a population bigger than that of Pennsylvania.

Multiple Terminals

In this section, multiple terminals opportunities in the four Los Angeles regions—San Fernando, Eastern, Orange County, and Central—are examined.

San Fernando

All seven LTL carriers currently serve the valley from the Pacoima–Sun Valley location. This location places the carriers strategically at the intersection of the Golden State (I-5) and Hollywood (I-170) freeways and near the manufacturing center of the valley. It would be difficult to improve on the current location.

Orange County

If a carrier can only provide one terminal in Orange County, then Orange—used by all seven LTL carriers surveyed—is a reasonable choice. Orange provides immediate access to the I-5–91–57 triangle and close access to the manufacturing center in Irvine. However, given the size of the Orange County market and the congestion on roadways heading toward Irvine, two terminals are warranted, with the second in Irvine.

Eastern

The concern in Eastern is not so much with existing congestion but with future congestion. The Ontario-Fontana area, along the Devore Freeway (I-15), is prime for development, and the larger carriers have wisely established terminals there. Because the area is large, reasonable locations for a pair of terminals are San Bernadino and Fontana.

Central

The central region contains at least three distinct manufacturing centers—San Diego corridor, Santa Ana corridor, and East Los Angeles County—which suggests that a minimum of three terminals is needed to serve the region.

San Diego Service to the northern part of the corridor would improve if the smaller carriers moved farther north to the Gardena-Torrance area (as Viking has already done). Gardena provides good reverse commute access to the ports and improved access to El Segundo, Los Angeles Airport, and Santa Monica. Ideally, the corridor would be served by two terminals, one to the north in the Inglewood area and one to the south in the Carson–Long Beach area.

Santa Ana Vernon-Commerce in south central Los Angeles provides immediate access to the industrial core and the downtown and reverse commute access to Santa Fe Springs and South Gate. Unfortunately, the lack of vacant land is an obstacle to locating there, and some carriers have moved farther south. Ideally, two terminals would be used, with the second in the Santa Fe Springs area.

East County El Monte, near the intersection of the San Gabriel (I-605) and Pomona (60) freeways, provides excellent reverse commute access to the manufacturing centers of Azusa-Irwindale and Industry–La Puente.

Addition of Terminals

The ideal way to avoid road congestion would be for motor carriers to establish more terminals, closer to their customers. In this regard, Consolidated Freightways, with 16 terminals, serves as a model. Although other individual carriers may not have the volumes to justify 16 terminals, they could justify more terminals if they pooled their traffic. Following are some possibilities:

- Jointly operated terminals for several LTL carriers, to minimize the investment of each;
- A pooled fleet of pickup and delivery trucks to be shared by carriers; and
- Independently operated terminals to serve multiple LTL carriers, with terminals acting as agents for pickup and delivery (LTL carriers would be responsible for line haul between terminals).

In this era of deregulation, trucking is a highly competitive but fragmented industry, with many small operators. Although shippers have benefited from low prices, some efficiency has been lost. From the standpoint of highway congestion, a reduction in the number of competitors, or cooperative agreements among competitors, would allow carriers to strategically locate terminals in more locations and to reduce truck travel over congested roads. This direction should be encouraged.

CONCLUSIONS

Los Angeles has been called a city without a center. Yet it has a fast-growing downtown that influences traffic patterns throughout the city. Los Angeles has also been called a city with perpetual traffic congestion. Yet most freeway segments only operate at capacity for a few hours each day, in one direction at a time. Combined, these factors indicate that even automobile-dominated cities, such as Los Angeles, have surplus road capacity ready to be exploited. By strategically locating terminals to make trucks act as reverse commuters, and by strategically locating terminals to reduce travel distance, congestion-related delays can be reduced for trucks and automobiles alike.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to Paul Chow, Fred Gey, Don Bain, Ned Devlin, Joe Finney, Mike Jordan, and the California Trucking Association, all of whom provided data for this paper.

REFERENCES

1. *Urban Freeway Gridlock Study*. Cambridge Systematics, Inc., Cambridge, Mass., 1988.
2. R. F. Teal. *Estimating the Full Economic Costs of Truck Incidents on Urban Freeways*. Institute of Transportation Studies, University of California, Irvine, 1988.
3. D. Gerard and R. Wunderlich. Truck Operations on Arterial Streets. *Strategies to Alleviate Traffic Congestion*, Institute of Transportation Engineers, Washington, D.C., 1988.
4. R. W. Hall and W. H. Lin. *LTL Trucking in Los Angeles: Congestion Relief Through Terminal Siting*. Working Paper 43. University of California Transportation Center, Berkeley, Oct. 1990.
5. *1989 Traffic Volumes on California State Highways*. Division of Traffic Engineering, Business, Transportation and Housing Agency, State of California, Sacramento, 1988.
6. *1987 California State Highway Log, District 7 and 12*. Office of Traffic Engineering, Department of Transportation, Business and Transportation Agency, State of California, Sacramento, 1987.
7. *Distribution of Manufacturing Employment in the Los Angeles Five County Area*. Western Economic Research Co., Encino, Calif., 1988.
8. *Routing and Zip Code Directory, 1989–1990*. ABF Freight System, Inc., Fort Smith, Ark.
9. *1989 Service Guide*. ANR Freight System, Inc., Denver, Colo.
10. *Customer Service Guide*. Consolidated Freightways, Menlo Park, Calif., Jan. 1990.
11. *Service Directory*. P.I.E. Nationwide, Jacksonville, Fla., March 1988.
12. *1990 Routing Guide*. Roadway, Akron, Ohio.
13. *Direct Points via Viking*. Viking Freight System, San Jose, Calif., 1991.
14. *1990 Service Guide*. Yellow Freight System, Overland Park, Kans.
15. *Los Angeles Basin Real Estate 1989*. Grubb and Ellis Realty, Los Angeles, Calif., 1989.

Publication of this paper sponsored by Committee on Urban Goods Movement.

ALINEA: A Local Feedback Control Law for On-Ramp Metering

Markos Papageorgiou, Habib Hadj-Salem, Jean-Marc Blosseville

ALINEA, a new local traffic-responsive strategy for ramp metering, is presented. The new control strategy is based on a feedback structure and is derived by use of classical automatic control methods. ALINEA is a simple, robust, flexible, and effective local strategy for ramp metering. Real-life results from application of the new strategy to a single on-ramp of the Boulevard Périphérique in Paris are provided.

Traffic-responsive on-ramp metering systems are currently operating at several freeways in the United States (*1*). Furthermore, on-ramp metering has been tested or introduced in some freeways in France (*2*), Italy (*3*), West Germany (*4*), New Zealand, (*5*), United Kingdom (*6–8*), and the Netherlands (*9*).

More or less sophisticated coordinated on-ramp control strategies have been proposed and tested by simulation (*10,11*), and some installations have been reported. Nevertheless, operating ramp metering systems are mainly of a local, traffic-responsive character. Real-time measurements used in local control strategies are provided by loop detectors that are located in the vicinity of the corresponding on-ramps.

An extensive review of local traffic-responsive strategies was provided by Masher et al. (*1*). Strangely enough, many local strategies presented or implemented so far are based on a disturbance compensation feedforward philosophy. However, direct disturbance compensation is known to lead to fairly sensitive control strategies. In fact, in the last 40 years, an independent scientific discipline, automatic control, has evolved from the need to eliminate disturbances in a robust and efficient way by use of the basic notion of feedback.

A new local, traffic-responsive strategy for ramp metering applying a feedback control structure is presented. The feedback law is derived by use of classical methods of automatic control theory and is characterized by remarkable simplicity, high efficiency, and robustness. The main aim is to thoroughly present the theoretical background for the strategy's development and to discuss qualitative features such as simplicity, robustness, and flexibility. On the other hand, the new strategy for ramp metering has been tested and compared to five known strategies during an experimentation period of 6 months on the Boulevard Périphérique in Paris. Extensive results of this experiment were reported by Hadj-Salem et al. (*2,12*), and are not included. More recently, ALINEA has been tested during an experimentation period of several months on the ramp of the Coen Tunnel near Amsterdam. Results of this investigation were reported by Middelham and Smulders (*13*).

SOME GENERAL CONTROL STRUCTURES

Consider a process under control (e.g., house heating), and a selected output (e.g., inner temperature), shown schematically in Figure 1. The process is affected by some process inputs. Process inputs that can be manipulated are called controllable inputs, or simply inputs (e.g., heating valves), whereas process inputs that cannot be manipulated are called disturbances (e.g., outer temperatures). Disturbances may be predictable or nonpredictable, measurable or nonmeasurable, etc. The control problem is to appropriately select the controllable inputs so as to achieve—despite the impact of disturbances—a desired process output value called the "set value" (e.g., 20°C).

Assume that a mathematical model of the process is available, and furthermore that all essential disturbances are time variant but measurable. Then, at any instant of time, inputs can be calculated on the basis of the process model using disturbance measurements to achieve a given set value (Figure 1a). This feedforward control procedure is broadly known as disturbance compensation. Because of inevitable inaccuracies of mathematical models and occurrence of other unexpected, nonmeasurable disturbances, disturbance compensation is known to be a particularly sensitive control structure.

In the heating example, disturbance compensation would correspond to measuring the outer temperature and controlling the heating valves to achieve a constant inner temperature of, say, 20°C.

Consider a traffic flow process around an on-ramp as shown in Figure 2. Assuming absence of congestion, the downstream traffic volume q_{out} may be declared as the process output with a set value $\hat{q}$ (e.g., equal to capacity), whereas the on-ramp volume r is identified as the controllable input and the upstream traffic volume q_{in} as a measurable disturbance. In order to keep q_{out} near $\hat{q}$, an intuitive way to do it is to calculate $r = \hat{q} - q_{in}$ using current measurements of the disturbance q_{in}. Obviously, this feedforward procedure, which is essentially applied by many known local methods for ramp metering, corresponds to the disturbance compensation of Figure 1a. The quality of results depends on the accuracy of the applied process model, but complicated models lead to complicated control algorithms. However, even for highly complex strategies, sensitivity with respect to existing inaccuracies and unexpected disturbances remains a structural drawback.

M. Papageorgiou, Lehrstuhl für Steuerungs- und Regelungstechnik, Technische Universität München, P.O. Box 20 24 20, D-8000 München 2, West Germany. H. Hadj-Salem and J.-M. Blosseville, Institut National de Recherche sur les Transports et leur Sécurité, Département Analyse et Régulation du Trafic, 2, av. du Gl. Malleret Joinville, 94114 Arcueil Cedex, France.

sumed to be constant during the time interval $[(k - 1)T, kT]$. The constant parameter β results from the discretization procedure to be $\beta = \exp(-\hat{Q}'\alpha T/\delta)$.

The parameter β may be neglected if the ratio δ/T is sufficiently small. This will be the case if traffic volume entering the freeway reaches Site 2 during the time interval T. In this case, the effect of entering traffic $r(k)$ will be visible at Site 2 by the end of the corresponding time interval. Setting $\beta = 0$ in Equation 5,

$$\Delta o_{\mathrm{out}}(k + 1) = [\Delta q_{\mathrm{in}}(k) + \Delta r(k)]/\hat{Q}' \qquad (6)$$

In the next section, a feedback law is developed under the assumption $\beta = 0$. An extension applying to the case of higher ratios δ/T will be derived later. Finally, a complementary disturbance compensation mechanism will be presented.

Regulator Design

An appropriate feedback law for the process of Equation 6 is given by the following integral regulator

$$r(k) = r(k - 1) - K_{\mathrm{R}}[\hat{o} - o_{\mathrm{out}}(k)] \qquad (7)$$

where K_{R} is a constant and positive regulator parameter. After applying this regulator to Equation 6, the following z-transfer function is obtained for the closed-loop system:

$$H(z) = o_{\mathrm{out}}(z)/\hat{o} = \frac{K_{\mathrm{R}}/\hat{Q}'}{z - 1 + (K_{\mathrm{R}}/\hat{Q}')} \qquad (8)$$

A time-optimal deadbeat regulator is obtained by choosing $K_{\mathrm{R}} = \hat{Q}'$. Because of the sign of $\hat{Q}'$, the linearization of Equation 6 is strictly valid only on the left-hand side of the fundamental diagram. However, even for congested traffic the feedback law of Equation 7 leads to traffic occupancy reduction and can thus be applied in the same way.

Extension for Bottleneck Further Downstream

If Site 2 is located at a bottleneck further downstream and if T is chosen accordingly short, the entering traffic may not reach Site 2 at the end of each time interval, in which case β cannot be neglected. Applying the regulator of Equation 7 to the original process model of Equation 5, the closed-loop z-transfer function becomes

$$H(z) = o_{\mathrm{out}}(z)/\hat{o} = \frac{z(1 - \beta)}{z^2 - 2\beta z + \beta} \qquad (9)$$

with the eigenvalues $\zeta_{1,2} = \beta \pm j[\beta(1 - \beta)]^{1/2}$. Although the closed-loop system remains stable for any $\beta < 1$, the transient behavior may be slow. An amelioration may be achieved by application of the following proportional-plus-integral feedback law

$$r(k) = r(k - 1) - K_{\mathrm{R}}[\hat{o} - o_{\mathrm{out}}(k)] - K_{\mathrm{p}}[o_{\mathrm{out}}(k) - o_{\mathrm{out}}(k - 1)] \qquad (10)$$

where K_{p} is a further constant and positive regulator parameter. A time-optimal deadbeat regulation is achieved by the choice $K_{\mathrm{R}} = \hat{Q}'$, $K_{\mathrm{p}} = \beta\hat{Q}'/(1 - \beta)$, as can be readily demonstrated. In summary, a further term has been added in Equation 9 as compared with Equation 7 for the particular case of, say, $\delta/T > \alpha\hat{Q}'$.

Elimination of Constant Disturbances

The regulator of Equation 7 is capable of eliminating constant disturbances. The z-transfer function of the feedback law is

$$\frac{r(z)}{-\Delta o_{\mathrm{out}}(z)} = \frac{K_{\mathrm{R}}z}{z - 1} \qquad (11)$$

Considering the process model of Equation 6 with $K = 1/\hat{Q}'$, the corresponding closed-loop linearized system is shown in Figure 3. Assume a disturbance d as indicated in Figure 3. The z-transfer function o_{out}/d is given by

$$\frac{o_{\mathrm{out}}(z)}{d(z)} = \frac{K(z - 1)}{z(z - 1 + KK_{\mathrm{R}})} \qquad (12)$$

If d is constant, this equation yields in the steady-state (i.e., for $z = 1$) $o_{\mathrm{out}}/d = 0$, i.e., any constant disturbance is eliminated in the steady state by the control system.

Because the upstream traffic volume q_{in} acts as the disturbance d in Figure 3, this statement holds true for constant (or slowly varying) upstream traffic volumes.

Elimination of Biased On-Ramp Volume Realization

What happens if the implemented on-ramp volume is biased as compared with the on-ramp volume ordered by the regulator? A bias in the on-ramp volume realization corresponds exactly to the disturbance d of Figure 3. Hence, the bias is automatically eliminated by the control system.

This statement does not hold true if the implemented on-ramp volume is used for the retarded value $r(k - 1)$ in the feedback law of Equation 7. Figure 4 shows the signal flow diagram of the closed-loop system in this case. The z-transfer function is now given by

$$\frac{o_{\mathrm{out}}(z)}{d(z)} = \frac{K}{z - 1 + KK_{\mathrm{R}}} \qquad (13)$$

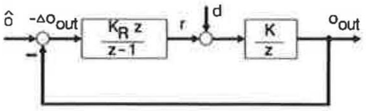


FIGURE 3 Linearized closed-loop system.

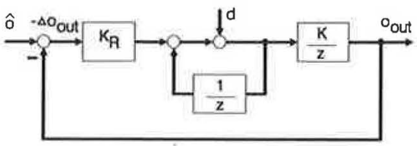


FIGURE 4 Modified closed-loop system.

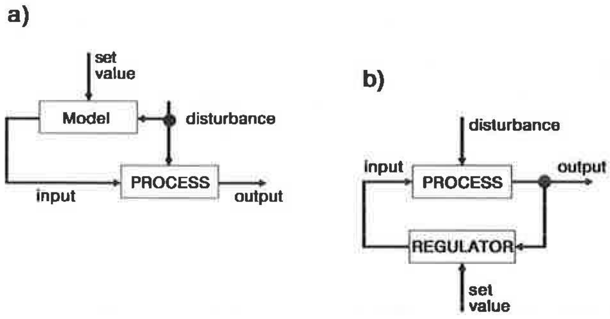


FIGURE 1 (a) Disturbance compensation and (b) feedback control.

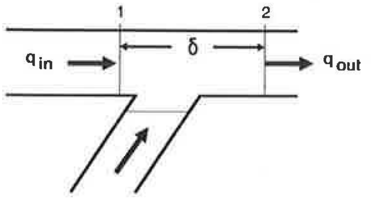


FIGURE 2 A traffic flow process.

Moreover, adjustment of threshold parameters during implementation becomes a difficult task for complicated strategies.

A much more elegant, robust, simple, and efficient way of solving the control problem is to introduce a feedback structure (Figure 1b). The measurable output is fed back, and the controllable input is permanently modified by an appropriate regulator to keep the output near its set value despite the influence of time-variant disturbances. Design of the regulator may be performed by use of well-known automatic control methods. Because of its feedback structure, a control system of this kind is much more precise and much less sensitive with respect to model inaccuracies and unexpected disturbances as compared with disturbance compensation.

In the heating example, feedback control corresponds to measuring the inner temperature and modifying the heating valves accordingly to achieve the desired set value despite the variations of the outer temperature.

The feedback methodology can now be transmitted to the problem of ramp metering shown in Figure 2. Because the same feedback law is to be applied both for congested and for free flowing traffic, it is preferable to consider occupancy o_{out} as an output variable instead of q_{out}. This is because traffic volume may have the same values both for light and congested traffic because of the characteristic form of the fundamental diagram. The corresponding set value $\hat{o}$ for traffic occupancy may be easily found on the basis of the fundamental diagram at the output line; alternatively, a desired downstream occupancy value may be provided directly.

An additional advantage of choosing o_{out} rather than q_{out} as an output variable arises from the fact that the critical occupancy o_{cr} seems to be less sensitive with respect to weather conditions and other operational influences compared with the capacity q_{cap} of a freeway stretch. This statement is supported by data material provided by Keen et al. (*14*). As a consequence, considering the set value $\hat{o} = o_{cr}$ is a more robust way of achieving capacity flow than considering $\hat{q} = q_{cap}$ because variations of q_{cap} caused by environmental or other conditions are stronger compared with variations of o_{cr}.

Thus, the next step is to derive a feedback control law $r = R(\hat{o}, o_{out})$ according to Figure 1b to keep o_{out} near $\hat{o}$. Derivation of this feedback law is the subject of the next section.

DERIVATION OF THE FEEDBACK LAW

Modeling

Consider the traffic flow process shown in Figure 2. Site 1 is assumed to be situated just upstream of the on-ramp. Site 2 is situated downstream of the on-ramp, at a distance δ from Site 1. Assume that the on-ramp volume r is updated every T time units, where $T = 10$ sec, . . . 1 min, or more. The conservation equation for the freeway stretch between Sites 1 and 2 (Figure 2) is

$$\dot{\rho}(t) = [q_{in}(t) + r(t) - q_{out}(t)]/\delta \tag{1}$$

where the traffic density ρ (veh/km) is defined as the number of cars included in the stretch, divided by length δ, t being the time argument. Because traffic density is not readily measurable, it is convenient to replace $\rho(t)$ in Equation 1 by the occupancy $o_{out}(t)$ using the approximate relationship $\rho = \alpha o_{out}$, where $\alpha = \mu/100\lambda$), μ being the number of lanes of the mainstream and λ being the mean effective vehicle length in kilometers.

Assuming $q_{out}(t)$ is given as a nonlinear function of $o_{out}(t)$ (fundamental diagram)

$$q_{out}(t) = Q[o_{out}(t)] \tag{2}$$

and substituting Equation 2 into Equation 1, a nonlinear first-order dynamic system model is obtained. This model may be linearized around a nominal steady state $(\bar{o}_{out}, \bar{q}_{in}, \bar{r})$ such that

$$\bar{o}_{out} = \hat{o};\ \bar{q}_{out} = Q(\bar{o}_{out});\ \text{and}\ \bar{r} = \bar{q}_{in} - \bar{q}_{out} \tag{3}$$

With the notation $\Delta o_{out}(t) = o_{out}(t) - \bar{o}_{out}$ used analogously for all variables, the linearization yields

$$\Delta\dot{o}_{out}(t) = [\Delta q_{in}(t) + \Delta r(t) - \hat{Q}'\Delta o_{out}(t)]/(\delta\alpha) \tag{4}$$

where $\hat{Q}' = dQ(\hat{o})/do_{out}$, i.e., $\hat{Q}'$ is the slope of the tangent of the fundamental diagram at $\hat{o}$ and hence its value is proportional to the speed of the corresponding kinematic traffic wave (*15*). Clearly, $\hat{Q}'$ is positive on the left-hand side of the fundamental diagram.

Because the control input r is updated every T time units, time discretization of Equation 4 with sample time interval T yields

$$\Delta o_{out}(k + 1) = \beta\Delta o_{out}(k) + [(1 - \beta)/\hat{Q}'] \times [\Delta q_{in}(k) - \Delta r(k)] \tag{5}$$

where $k = 0, 1, 2, \ldots$ is the sample time index. Thus $o_{out}(k)$ is the occupancy at time kT, and $\Delta q_{in}(k)$ and $\Delta r(k)$ are as-

If d is constant, this equation yields in the steady state $o_{\text{out}}/d = 1/K_{\text{R}}$, i.e., a bias $\bar{d}$ leads to a steady-state error (offset) of $\bar{d}/K_{\text{R}}$.

DISCUSSION OF THE FEEDBACK CONTROL LAW

The Regulator

The proposed feedback control law to be applied at the time instants kT, $k = 0, 1, 2, \ldots$, for any sample interval T (e.g., $T = 60$ sec), is

$$r(k) = r(k-1) + K_{\text{R}}[\hat{o} - o_{\text{out}}(k)] \tag{14}$$

where K_{R} is a positive, constant regulator parameter and $o_{\text{out}}(k)$ is the current occupancy measurement. The feedback law of Equation 10, which is much simpler than any other local metering strategy, was given the name asservissement linéaire d'entrée autoroutière (ALINEA).

Heuristic Interpretation

Equation 10 suggests a fairly plausible control behavior. If the measured occupancy $o_{\text{out}}(k)$ at time k is found to be lower (higher) than the desired occupancy $\hat{o}$, the second term of the right-hand side of Equation 10 becomes positive (negative) and the ordered on-ramp volume $r(k)$ is increased (decreased) as compared to its last value $r(k - 1)$. Clearly, the feedback law of Equation 10 acts in the same way both for congested and for light traffic (no switchings are necessary).

Note that some occupancy control strategies (*1*) react to excessive occupancies only after congestion is created and has reached an upstream measurement location, whereas ALINEA reacts smoothly even to slight differences $\hat{o} - o_{\text{out}}(k)$ and may thus prevent congestion in an elegant way.

On the other hand, some demand-capacity strategies react to excessive downstream occupancies only after a threshold value is exceeded. Typically, and in contrast to ALINEA, the reaction of these control strategies to excessive occupancies is rather crude: on-ramp volumes are set equal to their minimal values. In this way, a nonnecessary underload of the freeway may occur. On the contrary, the essential effect of ALINEA is to stabilize traffic flow at a high throughput level and eventually to reduce the risk of a breakdown without underloading the freeway.

The Value of K_{R}

A value of $K_{\text{R}} = 70$ veh/hr was found to yield good results in real-life experiments. K_{R} is the only parameter to be adjusted in the implementation phase because no thresholds or other constants are included in Equation 10. Moreover, from theoretical considerations,

- Results are insensitive for a wide range of K_{R} values;
- Increasing (decreasing) K_{R} values lead to stronger (smoother) reactions of the regulator, and regulation times get shorter (longer);
- For extremely high values of K_{R}, the regulator may have an oscillatory, unstable behavior.

In view of these statements, real-life calibration of the unique free parameter K_{R} is—if at all necessary—particularly easy.

Measurements

ALINEA requires only one detector station that measures occupancy o_{out} downstream of the merge area. This is equal or less than the measurement requirements of other local strategies. The measurement location should be such that a congestion originating from excessive on-ramp volumes be visible in the measurements. A distance of 40 m downstream of the on-ramp nose was found to be adequate at the Boulevard Périphérique in Paris. The strategy was also found to work adequately with a distance of 400 m in Amsterdam (*13*).

Disturbance Reduction

If upstream traffic volume q_{in} is constant, then the feedback law of Equation 10 is easily demonstrated to lead to $o_{\text{out}} = \hat{o}$ in the steady state. In other words, whatever the value of a constant (and not measured) upstream traffic volume might be, the feedback law leads occupancy to its desired value.

Similarly, if upstream traffic volume q_{in} is perturbed around a constant or slowly varying average, the feedback law of Equation 10 keeps downstream occupancy o_{out} close to $\hat{o}$ in the average. On the other hand, rapid oscillations of q_{in} around an average value are only slightly reduced by the control system.

Embedding in a Coordinated Control System

Realization of the feedback law of Equation 10 requires a set value $\hat{o}$ be provided by the user. The set value may be changed any time it is needed and hence, the feedback law may be embedded in a simple and natural way into a hierarchical control system with set values of the individual sections being specified (and changed in real time) by a superior coordination level or by an operator.

Restrictions, Override Tactics

The on-ramp volume values resulting from Equation 10 may be limited if some maximum or minimum values are exceeded. Moreover, override tactics (e.g., for preventing interference of the on-ramp queue with surface traffic) may be applied. When either a limitation or override tactic becomes active, the last on-ramp volume $r(k - 1)$ required in the feedback law of Equation 10 for calculating $r(k)$ should correspond to the actual number of cars that entered the freeway (because of the limitation or override) and not to the calculated but suspended value provided by Equation 10 in the last time interval.

Realization of On-Ramp Volumes

ALINEA is compatible with any kind of realization of the required on-ramp volumes (one-by-one, platoon, traffic cycle, etc.). Obviously, the implemented on-ramp volumes should be equal to the on-ramp volumes ordered by the control law of Equation 10 in the average unless a limitation becomes active according to the previous section. If this is not true, e.g., if for any reason the implemented on-ramp volumes are biased as compared with the calculated on-ramp volumes, then the control system is capable of eliminating the bias automatically. For this to be true, the value of $r(k-1)$ required in Equation 10 should be equal to the on-ramp volume calculated by Equation 10 in the last time interval and not equal to the on-ramp volume that actually entered the freeway, unless an override or a constraint has become active, as already mentioned.

Efficacy

A preliminary version of ALINEA and some popular previous control strategies have been implemented and tested on an on-ramp of the Boulevard Périphérique in Paris during an experimentation period of 6 months. Results of this lengthy experimentation period showed a clear superiority of ALINEA in preventing congestion and increasing traffic throughput as compared to other local traffic responsive strategies. Details were reported by Hadj-Salem et al. (*2,12*). More recent field results from the Netherlands confirm the superiority of ALINEA as compared with the demand-capacity type of strategies (*13*).

Extensions

Two possible extensions may be envisaged if the ratio δ/T is relatively high. This condition may hold if the output occupancy o_{out} is measured at a site far downstream of the on-ramp nose or the sample time interval is short (e.g., $T = 10$ sec). As a result of the theoretical development, these extensions are recommended if $\delta/T > \alpha K_R$, where α was defined earlier.

The first extension is to use the feedback law of Equation 9 rather than that of Equation 10 with the parameter value $K_p = \delta/(T\alpha) - K_R > 0$. Thus, even with this extension, K_R remains the only parameter to be calibrated during implementation.

The second extension for the case $\delta/T > \alpha K_R$ is to add to the feedback laws of Equations 9 or 10 the term $\gamma[q_{\text{in}}(k) - q_{\text{in}}(k-1)]$ if the value $q_{\text{in}}(k)$ can be predicted accurately enough by upstream measurements. This second extension corresponds to a disturbance compensation aiming at improving the feedback efficiency. The positive smoothing parameter $\gamma \leq 1$ should be appropriately chosen to avoid excessive oscillations of the on-ramp volumes caused by noise of the q_{in} measurements.

However, if the output Site 2 is being located far downstream of the on-ramp (at a downstream bottleneck) without any intermediary measurements, there is a risk of congestion being built up (because of incidents or other disturbances) in the interior of the stretch (1,2) without being visible at the output measurement Site 2. In such cases, the feedback (or any other) control system may be of little help for eliminating the congestion if no additional measurement stations are added.

REAL-LIFE APPLICATION RESULTS

ALINEA was first implemented and tested at the on-ramp Brançion of the Boulevard Périphérique in Paris. On-ramp volumes are realized on the basis of a traffic cycle of 40 sec with a minimum green phase of 10 sec. The set value $\hat{o}$ for downstream occupancy is 29 percent, which corresponds to capacity flow. Override tactics are applied to avoid interference of the on-ramp queue with the surface street traffic.

The on-ramp volumes ordered by the feedback law of Equation 10 are transformed into corresponding green-phase durations by use of a saturation flow value that is estimated in real time. The particular algorithm used for estimation of the saturation flow leads to a biased implementation of the ordered on-ramp volumes. The average difference (bias) between ordered and implemented on-ramp volumes is 2 veh/cycle or −180 veh/hr (which is completely independent of ALINEA). The closed-loop system was implemented as indicated in earlier sections (see also Figure 3) with $K_R = 70$ veh/hr. The bias of −180 veh/hr acts as an additional disturbance that is automatically eliminated by the feedback system.

Typical results of a 30-min time period are shown on Figures 5–7. Figure 5 shows the upstream and downstream occupancies and the periods of active minimum (10 sec) or maximum (40 sec) green-phase constraints. Figure 6 shows the corresponding upstream and downstream traffic volumes. Figure 7 shows the ordered and implemented on-ramp volumes, and the corresponding green phase duration.

Results of Figure 5 indicate a proper functioning of ALINEA: on-ramp volumes ordered by the feedback law of Equation 10 keep the downstream occupancy o_{out} near its set value in the average, despite the variation of upstream traffic volume and despite the bias in the on-ramp volume realization.

Clearly, when constraints become active because of lacking demand (green phase equal to 40 sec) or because of traffic

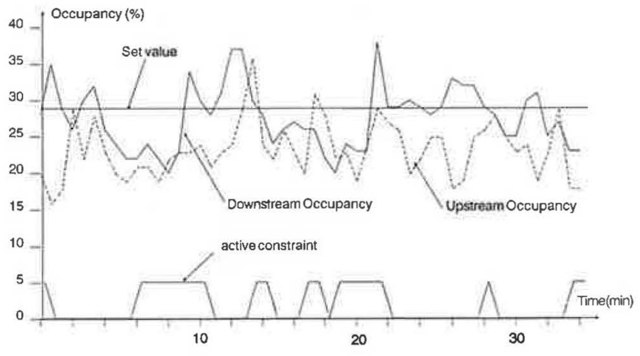


FIGURE 5 Occupancies and constraints activation.

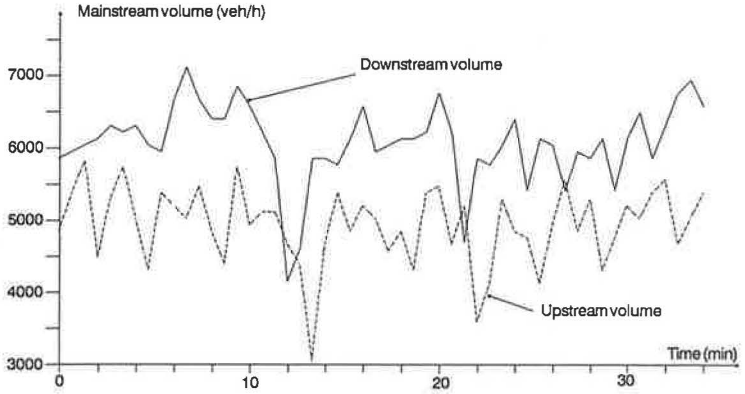


FIGURE 6 Mainstream traffic volumes.

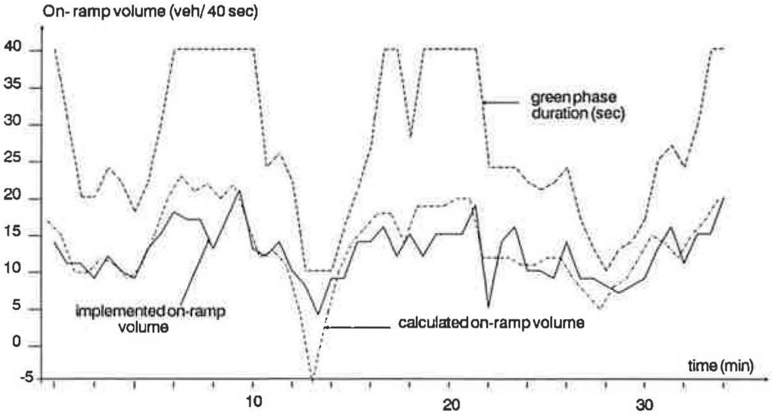


FIGURE 7 On-ramp volume and green-phase duration.

shock waves arriving from downstream (green-phase duration equal to 10 sec), some deviations from the set value may be observed. Figure 6 shows that the downstream traffic volume attains values around the capacity (6,300 veh/hr) of the three-lane freeway stretch. Thus the feedback law acts elegantly to avoid breakdowns and to fully utilize the freeway capacity on the basis of one single equation (Equation 10). Extended long-term results are reported elsewhere (*2,12,13*).

ALINEA is currently operational at the on-ramps Bran- çion, Chatillon, and Italie of the Boulevard Périphérique in Paris. Its implementation on a number of further freeway on-ramps of the Paris region is currently under way. Moreover, ALINEA has been tested at the on-ramp of the Coen Tunnel in the Netherlands.

CONCLUSION

ALINEA, a new local traffic-responsive strategy for ramp metering has been presented. The new control strategy is based on a feedback structure, which was derived by classical automatic control methods. ALINEA

- Is simpler than other known algorithms,
- Requires a minimal amount of real time measurements (detectors),
- Is easily adjustable to particular traffic conditions because only one parameter is to be adjusted in a prescribed way,
- Has proven in real life experiments to be more efficient in preventing congestion and preserving capacity flow compared to some known algorithms,
- Can be embedded in a coordinated on-ramp control system in a natural way,
- Can be easily modified in case of changing operational requirements,
- Is highly robust with respect to inaccuracies and different kinds of disturbances, and
- Is theoretically supported by automatic control theory.

ACKNOWLEDGMENT

This work was partially supported by the DRIVE Programme of the European Commission.

REFERENCES

1. D. P. Masher, D. W. Ross, P. J. Wong, P. L. Tuan, H. M. Zeidler, and S. Petracek. *Guidelines for Design and Operation of Ramp Control Systems*. Stanford Research Institute, Menid Park, Calif., 1975.

2. H. Hadj-Salem, M. M. Davée, J. M. Blosseville, and M. Papageorgiou. *ALINEA: Un Outil de Regulation d'Accés Isolé sur Autoroute*. Rapport INRETS 80, Arcueil, France, 1988.
3. R. La Pera and R. Nenzi. TANA—An Operating Surveillance System for Highway Traffic Control. *Proc., Institute of Electrical and Electronic Engineers*, Vol. 61, 1973, pp. 542–556.
4. N. Helleland, W. Joeppen, and P. Reichelt. Die Rampendosierung an der A5 Bonn/Siegburg der BAB 3 in Richtung Köln. *Strassenverkehrstechnik*, Vol. 22, 1978, pp. 44–51.
5. NN: Ramp Metering in Auckland. *Traffic Engineering and Control*, Vol. 24, 1983, pp. 552–553.
6. NN: Ramp Metering on M6. *Traffic Engineering and Control*, Vol. 27, 1986, p. 32.
7. NN: Access Control on M6. *Traffic Engineering and Control*, Vol. 27, 1986, p. 469.
8. D. Owens and M. J. Schofield. Access Control on the M6 Motorway: Evaluation of Britain's First Ramp Metering Scheme. *Traffic Engineering and Control*, Vol. 29, 1988, pp. 616–623.
9. H. Buijn and F. Middelham. Ramp Metering Control in the Netherlands. *Proc., 3rd IEE International Conference on Road Traffic Control*, London, United Kingdom, 1990, pp. 199–203.
10. M. Papageorgiou. *Applications of Automatic Control Concepts to Traffic Flow Modeling and Control*. Springer-Verlag, New York, 1983.
11. M. Papageorgiou. *Freeway On-Ramp Control Strategies: Overview, Discussion, and Possible Application to Boulevard Périphérique de Paris*. Internal Report, INRETS, DART, Arcueil, France, 1986.
12. H. Hadj-Salem, J. M. Blosseville, and M. Papageorgiou. ALINEA: A Local Feedback Control Law for On-Ramp Metering—A Real Life Study. *Proc., 3rd IEE International Conference on Road Traffic Control*, London, United Kingdom, 1990, pp. 194–198.
13. F. Middelham and S. Smulders. *Isolated Ramp Metering: Real-Life Study in the Netherlands*. DRIVE-Office, Brussels, Belgium, 1991.
14. K. G. Keen, M. J. Schofield, and G. C. Hay. Ramp Metering Access Control on M6 Motorway. *Proc., IEE 2nd International Conference on Road Traffic Control*, London, 1986.
15. M. J. Lighthill and G. B. Whitham. On Kinematic Waves II. A Theory of Traffic Flow on Long Crowded Roads. *Proc., Royal Society, London*, Series A, Vol. 229, 1955, pp. 317–345.

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

Traffic Operation of Bicycle Traffic

HEIN BOTMA AND HANS PAPENDRECHT

Knowledge has been provided about one-directional traffic streams of bicycles and mopeds on cycle paths. This knowledge is useful for the perfection of the geometric design of bicycle facilities. Data on bicycle traffic were collected on rather heavily used paths at four locations in built-up areas and one in a rural area. A special developed system of sensors was used to measure times of passage, speeds, and lateral positions of the vehicles at a path cross section. Speeds of bicycles and mopeds had the same magnitude at the different town locations. They were of the same magnitude as measured 10 years ago. Only a weak relation was found between the rate of flow in platoons and mean speed. The distribution of the lateral positions was used to investigate the influence of the geometrics of the path's cross section. The average length of passing maneuvers and the percentiles of the lateral positions during passing were determined. On the narrow path, the length of the passing maneuver is shorter than that on the wider paths. Also, lateral clearance between paired cyclists is less on narrower than on wider paths. Paired cyclists keep less lateral distance to each other than do passers. As a function of volume, the frequency of passings conforms to the theory that describes free flow conditions. However, in absolute terms, the number of passings detected was lower. Estimated values of path capacities derived from the headway distribution are much larger than those mentioned in the literature.

Knowledge about traffic operation of bicycle traffic is not advanced. For instance, Chapter 14 of the Highway Capacity Manual of 1985 about bicycles consists of only four pages. Still, bicycle traffic is or can become an important transportation mode in many areas. In the Netherlands, the contribution of bicycle traffic to the modal split in medium-sized cities (50,000 to 200,000 inhabitants) during rush hours is 40 percent of total internal vehicle trips and is still slowly increasing.

In a Dutch study (*1*) that investigated the possibilities of promoting bicycle traffic by means of special facilities, a maximum contribution to the modal split of 55 percent was found. Despite this high contribution to mobility, little attention has been paid to fundamental research on traffic operation of bicycle traffic needed to base bicycle infrastructural designs and bicycle-promoting actions on a more scientific understanding. This study is a contribution towards this better understanding.

To describe the situation in the Netherlands, some data about car and bicycle ownership from the Dutch Statistical Bureau CBS (*2*) are presented in Table 1, which indicates

- Growing car and motorbike ownership,
- Decreasing and then stabilizing moped ownership, and
- Increasing and then stabilizing bicycle ownership.

The results of a study of bicycle traffic in one direction on a bicycle path accessible for bicycles and mopeds are described. In the Netherlands, mopeds are authorized to use such a path. In general, the proportion of mopeds is limited, but because of their higher speed the influence on bicycle path operation and on traffic safety can be substantial. Consequently, moped traffic was included in this study.

The study is based on experimental data collected at a path cross section at five different locations.

RESEARCH GOALS AND METHODS

In the Netherlands, bicycle paths are used by two types of vehicles: bicycles and mopeds. Mopeds can be described as light motorbikes and compared to bicycles have more mass and a higher speed. The official speed limit for mopeds is 30 km/hr inside built-up areas and 40 km/hr outside. However, there is little enforcement and these speed limits are not much respected.

Differences in speed between the two vehicle types and between vehicles of the same type lead to passing demand. As long as volume is low and the width of the path sufficient, the demand for passings can be easily satisfied. With increasing volume, it becomes more difficult to carry out the passings at any desired moment. The quality of operation decreases. In order to get a quantitative insight into this phenomenon, the passing demand and the passing possibilities must be studied. Passing possibilities depend on the space needed for passings and on the space available.

It is evident that lateral behavior is an important aspect of passing. So speeds, lateral positions, and lateral position during passing maneuvers were studied.

The quality of traffic operation and its relation to volume were also investigated at a macroscopic level. The questions to be answered are: What is the relation between mean speed and volume? and What is the capacity of a path in relation to its width?

Method

It was decided to collect the data from ordinary bicycle traffic and not to use test persons.

To get data on speeds and lateral positions, it is sufficient to make measurements at one cross section. Characteristics of the average shape of a passing maneuver could also be deduced from these local data.

In order to get sufficient interaction between path users, it was needed to select high-volume locations. To rule out interaction between effects of volume and path characteristics, locations were selected with more or less ideal geometric char-

Department of Transportation Planning and Highway Engineering, Faculty of Civil Engineering, Delft University of Technology, P.O. Box 5048, 2600 GA DELFT, the Netherlands.

TABLE 1 DATA ON VEHICLE OWNERSHIP IN THE NETHERLANDS

	1975	1980	1985	1986	1987	1988
population	13.600	14.100	14.450	14.530	14.620	14.710
passenger cars	3.399	4.515	4.901	4.950	5.118	5.251
motorbikes	68	103	128	127	131	135
mopeds	650	814	534	564	516	516
bicycles	8.600	10.580	11.179	11.517	11.441	11.695

acteristics (no curves and grades and a smooth pavement) and with no discontinuities immediately downstream or upstream. In fact, the only geometric characteristics varied were the width of the path and the height of the curb.

Measuring Device

The measuring device used was a specially designed mat in which several strings of tape switches were installed. Two strings at a longitudinal distance of 30 cm with switches of 50-cm length were used to measure moments of passage and speeds; a string with 16 switches of 15.6 cm was used to measure the lateral position more precisely.

To check the functioning of the device, a video registration of the bicycle traffic was made. A few samples of this registration were analyzed in detail to determine some parameters used in the analysis.

The mat had a maximum thickness of only 8 mm in the middle and had no perceivable influence on cyclists' behavior. It was visible but not obtrusive. The video camera and the van with the registration equipment were rigged in such a way that they were not visible for passing cyclists.

Data Collection

It proved to be difficult to find locations with the required geometric and volume characteristics. In particular, a location where the capacity of the path was reached could not be found. Eventually four locations inside a town were selected: one had a width of only 180 cm; three had widths of around 250 cm. In order to investigate high volumes, one exceptional location was added. It was at a cross section of a 3-m-wide bicycle path, which at the time was part of the route of a long-distance tour (230 km) for amateur cyclists. No mopeds were present in this case and speeds could be expected to be higher than those of ordinary bicycle traffic.

In the discussion of the results of the study, the five locations mentioned are indicated by T-Narrow, T1-Wide, T2-Wide, T3-Wide, and Tour.

ANALYSIS AND RESULTS

Volumes and Vehicle Composition

From an earlier investigation, it was known that bicycles and mopeds could be discriminated by speed and that a reasonable limit was 30 km/hr. That means a detected two-wheeled vehicle was registered to be a bicycle (bic) when its speed was less than 30 km/hr; otherwise it was registered as a moped (mop). Few bicycles have speeds higher than this limit and even less mopeds have speeds that are lower. The inevitable error was considered to be acceptable.

Measurements were carried out over a period of 3 hr in which a rush hour was included. Table 2 presents the detected number of vehicles, the vehicle composition, the average volume, and the rate of flow over the busiest half-hour. It can be noted that the mopeds were a small minority of the volume and that only the volumes of the Tour were rather high.

Distribution of Speeds

Table 3 presents the most important statistics of the speeds of bicycles and Table 4 those of the mopeds.

The speeds of bicycles were virtually the same at the four locations in towns. There seemed to be no effect of the width of the path. Generally speaking, the mean value is 19 km/hr

TABLE 2 VOLUMES AND VEHICLE COMPOSITION

Location	Total Number of veh	Number of Bic	Number of Mop	Mop %	Mean Volume veh/h	Rate of Flow busiest .5 h veh/h
T-Narrow	1,257	1,199	58	4.6	419	666
T1-Wide	1,792	1,693	99	5.5	600	864
T2-Wide	1,862	1,671	171	9.2	610	1,606
T3-Wide	1,510	1,481	29	1.9	503	1,034
Tour	8,860	8,860	-[1]	-[1]	2,953	3,328

[1] No mopeds present at location Tour

TABLE 3 STATISTICS OF BICYCLE SPEEDS

	Path		Speed			
	Width		Mean	St. dev.		85 %
Location	cm	Number	km/h	km/h	%	km/h
T-Narrow	180	1,199	19.6	3.4	17	22.8
T1-Wide	240	1,693	19.0	3.1	14	22.3
T2-Wide	250	1,691	19.0	2.9	15	21.8
T3-Wide	270	1,481	18.9	2.5	13	21.4
Tour	300	8,860	24.9	3.2	13	28.4

TABLE 4 STATISTICS OF MOPED SPEEDS

	Path		Speed			
	Width		Mean	St. dev.		85 %
Location	cm	Number	km/h	km/h	%	km/h
T-Narrow	180	58	36.9	4.4	12	42.5
T1-Wide	240	95	38.2	4.7	12	43.7
T2-Wide	250	171	39.9	4.9	12	45.0
T3-Wide	270	29	39.7	7.4	19	47.1

and the standard deviation around 3 km/hr. The speeds at the Tour are definitely higher, as was expected. A simple test indicated that the speeds were normally distributed.

The speeds of mopeds seemed to be influenced a little by the width of the path. The overall mean speed was around 38 km/hr and the standard deviation was about 5 km/hr.

An earlier study (*3*) carried out in 1979, with measurements at a comparable location, indicated approximately the same values of speeds of bicycles (a mean of 19.2 km/hr) and mopeds (a mean of 36.6 km/hr). Consequently, it can be concluded that speeds have not changed appreciably during the last 10 years.

Relation Between Mean Speed and Volume

A level of service (LOS) for bicycle traffic may be defined by using the mean speed as a criterion for the quality of the flow, as was done for motor vehicles in the HCM of 1960. This approach is feasible only when the mean speed varies with volume. In order to investigate this relation, it is necessary to differentiate between bicycles and mopeds. Both the volumes of bicycles and mopeds can affect the mean speeds of bicycles and the mean speed of mopeds.

For the analysis with linear regression, 5-min values were calculated for the following:

- Rates of flow of bicycles and mopeds; and
- Mean speeds of bicycles and of mopeds.

Mean speeds of both vehicle types were then related to both rates of flow. Some relations were statistically significant but only one was considered to be of any relevance; on the narrow path of 180-cm width a higher volume of bicycles leads to a lower mean speed of mopeds. Volumes had at most a minor effect on speeds for the range of volumes that could be observed (50 to 1,500 bic/hr).

Relation Between Mean Speed of Bicycles and Volume on a Smaller Time Scale

Closer inspection of the pattern of arrival within 5-min periods indicated that large gaps between vehicles occurred. This was an incentive to investigate a possible relation between rate of flow and mean speed using smaller intervals than the 5-min intervals, by which the vehicle population was split up before.

Groups of at least 10 bicycles were selected with headways of less than 5 sec, except for the first one. Such a group was defined as a platoon although it was realized that the word platoon for such a group would have a slightly different meaning than the word platoon used when describing a stream of cars. Bicycles with a headway of, say, 4 sec are still much freer in maneuvering than cars in the same situation, due to their greater lateral freedom.

A similar analysis was not possible for the detected mopeds because from a statistical point of view their number was too small to yield reliable results. In addition, on the narrow path only three platoons could be detected that fulfilled the conditions—too small a number for proper analysis. The analysis results for the three other town locations were sufficiently similar to allow the combination of the data to a single population. Their analysis led to Equation 1, which describes the mean speed U (km/hr) as a function of volume Q (bic/hr):

$$U = 20.8 - 6.8 \times 10^{-4}Q \qquad (1)$$

$$R^2 = 0.20 \qquad N = 66$$

where

R^2 = explained variance, and
N = number of observations.

The standard errors of the first and second coefficients, which were 3 and 25 percent, respectively, indicated their statistical significance.

Thus under conditions of high rates of flow, the mean speed decreases with volume as shown in Figure 1. No significant decrease in speed was found at location Tour, although rates of flow reached a maximum of 24,000 bic/hr as shown in Figure 2.

The finding that mean speed decreases with volume made it possible to estimate path capacity by fitting a model of volume Q as a function of density K (bic/km), resulting in Equation 2.

$$Q = 22.0K - 1.88 \times 10^{-2}K^2 \qquad (2)$$

$$N = 66$$

The standard errors of the first and second coefficients were 2 and 9 percent, respectively. Extrapolating this function, a capacity value of around 6,000 bic/hr was calculated as shown in Figure 3. However, it should be realized that this value is only an indication of the order of magnitude of bicycle path capacity.

Equations 1 and 2 are mathematically incompatible, in the sense that one equation cannot be derived from the other. However, they both model the finding that mean speed decreases with volume.

Lateral Positions

The lateral position of a vehicle has been defined as the distance to the right-hand edge of the path's pavement. Of special importance is the repelling effect of any curb on the lateral position. Cyclists tend to shy away from the curb, depending on its height. The effect of obstacles on the curb or adjacent to the path edge were not investigated in this study. Dutch design guidelines give values for this effect for various curb heights. However, so far they have been based on practical experience rather than on research data.

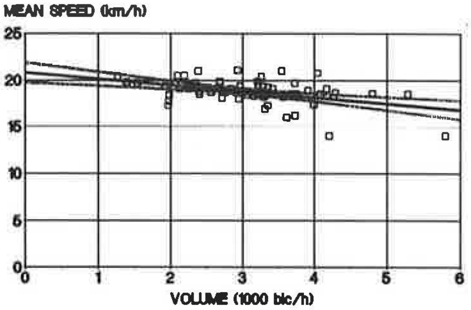


FIGURE 1 Mean speed as a function of volume for the three wide paths (linear relation and 95 percent confidence interval).

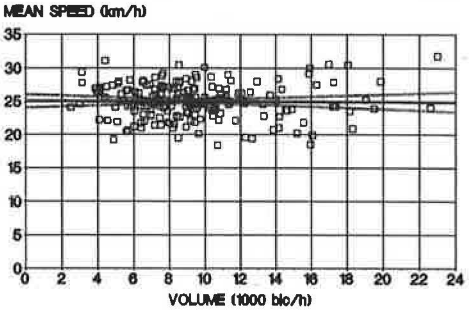


FIGURE 2 Mean speed as a function of volume for the Tour location (linear relation and 95 percent confidence interval).

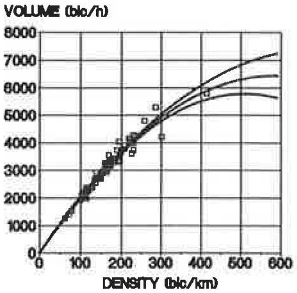


FIGURE 3 Volume as a function of density for the three wide paths (quadratic relation and 95 percent confidence interval).

In analyzing the distribution of lateral positions, a distinction was made between bicycles and mopeds and between free and nonfree cyclists. Free was defined in this context as having a headway in front and to the rear of at least 2.5 sec, a value obtained from visual observation.

Analysis of the distributions exhibited a difference in lateral position between bicycles and mopeds. Mopeds are clearly riding more to the left than bicycles (right-hand traffic), as presented in Table 5.

Because mopeds had to pass the slower bicycles often on these busy paths, it is assumed that they chose their lateral position so to make frequent passing easier rather than that this position was influenced by the height of the curb.

For percentiles lower than 50 percent, there was little difference between the free and nonfree cyclists. This means that for cyclists the curb is the determining factor in choosing the observed lateral position.

Table 6 presents for bicycles the lower percentiles of the distribution of the lateral position. The most important finding has been that the shying-away effect for a curb of 10 cm height was not more than for the lower curbs. It still must be decided which percentile should be chosen in incorporating the effect of the curb in design practice.

Passing and Paired Riding

Passing by faster bicycles or mopeds and riding in pairs can both cause serious disturbance and thus have a great influence on the level of service on bicycle paths. Riding in pairs over longer distances is rather common behavior of dutch cyclists, which is legal in most circumstances.

Thresholds for Passing and Paired Riding

The data at the measuring sites were collected at one cross section (one-dimensional). A special procedure was followed

TABLE 5 50-PERCENTILE LATERAL POSITIONS (cm) OF BICYCLES AND MOPEDS

Location	Width	Bicycle	Moped	Difference
T-Narrow	180	70	97	27
T1-Wide	240	83	166	83
T2-Wide	250	74	115	41
T3-Wide	270	84	134	50

TABLE 6 EFFECT OF CURB HEIGHT ON LATERAL POSITIONS OF BICYCLES (cm)

			Shying effect			
		Curb	Guide-	Percentiles		
Location	Width	Height	lines	p-1	p-5	p-10
T-Narrow	180	5	25	31	41	47
T1-Wide	240	10	50	32	42	49
T2-Wide	250	3	25	20	34	41
T3-Wide	270	0	25	21	38	47
Tour	300	0	25	28	41	51

in order to describe the two-dimensional passing maneuvers using the cross section data as a base. The procedure is described in the following paragraphs.

First, the thresholds for passing and for paired riding have been determined. From video pictures, 50 passing and 50 paired riding situations were selected and the thresholds calculated as a low percentile. The results were as follows:

- Passing takes place if the difference in speed between vehicles is more than 50 cm/sec (1.8 km/hr). For practical reasons, a second threshold was introduced, viz., the passing maneuver had to occur within a distance of 50 m from the measuring site.
- Paired riding occurs when the speed difference is less than 50 cm/sec and the headway is less than 0.125 sec.

With these thresholds, all bicycles that would pass within a certain distance from the measuring site could be determined by calculating the intersection of trajectories. This was indicated as the passing point, under the assumption that the speeds of the passing and the passed vehicles would be constant during the entire passing maneuver and equal to the speed observed at the measuring site. Passing maneuvers were only selected for further analysis if the value of the distance between the passing point and the measuring site was less than 50 m.

Shape of the Passing Maneuver

To determine the shape of the passing maneuver, the data of the three locations in town with the wide bicycle paths were combined. Of all the passing maneuvers that met the thresholds, the known data were as follows:

- Individual speeds and speed differences between passing and passed bicycles at the measuring site,
- Individual lateral positions of passing and passed bicycles at the measuring site and consequently also the differences in lateral position, and
- Distances from measuring site to passing point.

In the second step, the section from −50 to +50 m was divided into 33 subsections of 3 m each. The difference in lateral position between passing and passed bicycles was determined in these 33 subsections. These differences varied between −120 and +225 cm.

The distributions of the lateral positions at each subsection were compared on equality with the first subsection at −48 m from the measuring site with the help of the Kolmogorov-Smirnov test. The first subsections exhibited no significant differences. The passing maneuver appears to start at 27 m before the passing point, where the first significant difference was found, compared to the undisturbed first subsection, and to end 30 m beyond the passing point. In Figure 4, the 85-, 50-, and 15-percentile values are plotted, describing the shape of the passing maneuver. In Table 7, characteristic values of the passing maneuver for different widths of bicycle paths are presented.

The procedure applied to the data of the narrow path (1.80 m wide) was identical to that for the wide path. However,

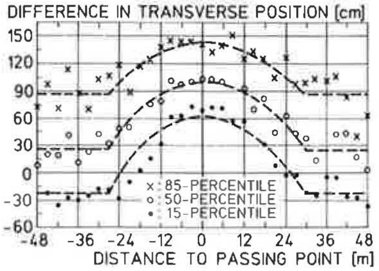


FIGURE 4 Shape of passing maneuver at wide bicycle paths.

TABLE 7 CHARACTERISTICS OF PASSING MANEUVERS

			Average	Pass.man.related		Average
Path	No. of	No. of	Number/	to Passing Point		Passing
Width	Passings	Subsect.	Subsect.	Start	End	Time[1)]
2.40 m	1,178	33(x3m)	36	-27 m	+30 m	11 s
1.80 m	192	15(x6m)	13	-12 m	+12 m	4.5 s

[1)] based upon average speed of bicycles.

the size of the sample and therefore the number of passing maneuvers was considerably smaller. The length of the subsections was therefore chosen at 6 m, which yielded a larger number of observations per subsection, necessary for a reliable distribution, but resulted in a less accurate determination of the shape of the maneuver.

Finally, the shape of the passing maneuver indicated in Figure 4 is determined by the passing bicycle. The passed bicycle keeps a straight path during the passing maneuver.

Paired Riding

On the wide bicycle path, 450 pairs of bicycles were analyzed. The important findings were that there was no influence of speed on the relative lateral positions of the two bicycles in a pair, nor on the lateral position of the pair as such. The difference in lateral position of paired bicycles is clearly smaller than when passing, an average 65 cm for paired bicycles instead of 90 cm for passing bicycles (see Figure 5).

Frequency of Passing Maneuvers

Theoretically, the number of passing bicycles can be calculated, assuming that bicycles don't impede each other anywhere and that the space distribution of speeds is normal, with the equation [see, e.g., OECD (*4*)]:

$$n = XT(V^2/U^2)S/(\pi)^{1/2} \tag{3}$$

where

n = number of passings,
X = length of road section,
T = length of time period,
V = volume of bicycles (bic/hr),
U = space mean speed (km/hr), and
S = standard deviation of speed (km/hr).

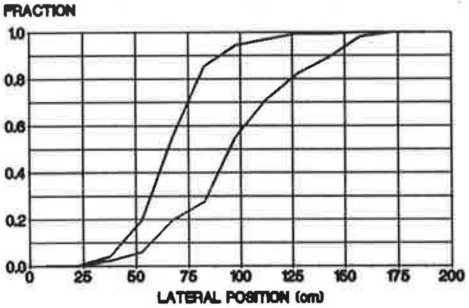


FIGURE 5 Distribution of difference in lateral position of paired (left curve) and passing (right curve) riders.

The number of passings and the volume per 10-min period were determined. With these data, several regression models were applied. The model with the best fit was

$$n = 3.0 \times 10^{-4} V^2 \tag{4}$$

$R^2 = 0.94 \qquad N = 73$

The standard error of the coefficient was 2.6 percent.

To determine the frequency of passing maneuvers, all locations were first treated separately. The outcome of the models scarcely differed, so when applying the model the data of all four town locations were combined (see Figure 6). The number of passings derived from the model is smaller than the theoretical number based on perfect disorder. The coefficient in the theoretical model with correction for the threshold value becomes 4.2 × 10^{-4} instead of 3.0 × 10^{-4}. Obviously there is dependency and order in the bicycle stream.

Frequency of Paired Riding

The relation between frequency of paired riding and volume was analyzed with the same method applied for passing frequency. It was found that the number of paired bicycles is a function of volume. However, the measure of dependence differed with location. Paired riding is clearly dependent on the type of cyclist. For instance, children going to school ride more often in pairs than adults going to work.

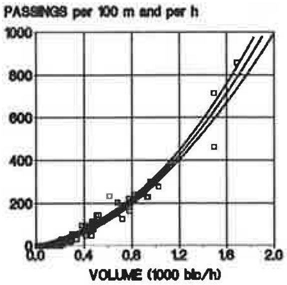


FIGURE 6 Number of passings as a function of volume on wide bicycle paths (quadratic relation and 95 percent confidence interval).

Estimation of Capacity Based on Headways

Earlier, the capacity was roughly estimated by extrapolating a relation between volume and density. In this section, the capacity is estimated using a decomposition of the flow into two types of cyclists: free and nonfree. The other assumptions of this method are as follows:

- With increasing volume, more and more cyclists are obliged to follow a bicycle in front. At capacity everybody is nonfree.
- The distribution of the headways of nonfree cyclists at volumes under capacity is the same as at capacity.
- The method of dividing the cyclists into the two categories of free and nonfree is functioning correctly.

This method has been applied to car traffic at a two-lane road (*5*). Details of the decomposition method are given by Wasielewski (*6*); the method for estimating the parameters of the model was improved by Groeneboom (*7*).

For bicycle traffic, an adjusted definition of headway is required because a stream of bicycles cannot be allocated to well-defined lanes. Two paired bicycles riding next to each other would exhibit a small headway, yet the cyclists would not feel restrained in their movements.

The following solution was chosen. The cross section at the measuring site was divided into sublanes of 15.6-cm width. The sublane touched by a bicycle was known. The bicycle closest ahead that had touched any of five sublanes around the sublane of the bicycle in question was defined as the "bicycle in front" and the headway with respect to this bicycle was used. The definition is shown in Figure 7, in which Bicycle 7 is in front of Bicycles 6 and 9, etc. A property of this definition is that every bicycle has a headway but some have no follower; e.g., Bicycles 1 and 11 in Figure 7.

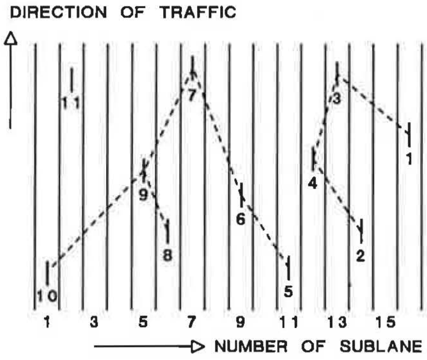


FIGURE 7 Definition of headway for capacity estimation.

The definition implies a lane of 5 * 15.6 = 78 cm. Consequently, the estimated capacity refers to that width in the first instance. The capacity for the total path width can be calculated by multiplying with the number of sublanes and then dividing by 5. The calculated capacities per path width are presented in Table 8.

The capacity per 78-cm lane at the town locations varied from 3,000 to 3,500 bic/hr. The calculated capacity per 78 cm at the Tour location was higher, as could be expected. This type of cyclist usually follows one other closely. The standard error of the estimated capacities is small. However, it should be noted that the real error might be larger, because actually the used model for the headways, on which the decomposition is based, does not fit the data too well.

It is of interest to compare the estimated capacities with the values found in literature. For this comparison, the number of lanes (of 78 cm) is fixed at three, corresponding to a path width of 250 cm and an effective width of (14 * 15.6), or 218 cm. This means that the elevated curbs have an estimated shying-away effect of 32 cm.

Investigators	*Capacity*	*Equivalent Capacity (bic/hr)*
Homburger & Kell (*8*)	13.3 bic/ft-min	5,700
Beukers (*9*)	1,800 bic/lane-hr	5,400
Botma & Papendrecht	From flow-density curve	6,400
Botma & Papendrecht	From headways	8,900

Consequently, a value of around 5,500 bic/hr is mentioned in the literature, whereas this study arrives at 6,500 to 9,000 bic/hr.

CONCLUSIONS AND FURTHER RESEARCH

- The percentage of mopeds in the total volume on busy bicycle paths located in built-up areas is about 5 percent.
- The speed distribution of bicycles is not dependent on location or width of the path and in the Netherlands has not changed in time.
- Bicycles have an average speed of 19 km/hr and a standard deviation of 3 km/hr.
- Bicycle speeds are normally distributed.

TABLE 8 RESULTS OF ESTIMATION OF CAPACITY

Location	Path Width cm	Cap. of 78 cm bic/h	Stand. Error %	No. of Sublanes used	Capacity of Path bic/h
T-Narrow	180	3,300	2.9	9	5,900
T1-Wide	240	2,990	2.2	14	8,400
T2-Wide	250	3,490	1.8	14	9,800
T3-Wide	270	3,090	2.4	14	8,600
Tour	300	5,300	1.0	16	17,000

- Speeds of mopeds are slightly influenced by path width.
- The overall mean speed of mopeds is 38 km/hr; the standard deviation is 5 km/hr.
- No systematic effect of volumes on mean speeds of bicycles or mopeds has been found.
- Only on the narrow bicycle path, higher volumes of bicycles reduced the average speed of mopeds.
- On the basis of the analysis of groups with small headways, there is a small decrease of average speed with volume. The estimated capacity determined by extrapolation is 6,500 bic/hr.
- Even at the Tour ride with values of volumes in groups of 23,000 bic/hr, still no influence was found on the average speed.
- The lateral position of mopeds is clearly more to the left than that of bicycles. It is determined more by the passing maneuver than by the height of the curb.
- The height of the curb does not influence the lateral position of bicycles. No difference in effect between curb heights of 5 and 10 cm was found, as indicated in the Dutch design guidelines.
- The length of a passing maneuver of bicycles on a wide path is 57 m; the passing time is 11 sec. At the passing point, the average lateral distance between vehicles is 100 cm.
- On a narrow path, the length of a passing maneuver is 24 m and the passing time is 4 sec. At the passing point, the average lateral distance between vehicles is 75 cm.
- On a wide path, the average lateral distance between paired bicycles is 65 cm, which is 25 cm less than between passing bicycles.
- The frequency of passing maneuvers is proportional to the volume squared. However, the actual frequency is less than the number derived from a theoretical model based on perfect disorder, which suggests some dependency between cyclists.
- The estimated capacity of a bicycle path of 2.50 m, based on decomposition in free and nonfree cyclists, leads to a value of around 9,000 bic/hr. That is higher than the capacity values mentioned in the literature of around 5,500 bic/hr.

Further research is planned to develop criteria to indicate the level of service on a bicycle path. From this study, it has become clear that mean speed is unsuited as a quality-of-flow indicator, because mean speed is constant over a large volume range.

The planned research should develop a quality-of-flow indicator related to the degree of freedom to maneuver and the possibility to make unrestricted passing maneuvers. This will be done most likely by the development of a simulation model based on the findings in this study.

ACKNOWLEDGMENTS

The study described in this paper was sponsored by the Netherlands Ministry of Transport and Public Works, more specifically, the Transportation and Traffic Engineering Division of Rijkswaterstaat. We wish to thank the sponsor for his permission to publish these results.

REFERENCES

1. P. H. L. Bovy and M. J. P. F. Gommers. Mobility and Use of Means of Transportation in Middle Sized Cities: Some Recent Figures. (In Dutch.) In *Proc., Colloquium Vervoersplanologisch Speurwerk*, E. J. Verroen, ed., Delft, C.V.S., Netherlands, 1987.
2. *Statistisch Zakboek 1984 and 1990.* (In Dutch.) Central Bureau for Statistics CBS, Staatsuitgeverij, The Hague, Netherlands.
3. J. de Vries. *The Relation Between Speed and Volume for Bicycle Traffic.* (In Dutch.) Transportation Research Laboratory, TU Delft, Netherlands, 1979.
4. *Speed Limits Outside Built-Up Areas.* Organization of Economic Cooperation and Development, Paris, 1972.
5. H. Botma, J. H. Papendrecht, and D. Westland. *Validation of Capacity Estimators Based on the Decomposition of the Distribution of Headways.* Transportation Research Laboratory, TU Delft, Netherlands, 1980.
6. P. Wasielewski. Car-Following Headways on Freeways Interpreted by the Semi-Poisson Headway Distribution. *Transportation Science*, Vol. 13, No. 1, 1979, pp. 36–55.
7. P. Groeneboom. *Research on Time Headways on 2×2-Lane Freeways.* (In Dutch.) Centrum voor Wiskunde en Informatica, Amsterdam, Netherlands, 1984.
8. W. S. Homburger and J. K. Kell. *Fundamentals of Traffic Engineering.* Institute of Transportation Studies, Berkeley, Calif., 1988.
9. B. Beukers. Bicycles and Mopeds as Alternative Modes of Transportation. *Proc., 13th International Study Week in Traffic Engineering and Safety*, Montreux, Sept. 1978.

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

Examination of the Speed-Flow Relationship at the Caldecott Tunnel

Hoong C. Chin and Adolf D. May

Recently, a new procedure for analyzing multilane highways was proposed, together with a set of speed-flow curves that were rather different from those in the *Highway Capacity Manual* (HCM). Consideration is given to whether a freeway basic section treated as a multilane divided highway section with no access points can be analyzed using this procedure. The speed and flow characteristics of a freeway section (California State Highway 24) at the Caldecott Tunnel are examined and compared with the speed-flow curves in the HCM for freeways and those in the new procedure for multilane highways. The problems associated with traffic data gathering, reduction, and analysis are also reviewed. It was found that the new procedure reflects the speed-flow characteristics at the Caldecott Tunnel better than the HCM procedure for freeways.

The relationship among speed, density, and flow of traffic on a highway has been the subject of many years of intensive research, resulting in numerous publications on theoretical models and empirical investigations of the traffic characteristics of highway sections. The first traffic flow model postulated by Greenshields assumed a linear relationship between speed and density, giving a parabolic speed-flow function. This shape has influenced the understanding of the relationship between speed and flow on highways for many years.

Traditional speed-flow curves as represented by those in the different versions of the *Highway Capacity Manual* (HCM) (*1*,*2*) are taken to be smooth, that is, differentiable and continuous functions over the entire range of density values. Although the 1985 HCM has noted some studies (*3*,*4*) indicating that speed-flow relationships may be better described by nondifferentiable and perhaps discontinuous speed-flow functions, the results of these studies have not been incorporated in the HCM calibrated curves. Adopting a continuously differentiable speed-flow function necessitates the bending of the speed-flow curve and a precipitous speed drop near capacity. This condition is clearly seen in the HCM speed-flow curves, in which the optimum speed is always at about 30 mph for freeways and multilane highways, regardless of the geometrics.

Recently, a set of speed-flow curves for multilane highways has been proposed (*5*) as a revision to Chapter 7 of the HCM. This new method is referred to here as the JHK method. Figure 1 shows the set of speed-flow curves proposed. Several interesting differences become apparent when these curves are compared with those currently used in the HCM for multilane highways and freeways, as shown in Figure 2. These differences include the following:

- The mean passenger car speed, instead of the average travel speed, is used as the speed variable in the speed-flow relationship.
- Only the upper branch of the speed-flow curves dealing with uncongested flows is presented in the JHK method. Comparing this method with the HCM curves indicates that the former predicts a drastically smaller change in speed with increased flows. The mean passenger car speed remains constant for flows up to 1,400 passenger cars per hour per lane (pcphpl) and subsequently drops gradually and nonlinearly as capacity is reached. The total drop in speed in the upper branch is always 5 mph for all curves. In contrast, the HCM curves predict a gradual drop in speed as flow increases, followed by a precipitous drop in speed near capacity. Clearly, the shape of the JHK curves implies that the speed-flow function is nondifferentiable at capacity.
- In the set of JHK curves, capacity is at 2,200 pcphpl under ideal conditions and is independent of the design speed of the highway. Indeed, capacity is constant regardless of the effects of geometric inadequacies, such as reduced lane width and lateral clearance. The HCM curves suggest that the capacity of multilane highways and freeways under ideal conditions is 2,000 pcphpl except when the design speed is 50 mph, in which case the capacity is 1,900 pcphpl.

Besides these differences in the speed-flow curves, there are several other differences related to the two methods of the operational analysis of multilane highways:

- Development environment, which has been classified as rural and suburban in the HCM procedure for multilane highways, is replaced by a more precise measure of access density in terms of the number of access points per mile in the JHK method.
- As in the HCM procedure, multilane highways are classified as divided or undivided in the JHK procedure. However, highways with continuous left-turn lanes separating opposing flows are considered specifically as divided highways in the JHK method instead of being somewhat in between divided and undivided highways in the HCM procedure.
- In the JHK method, lane width, lateral clearance, type of median, and number of access points affect not the capacity but the operating speed of the highway. In the HCM method, reduction in lane width, insufficient lateral clearance, or less-than-ideal median type and development environment result in a reduction in highway capacity. Hence, adjustment because of nonideal geometrics (not including terrain type) causes

H. C. Chin, Department of Civil Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 0511. A. D. May, Institute of Transportation Studies, University of California at Berkeley, 109 McLaughlin Hall, Berkeley, Calif. 94720.

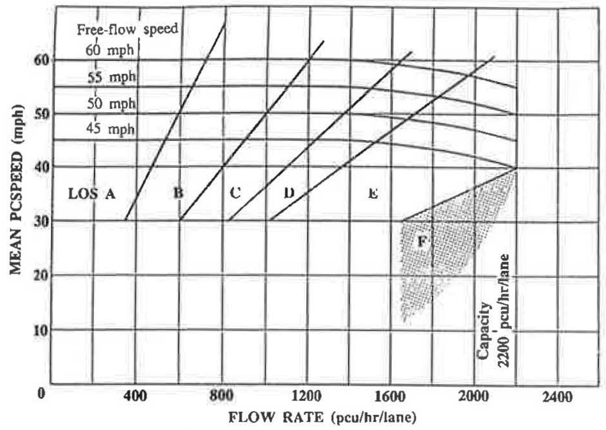


FIGURE 1 JHK speed-flow curves for multilane highways (5).

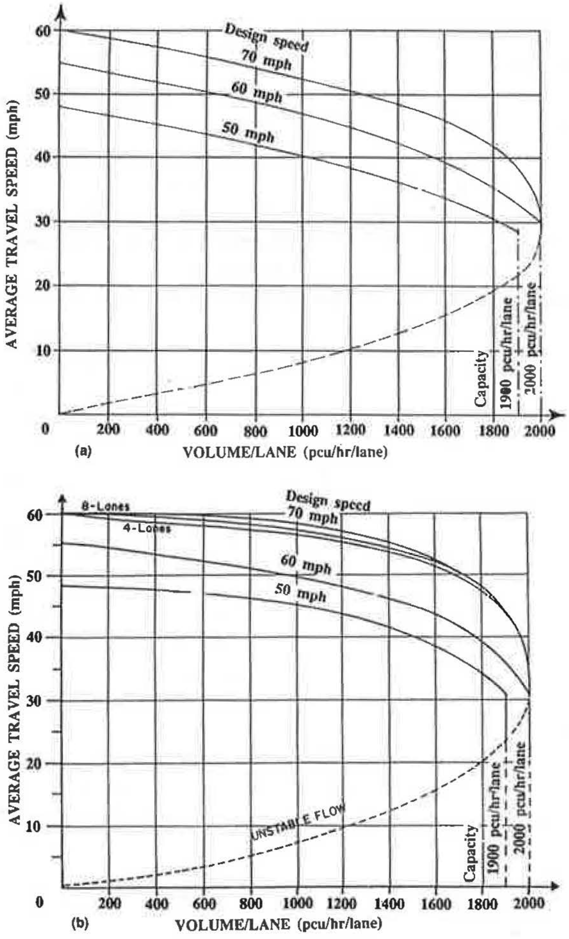


FIGURE 2 HCM speed-flow curves: (a) multilane highways; (b) freeways (2).

a vertical translation of the ideal speed-flow curve in the JHK method but a change in the horizontal scale of the speed-flow curve in the HCM method.

- The JHK method uses a new set of values for heavy-vehicle factors. In general, trucks and recreational vehicles have lower equivalent factors than those in the HCM. Furthermore, no distinction is made between trucks and buses for the purpose of passenger car unit (pcu) conversion.
- Adjustment for noncommuter traffic in the JHK method is considered unnecessary and therefore is not included.
- The density values that divide the various levels of service are also revised, with the exception of levels of service (LOS) A and B. Furthermore, because the capacity may occur at any speed in the new method, the maximum density value for LOS E is not fixed as it is in the HCM.

If the JHK method, which is meant for multilane highways, is accepted in the HCM, some disparity may exist between the design of multilane highway sections following the JHK procedure and the design of freeway sections following the procedure of Chapter 3 in the HCM. For example, a multilane divided highway with less than one access point per mile and designed for 70 mph under ideal conditions is expected to yield a higher capacity than a freeway section under ideal conditions with the same design speed. Consequently, if a freeway basic section were analyzed as a multilane divided highway section with no access points according to the new procedure, a better level of service and higher capacity could be obtained.

The speed-flow characteristics of a freeway section (California State Highway 24) near the Caldecott Tunnel are examined to see how well the observed conditions are predicted by the speed-flow curves in the two procedures: (a) by analyzing the freeway section according to the current HCM method for freeways and (b) by treating the freeway section as a multilane divided highway with no access points and analyzing it according to the JHK procedure.

First, recent studies of speed-flow relationships are reviewed, highlighting the problems encountered in making such studies. This review is presented in the next section, followed by a description of the Caldecott Tunnel site and the procedures of data collection and reduction. The results of the speed-flow analysis are then discussed, together with a comparison of the HCM and JHK curves.

RECENT STUDIES ON SPEED-FLOW RELATIONSHIPS

In recent years, several studies on the traffic characteristics of freeways (*6–10*) have challenged some of the traditional views of the relationship among speed, flow, and density. Except for Banks' work in San Diego (*9*), these studies relate Canadian conditions, with data taken mostly from Queen Elizabeth's Way near Toronto.

Several important issues have been raised in these studies, but basically these issues have to do with the way in which speed, flow, and density data are gathered, reduced, and analyzed. The common contention has been that serious misinterpretations of speed-flow relationships can result if unsuitable data reduction and analytical procedures are employed and if the influence of the study location on the nature of the data is disregarded. The problems associated with unsuitable analytical procedures, data gathering, and reduction are reviewed in this section.

The traditional method of analyzing speed, flow, and density data is to fit a function to the data. The problems related to fitting a curve to a set of speed-flow data have been pointed out by Hurdle and Datta (*4*) and Allen et al. (*6*). The form of the speed-flow curve, whether it is continuously differentiable, continuous but not continuously differentiable, or discontinuous, cannot be normally discerned from data alone. Yet the choice of the speed-flow function can have a significant impact on the interpretation of the speed-flow relationship. In order to avoid this problem, these researchers have departed from the usual way of fitting a curve to the data and have instead resorted to interpreting the speed-flow relationship by merely observing the scatter of data.

One consequence of not fitting a single function to the speed-flow data is that the speed-flow curve is examined in two portions: (a) the upper branch for the uncongested regime, in which the speed is seen to be relatively insensitive to flow, and (b) the lower branch for the congested regime. The main problem then lies in distinguishing the field data belonging to each portion—a complicated task when the flow is near capacity. Hall and Gunter (*7*) have even considered whether the speed-flow states can be between the two branches of the speed-flow curve by examining the issue of transition between the two regimes: flow breakdown and recovery.

Apart from the problems resulting from predetermining the speed-flow function, there are several other problems related to data gathering. All the researchers have reasoned that the location at which the data are gathered affects the nature of the data. Hall and Hall (*10*) have pointed out that the study location with respect to the bottleneck of the freeway is important—a point that was made long ago (*3*) but has not been seriously considered. A location upstream of a bottleneck may never experience flows near capacity because the flows are limited by the capacity of the bottleneck. A truncation in the speed-flow curve may result in this case. On the other hand, flows on a location downstream of a bottleneck may also be constrained by the bottleneck so that the true capacity of the location may not be recorded. Banks (*9*) reiterated this theory and noted that it may be difficult to find a location in which flows at capacity are observed. As pointed out by Hall and Hall (*10*), even though a location is characterized by the two portions of the speed-flow curve, in practice to obtain both of them for a particular location is difficult.

Persaud and Hurdle (*8*) have also indicated that speed data gathered in the acceleration zone downstream of a queue may be seriously misinterpreted. Because vehicles discharging from the queue accelerate over a finite distance, speeds observed in this zone may be influenced by the presence or absence of a queue. On this point, Hall and Hall (*10*) observed a reduction in speed in their site when there was an upstream queue, but Banks (*9*) observed no discernible change in speed in his site.

Serious problems may also arise because of data reduction procedures. For example, Allen et al. (*6*) have analyzed the data lane by lane, supposing that there are different speed-flow relationships between the different lanes, whereas Banks (*9*) has chosen to combine the data for all lanes, arguing that some portions of the speed-flow curve may not be observed otherwise. Persaud and Hurdle (*8*) also suggested that the averaging effect of taking data from different locations in a bottleneck may be the reason for the precipitous drop in speed near capacity observed in previous studies.

Related to the problems of data gathering and data reduction is the quality of the data gathered. Banks (*9*) has made use of data obtained from loop detectors installed on the freeways. These data are in the form of flow counts and occupancy. Speed is then determined indirectly, a computation that is found to be rather unreliable (*11*). On the other hand, Hurdle and his associates (*4,8*) have made use of time lapse photography to obtain a more direct, though not necessarily more accurate, measurement of speeds. The disadvantage they faced is that the sample size can be rather small. In order to increase the sample size for the study, the sampling time interval is shortened, to 2 min in their case. However, this change results in a greater variance in both speed and flow on the speed-flow plot, which may hamper a good evaluation of the change of speed with flow.

Another problem encountered by these researchers is the question of truck percentages. Persaud and Hurdle (*8*) selected a site at which trucks were not present, but, in the other studies, the proportion of trucks was estimated. All studies have assumed a truck equivalent factor of 2, but Banks (*9*) has avoided the problem by expressing flow in vehicle units rather than pcu.

In the light of these problems, it is extremely important that data be obtained and analyzed properly if a meaningful comparison is to be made between observed speed-flow conditions and predicted relationships.

DATA COLLECTION AND REDUCTION

The data used in this study are taken from the two westbound lanes that emerge from the northern bore of the Caldecott Tunnel through which California State Highway 24 passes. The Caldecott Tunnel has three bores: the southern bore is devoted to eastbound traffic, the northern bore to westbound traffic, and the center bore to reversible traffic flow. Typically, the center bore is open to westbound traffic in the morning and eastbound traffic in the afternoon.

The section under investigation is about 200 ft downstream of the tunnel exit. At this point, the two lanes are each 12 ft wide, and the lateral clearance is 6 ft on each side. The lanes are sloping, with a downgrade of 5.5 percent. An offramp from the freeway is located about 900 ft after the tunnel exit, and an onramp is located 600 ft further downstream. The flow on the off- and onramps, which primarily serve the control center above the tunnel, is so low that traffic on the main line is unaffected by the traffic on the ramps.

When the westbound flow exceeds the capacity of the tunnel, traffic backs up from the entrance of the 3,000-ft-long tunnel. Under normal conditions, traffic emerging from the tunnel is free flowing.

At the time of the study, two sets of detectors were installed about 200 ft downstream of the tunnel exit along the westbound lanes. These detectors were installed as part of a project funded by the California Department of Transportation (Caltrans) to study detector technologies for freeway surveillance and control. The first set of detectors are inductive loop detectors buried on the roadway on each of the two lanes. (These two lanes are referred to here as the shoulder and middle lane. There are other westbound lanes, but they use the other tunnel bore.) The second set of detectors are ultrasonic detectors suspended from a roadway overpass about 100 ft downstream of the loop detectors. Both sets of detectors are wired to a roadside junction box and connected to Fisher-Porters recorders and Caltrans 170 controllers located in the Caldecott Tunnel control room just above the study site. Every 1/60 sec, the 170 controllers would scan the various detector stations for the presence of a vehicle within the detection zones. Detector information is transmitted from the control center via a telephone line to a microcomputer for data logging at the Institute of Transportation Studies at the University of California in Berkeley.

The detector signals are transformed to pulses, from which durations of "on" and "off" times at each detector station can be determined. Irregular signals have been eliminated, and the possibility of detectors miscounting the vehicles has been studied by comparing the detector pulses with visual observation of the traffic recorded on a video recorder over a period of 2 hr. An average overcount of 0.4 percent was made by the loop detectors, whereas the ultrasonic detector overcounted on average by 2.0 percent. The computation of flow across the detectors is therefore based on the loop information.

Speed measurements are obtained not from occupancy data but by measuring the time of travel of a vehicle between the loop and ultrasonic detectors. Provided that the correct pair of pulses can be identified, it is possible to determine the speed of the vehicle by noting the time difference between the onset of the pulses. Matching the pulses is crucial to the correct evaluation of the speeds because any mismatch can lead to biased results. The criterion to identify the correct match is based on the computed speed of the vehicle associated with the pulses. If the pulses are wrongly matched, the computed speed will either be too high or too low. An arbitrary, though not unreasonable, upper limit of 100 mph and a lower limit of 20 mph are set to test the pulse matching. Matching proved to be rather insensitive to the values of the cut-off levels when they are near the two limits set. In most cases, the pulses matched sequentially. On the basis of observations made from the 2-hr video recording, the number of matched pulses accounts for about 99.5 percent of the cases. The computed speed values are certainly not affected by ignoring the unmatched cases.

In order to express flows in pcu, the type and proportion of heavy vehicles must be known. However, these characteristics cannot be obtained directly from the detector information. On-site observation shows that the number of recreational vehicles and buses is small compared with the number of trucks. It is therefore sufficient to consider just one class of heavy vehicles. Furthermore, it is possible to estimate the percentage of heavy vehicles by treating long vehicles as heavy ones. This calculation requires that the length of vehicles be known, which is obtained by noting the computed speed of individual vehicles and the corresponding recorded "on" times. Investigation into the distribution of "on" times of the loop and ultrasonic detectors shows that the loop detectors gave more reliable "on" times. The length of vehicles is therefore computed on the basis of the "on" times recorded by the loop detectors. Although the computed values of vehicle length are not precise, the estimated percentage of heavy vehicles is not significantly affected by this. According to a comparison of the detector pulses and the video recording, a threshold of 27 ft is found to distinguish long vehicles, including trucks, buses, and recreational vehicles, from passenger cars. Instead of applying a fixed pcu equivalent for the heavy vehicles (such as 2), the pcu equivalent factors used in this study are based on the recommended values given by the HCM and JHK methods.

A total of 131 hr of data was collected over several occasions from March 16 to April 18 of 1990. On the basis of detector information, the space-mean speeds (harmonic mean of individual speeds), vehicle counts, and proportion of heavy vehicles on each of the two lanes for various time intervals were obtained.

ANALYSIS OF SPEED-FLOW DATA

Speed-Flow Relationship

Figure 3 shows the speed-flow plot for the average lane using data grouped in 2-min, 15-min, and 1-hr sampling periods. For shorter periods, there is a greater scatter in the data in the measurements of speed and flow. However, the abundance of data points and the controlled measurements of speed have produced a tight band of the data in all three graphs (except where the flow is extremely low, resulting in a corresponding small number of speed measurements). Clearly, there is a well-defined relationship between speed and flow in the upper branch of the speed-flow curve.

Speed is insensitive to flow in the low-flow region for flows up to about 800 vehicles per hour (vph), beyond which the speed drops gradually but definitely until the maximum flow is observed. The graphs indicate that it is not correct to suggest that speed is insensitive to flow for the entire upper branch of the speed-flow curve. However, a precipitous drop in speed near capacity is clearly absent, and speed remains above 50 mph even for flows exceeding 2,000 vph. There is also little evidence that the second derivative of the speed-flow function is negative. Indeed, the upper branch of the speed-flow curve can be considered a bilinear function, with speed independent of flow in the low-flow range and a linear drop in speed in the moderate- and high-flow range. The speed drop observed is about 0.7 mph for every increase in flow of 100 vph, a little more than what Banks reported in San Diego (*9*).

The speed-flow plots also suggest that flows in excess of 2,200 vph are possible for 1-hr periods and certainly for 15-min and 2-min periods. This finding confirms previous observations in Canada and San Diego that flows substantially higher than the HCM capacity value of 2,000 pcphpl are possible even over a sustained period.

Effect of Different Lanes

The speed-flow relationships are plotted on the basis of the 15-min grouping for individual lanes and shown in Figure 4. Under low-flow conditions, there seems to be little difference in the speeds measured on the shoulder lane and those measured on the middle lane. As flow increases, the difference in speed becomes more apparent, with traffic on the shoulder lane moving at about 3 to 5 mph lower than that on the middle lane. Also, a higher value of maximum flow is recorded in the middle lane than in the shoulder lane. The difference in speeds and maximum flow attained in the two lanes may be

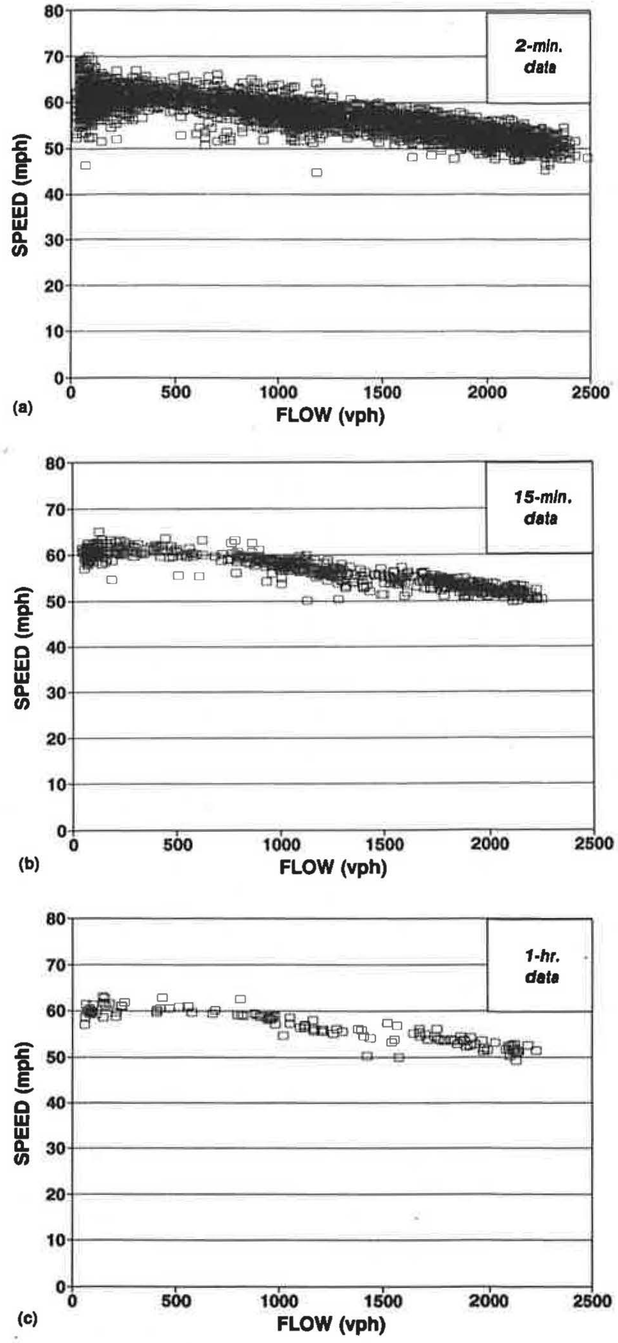


FIGURE 3 Speed-flow plot at Caldecott Tunnel for different sampling intervals.

caused by a larger proportion of heavy vehicles and other slower moving vehicles in the shoulder lane.

Figure 5 shows the distribution of the heavy vehicles for the two lanes. Most of the time, the percentage of heavy vehicles is small, with 45 percent of the cases without heavy vehicles in the middle lane. On the other hand, there is a wide range in the proportion of heavy vehicles in the shoulder lane; in a number of cases, this proportion exceeds 10 percent. On average, 2.7 percent of heavy vehicles travel on this stretch of the roadway throughout the day, with the mean speed of heavy vehicles about 5 mph lower than that of the passenger cars.

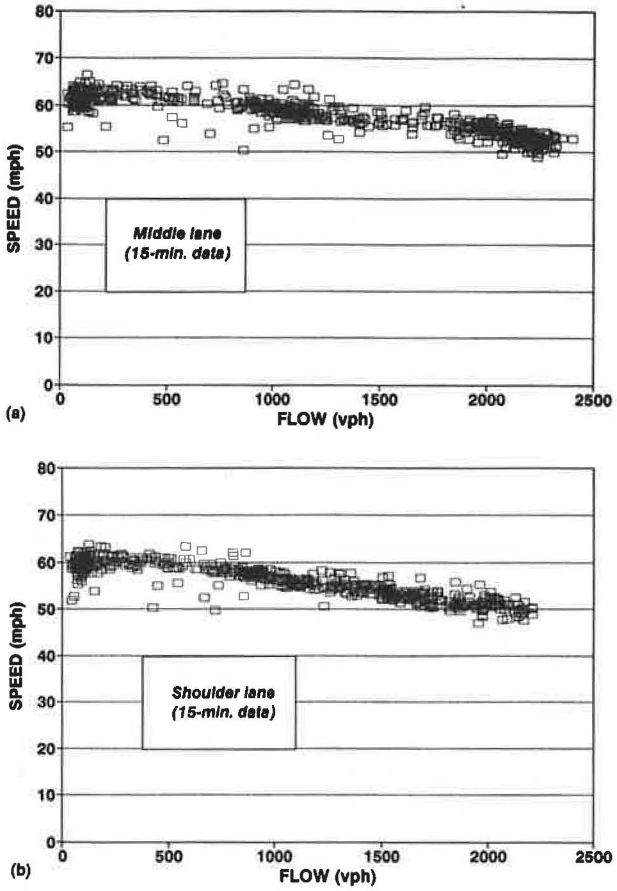


FIGURE 4 Speed-flow plot at Caldecott Tunnel for different lanes.

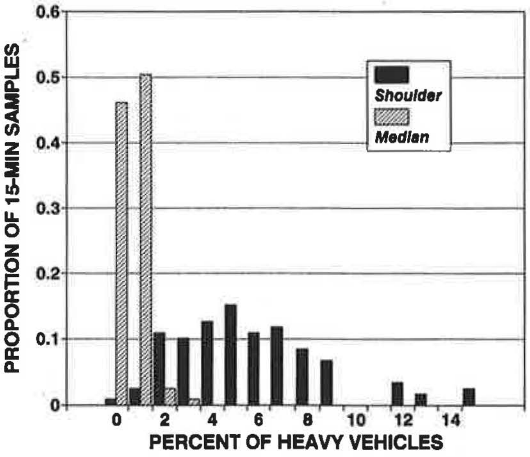


FIGURE 5 Distribution of heavy vehicles.

Effect of Upstream Queue

The effect of an upstream queue on the speed-flow characteristics at the site is also investigated by isolating the speed-flow data corresponding to periods with an upstream queue. The precise times when the queue forms and vanishes are not known. However, in general, an upstream queue is present at the entrance of the tunnel during the morning peak hours.

Figure 6 shows speed-flow data in which cases with an upstream queue are plotted differently than those without an upstream queue. The data are gathered at 15-min intervals for the middle and shoulder lanes. Judging from the plots for both the lanes, there is no evidence that the speed-flow characteristics are any different when there is an upstream queue than when there is not. That the speed is not different between the two groups of data is not surprising because vehicles discharging from the queue would have returned to their normal speed at the end of the tunnel, which is about 3,000 ft long. The maximum flow attained when there is a queue is also comparable with the high flows attained otherwise. However, it is not clear why flows dropped significantly below the maximum possible on a number of occasions.

Comparison with HCM Curves for Freeways

To compare the observed speed-flow data with the HCM curves, the speed, flow, and proportion of heavy vehicles must be grouped in 15-min periods in accordance to the time period of measuring flow used in the HCM. Because it is not possible to distinguish the different types of heavy vehicles from the data available, a single value of truck equivalent factor is used to convert the vehicle flows to passenger car flows. For downgrades, the equivalent factors given in the HCM are rather imprecise. The recommended procedure is to establish the average truck speed and to obtain the equivalent grade of the same length as outlined in Appendix I of Chapter 3 in the HCM (2). On the basis of an average truck speed of 50 mph, an equivalent factor of 2 is used. Because the lane width and lateral clearance of the site are ideal, no adjustment is made for these factors. In addition, no adjustment is made for the influence of driver population, which is assumed to be commuter traffic. Because the amount of adjustment varies for each 15-min period, adjustments are made on the data points rather than on the speed-flow curves. The adjusted 15-min speed-flow data are plotted in Figure 7, together with the HCM speed-flow curves corresponding to four- and eight-lane freeways with a design speed of 70 mph.

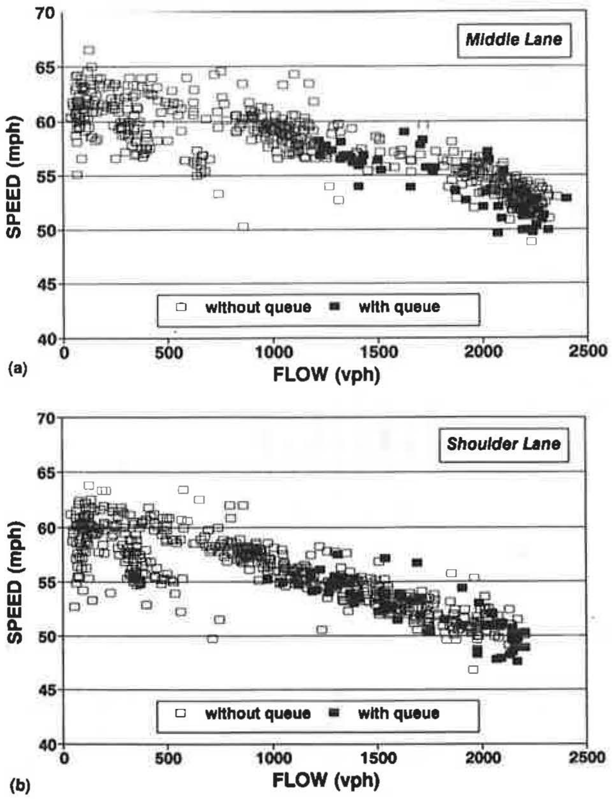


FIGURE 6 Speed-flow plot at Caldecott Tunnel with and without upstream queue.

The observed data deviate from the HCM curves, especially under high-flow conditions. On a number of occasions, a flow in excess of the capacity value of 2,000 pcphpl has been observed. If an operational analysis of the freeway section were performed for various 15-min intervals on the basis of HCM curves, LOS E would be encountered on a large number of instances in which the speed is higher than the 55-mph speed limit.

The discrepancy between the observed data and the HCM curves may result simply because the site under investigation is not an average one represented by the HCM. On the other hand, the HCM recommended values used may not be appropriate. If the latter is true, then there are three possibilities in explaining the difference between the observed and predicted data. First, the pcu factor may be too high. Although possible, this error is difficult to judge. Moreover, even assuming a passenger car–unit value of 1.0, the difference between observed and predicted remains high. Second, the HCM capacity value of 2,000 pcphpl may be too low. This possibility was already noted in the previous section. Third, the speed-flow curve may be inappropriate. If the uncongested speed-flow data found in this site are representative, they indicate that the second derivative in the speed-flow curve is unlikely to be negative and is definitely not so negative that it results in a precipitous drop in speed near capacity.

Comparison with JHK Curves for Multilane Highways

The basis of comparing the speed-flow characteristics obtained from the Caldecott Tunnel site with the JHK method of analyzing multilane highways is that the freeway section can be regarded as a multilane divided highway with no access

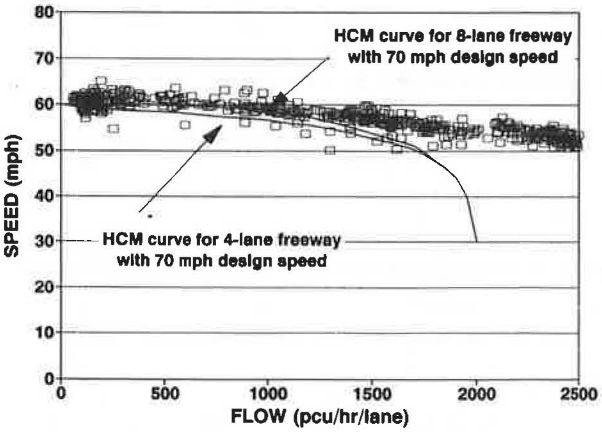


FIGURE 7 Comparison of speed-flow plot with HCM curves for freeways.

points. The number of differences between the current HCM method of evaluating freeways and the JHK method of evaluating multilane highways requires that the speed-flow data be adjusted differently. First, the speed of passenger cars is used in obtaining the speed-flow plot. For the set of data used, however, there is only a small difference between the passenger car speed values and the all-vehicle speed values. The method of equivalent grade is again used to estimate the heavy-vehicle factor, but in this case an equivalent factor of 1.7 is used as recommended.

In selecting the appropriate speed-flow curve, the JHK method requires that the free-flow speed of passenger cars in the section be known. Two ways of determining the free flow speed have been suggested. The first is to estimate the ideal free flow speed from the value of the posted speed limit and then to reduce the speed value by adjustments for nonideal geometric conditions. The second requires a site measurement of passenger-car speeds under low- or moderate-flow conditions.

The posted speed limit on this freeway is 55 mph. According to Table 7-2 in the JHK procedure (5), this limit gives a corresponding ideal free flow speed of 59.3 mph. Because the geometric conditions are ideal at this site, no reduction in the free-flow speed value is necessary. The speed-flow curve that is based on this value, together with the adjusted speed-flow data, is shown in Figure 8. The JHK speed-flow curve marginally underestimates the speeds under low-flow conditions but overestimates the speeds under moderate- and high-flow conditions. In general, this speed-flow curve gives a better reflection of the conditions at the Caldecott Tunnel site than the HCM speed-flow curves used in the previous section. However, a maximum flow of 2,300 pcphpl attained at a speed of about 50 mph is still higher than the JHK capacity value of 2,200 pcphpl at 55 mph. Furthermore, the speed at capacity is approximately 10 mph lower than the free flow speed, compared with a speed reduction of 5 mph as suggested in the JHK method.

If speed measurements are taken from the field to estimate the free flow speed and a 15-min sample is used, then the speed value in any of the data points falling within the low- or moderate-flow region shown in Figure 8 could have been used. In this case, the free flow speed may range from 55 to 63 mph, and any speed-flow curve with intercepts within that range may have been chosen for the analysis. Figure 9 shows the envelope of the speed-flow data, together with the curves corresponding to the upper and lower values of the measured free flow speed. There can be quite a difference in the choice of the curve used. However, the difference may not be significant in the evaluation of the level of service.

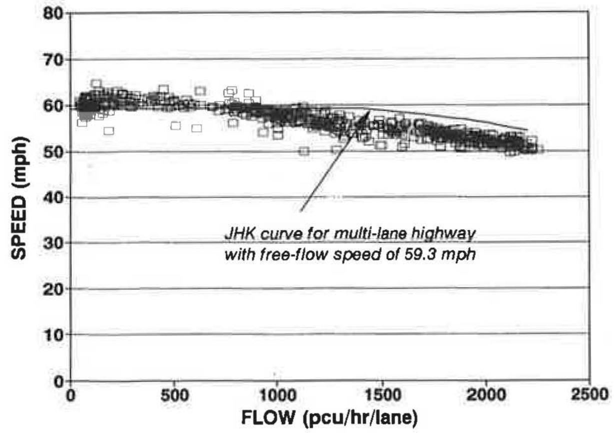


FIGURE 8 Comparison of speed-flow plot and JHK curve on the basis of posted speed limit.

CONCLUSION

It could be argued that the results discussed in the previous paragraphs are limited because only a single location was investigated. Certainly, some of the findings cannot be generalized and can only be confirmed with observations from other sites. Nevertheless, the findings not only gave some insight into the relationship between speed and flow but should also provide motivation for more investigations on the speed-flow interaction of traffic on the highways.

The manner in which the speed and flow data were gathered not only resulted in a sufficiently large number of samples for analysis but also ensured reliable measurement of traffic speeds. The results indicate that, under uncongested conditions, there is a well-defined relationship between speed and flow that can be described in the form of a bilinear function. The study confirms earlier reports that there is no precipitous drop in speed as flow approaches capacity and that flows in excess of 2,200 pcphpl are possible. However, the presence of an upstream queue does not seem to affect the speed-flow relationship at the location studied.

The results also show that the speed-flow conditions at the Caldecott Tunnel have not been well represented by the HCM speed-flow curves but were reasonably reflected by the JHK speed-flow curves for multilane highways. If the results obtained are representative, it may be better to evaluate freeway basic sections by treating them as multilane divided highways with no access points and following the JHK procedure for multilane highways.

ACKNOWLEDGMENTS

This work was performed while the first author was on sabbatical leave at the Institute of Transportation Studies at the

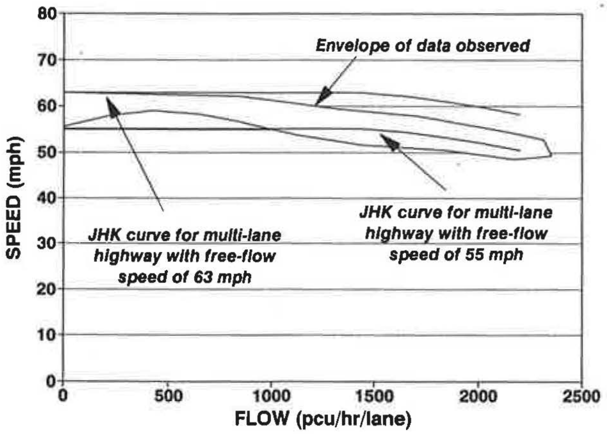


FIGURE 9 Comparison of speed-flow plot and JHK curve on the basis of estimated free flow speed.

University of California in Berkeley. It was made possible through the support provided by the National University of Singapore and the University of California. The authors acknowledge the use of detector data that were gathered under a project funded by the California Department of Transportation and FHWA. The assistance of Larry Labell in obtaining the detector data is also much appreciated.

REFERENCES

1. *Special Report 87: Highway Capacity Manual 1965*. HRB, National Research Council, Washington, D.C., 1965.
2. *Special Report 209: Highway Capacity Manual 1985*. TRB, National Research Council, Washington, D.C., 1985.
3. J. S. Drake, J. J. Schoefer, and A. D. May. Statistical Analysis of Speed-Density Hypothesis. *Proc., 3rd Symposium on the Theory of Traffic Flow*, Elsevier, N.Y., 1967.
4. V. F. Hurdle and P. L. Datta. Speeds and Flows on an Urban Freeway: Some Measurements and a Hypothesis. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 127–137.
5. W. R. Reilly, D. W. Harwood, J. M. Schoen, and M. F. Holling. *Capacity and Level of Service Procedures for Multi-Lane Rural and Suburban Highways*. JHK & Associates and Midwest Research Institute, May 1989.
6. B. L. Allen, F. L. Hall, and M. A. Gunter. Another Look at Identifying Speed-Flow Relationships on Freeways. In *Transportation Research Record 1005*, TRB, National Research Council, Washington, D.C., 1985, pp. 54–64.
7. F. L. Hall and M. A. Gunter. Further Analysis of the Flow-Concentration Relationship. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986, pp. 1–9.
8. B. N. Persaud and V. F. Hurdle. Some New Ideas that Challenge Some Old Ideas About Speed-Flow Relationships. In *Transportation Research Record 1194*, TRB, National Research Council, 1988, pp. 191–198.
9. J. H. Banks. Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989.
10. F. L. Hall and L. M. Hall. Capacity and Speed-Flow Analysis of the QEW in Ontario. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.
11. F. L. Hall and B. N. Persaud. An Evaluation of Speed Estimates Made with Single Detector Data from Freeway Traffic Management Systems. In *Transportation Research Record 1232*, TRB, National Research Council, Washington, D.C., 1989.

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

Two-Capacity Phenomenon at Freeway Bottlenecks: A Basis for Ramp Metering?

JAMES H. BANKS

The issue of whether ramp metering can increase the capacity of freeway bottlenecks by prevention or delay of flow breakdown on the freeway main line is considered. A summary of the results of four case studies of metered bottlenecks in San Diego is presented. The hypothesis that flow decreases when it breaks down is confirmed, provided the hypothesis applies to individual lanes. Flow decreases ranging from 10 percent to less than 1 percent were observed in the left lane at the various study sites. When averaged across all lanes, flow decreased by about 3 percent at one site; there was no significant change at the other sites. It is concluded that this phenomenon is unlikely to provide a basis for metering at more than a few locations, because decreases in flow across all lanes are very small, sometimes nonexistent. Even where there are decreases in total flow, there is risk that metering will be counterproductive if it is too restrictive or begins too early.

A previous paper (1) presented a study of flow processes in the vicinity of a high-volume freeway bottleneck in San Diego. Detailed detector data were analyzed and compared with videotapes of traffic flow. Evidence supported the hypothesis that capacities at this bottleneck decrease when queues form. In addition, it was found that queues formed about 1,500 ft upstream of the merge point, rather than at the merge point. This situation occurred despite merge rates of approximately 2,500 to 2,800 vehicles per hour (vph) and flows downstream of the onramp that commonly exceeded 2,400 vehicles per lane per hour (vplph) and sometimes reached 2,600 vplph.

The overall objective of this study was to determine whether ramp metering can increase the capacity of freeway bottlenecks by preventing flow breakdown. The project eventually included case studies of three additional bottlenecks. The results of the entire series of case studies are summarized in the following paragraphs, and the potential of the so-called "two-capacity" phenomenon (the existence of which was confirmed by all the case studies) is explored as a basis for ramp metering. A companion paper in this Record focuses on the process by which congestion was initiated at these locations and addresses the question of why capacity should decrease when flow breaks down.

BACKGROUND

Past case studies of ramp metering have indicated that even crude metering systems can sometimes achieve significant reductions in total travel time with no significant reduction in total vehicle-miles of travel (2,3). One common explanation is that maximum flow rates at bottlenecks decrease when queues form; hence, metering increases the capacity of the bottlenecks by preventing queueing on the freeway.

This two-capacity hypothesis is found in several standard works (4,5) and was previously used as a basis for a ramp metering strategy by Athol and Bullen (6). The hypothesis has often been related to two-regime or dual-mode traffic flow theories (6–9), but questions have been raised as to whether such theories might represent a misinterpretation of the data (10–12). Recent attempts to verify the two-capacity hypothesis were undertaken by Hurdle and Datta (13) and Persaud (14) but were proved inconclusive because of insufficient data.

Hurdle and Datta (13) and Persaud (14) took the proof of the two-capacity hypothesis to be a flow pattern in which the mean flow rate at the bottleneck drops abruptly just as the queue forms. Banks (1) extended this approach by considering how the distribution of short-term counts just downstream of the bottleneck should be affected by queue formation. Two situations were contrasted: one in which there is no change in capacity and one in which capacity decreases. It was assumed that individual lanes might have separate capacities and concluded that, if even one lane is found in which the mean flow rate is less after the queue forms or in which high counts are less frequent after queue formation, this finding would tend to confirm the two-capacity hypothesis. Increases in flow in noncritical lanes or in total flow across all lanes do not negate the hypothesis, however, because these increases might happen as a result of shifts in lane use. On the basis of these tests, it was concluded that the two-capacity hypothesis was confirmed at the first bottleneck studied.

The test of the two-capacity hypothesis, as a proposition about the facts of traffic flow, depends on decreases in flow in individual lanes. Its value as a basis for ramp metering, however, depends on the existence of a decrease in flow across all lanes when queues form. At the first study site, flow averaged across all lanes did decrease in eight of nine cases. One issue to be addressed by the remainder of the study was the extent to which this situation is typical.

In addition to the question of how often total flow decreases when queues form, there are other concerns related to the exploitation of the two-capacity phenomenon by ramp metering. In his previous paper, Banks (1) introduced a model relating the potential benefits of metering (i.e., time savings)

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

to several details of the metering strategy. The model clearly indicated that the potential benefit of metering depended, among other things, on (a) the relationship between flow and the time to flow breakdown, (b) the difference between the time metering begins and the time flow would have broken down without metering, and (c) the relationships between the metering rate and the maximum unmetered prequeue flow and queue discharge rates. This model is further developed here, and evaluated by means of sensitivity analyses, to provide a better idea of whether and how metering could take advantage of the two-capacity phenomenon.

METHODOLOGY

The study methodology involved comparison of detailed detector data taken in the vicinity of the various bottlenecks with videotapes of traffic flow. Analysis of detector data was based on an extension of the event-based averaging technique described by Allen et al. (*15*). Characteristics of 30-sec count distributions (including means and standard deviations) were determined for 12-min periods before and after flow breakdown. In addition, regressions of 30-sec count versus time were performed for these same time periods to determine whether flows were increasing or decreasing significantly, and frequency polygons of the count distributions (aggregated over all days included in each case study) were plotted to determine their shapes. Further details of the study methodology are documented elsewhere (*1,16*).

STUDY SITE CHARACTERISTICS

The following four bottlenecks were selected for study. In each case, the exact location of the bottleneck was determined as part of the field study.

1. Westbound Interstate 8 at College Avenue, morning peak period;
2. Northbound Interstate 805 at El Cajon Boulevard, morning peak period;
3. Eastbound Interstate 8 at College Avenue, evening peak period; and
4. Southbound State Route 163 at Washington Street, morning peak period.

Figures 1–3 show schematic diagrams of these sites, and selected characteristics of the sites are presented in Table 1. Further details can be found elsewhere (*1,16*).

As can be seen, all sites involve volumes (both on the main line and at the ramp terminals) that are considerably in excess of those that would be predicted by the *Highway Capacity Manual* (HCM) (*4*) for similar circumstances.

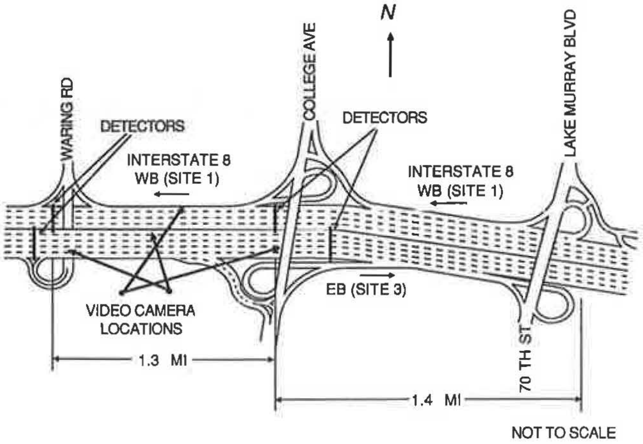


FIGURE 1 Schematic diagram of Sites 1 and 3.

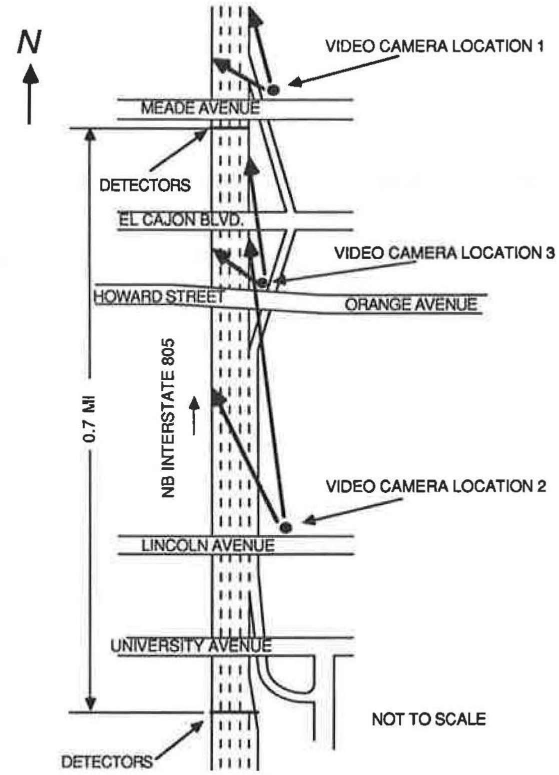


FIGURE 2 Schematic diagram of Site 2.

All sites were metered to some extent, although the effectiveness of the metering varies widely. Metering upstream of Site 1 is extensive, so that virtually all vehicles on the main line approaching the bottleneck have passed a meter. At Site 4, only vehicles entering from the Washington Street ramp itself are metered, and there is no attempt to control mainline flows approaching the bottleneck. Some metering exists upstream of Sites 2 and 3, but it is less extensive than at Site 1; that at Site 2 is probably more effective than that at Site 3.

Other important characteristics distinguishing the sites are grades, horizontal curves, and the presence of ramp terminals. In general, the severity of the grades increases from a downgrade at Site 1 to an upgrade of 6 percent at Site 4. There are also critical horizontal curves at Sites 1 and 4. At Sites 1, 2, and 4, maximum volumes per lane occur just downstream of onramp junctions; at Site 3, the maximum volume per lane occurs just upstream of an offramp junction.

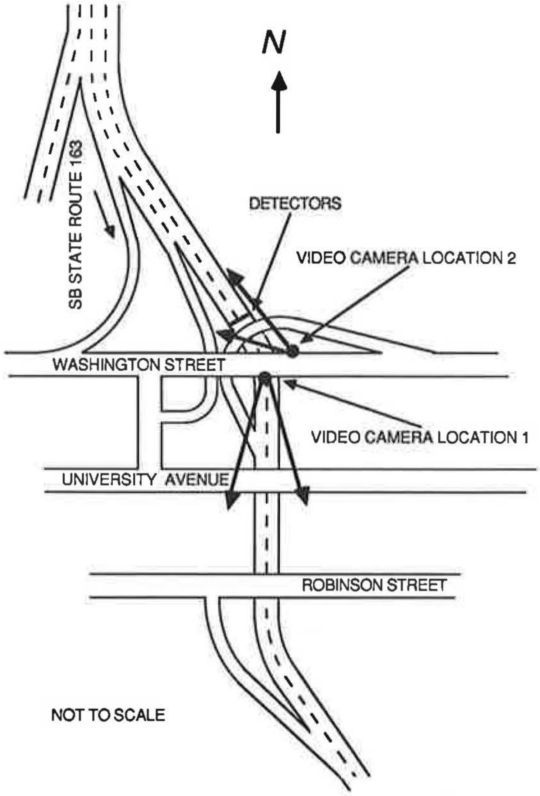


FIGURE 3 Schematic diagram of Site 4.

TABLE 1 STUDY SITE CHARACTERISTICS

Characteristic	Site 1	Site 2	Site 3	Site 4
Number of Lanes, One Direction	4	4	5	2
Mainline Flow, VPHPL				
Critical Point	2250	2175	2100	2325
Maximum	2400	2300	2100	2450
Critical Ramp (VPH, 6-min. Flow)				
Merge	2500-2800	2500-2800	n.a.	2500
Diverge	n.a.	n.a.	2000-2600	n.a.
Heavy Vehicles	4.0%	4.0%	4.5%	1.9%
Critical Upgrade				
Length, Mi.	n.a.	1.8	1.5	1.0
Average Grade	n.a.	1.7%	2.9%	4.9%
Steepest Grade	n.a.	2.0%	3.0%	6.0%
Critical Horizontal Curve Radius, Ft.	2000	n.a.	n.a.	600

RESULTS

Table 2 presents a comparison of the changes in the flow process that were observed to occur at the onset of congestion. Further details are given elsewhere (*1,16*).

Table 2 indicates that, in a majority of cases, the following changes occurred at each site:

- The mean flow in the left lane decreased. Flow decreases (averaged over all days) ranged from 10.5 percent at Site 1 to 0.6 percent at Site 4. On the basis of the sign test (i.e., assuming that the probability of an increase or decrease is 0.5 and determining the probability of occurrence, as given by the binomial distribution, of at least the number of decreases observed), all the decreases except that at Site 4 were significant at the 5 percent level.
- The mean flow across all lanes decreased. In this case, however, the flow change (averaged over all days) was actually positive in three of four cases, ranging from −3.2 percent at Site 1 to +1.2 percent at Site 4. On the basis of the sign test, only the decrease at Site 1 was significant at the 5 percent level.
- Where regression slopes of flow versus time tended to have positive slopes before the onset of queueing (thus indicating that flow was increasing despite the metering), the value of the regression line just before the onset of congestion was greater than that just afterward.
- The relative frequency of the highest counts decreased, whether left-lane counts were considered individually or counts for all lanes were averaged together.
- The percentage of flow in the left lane decreased, whereas that in the right lane increased.
- The variances of the 30-sec counts decreased. Variances and coefficients of variation varied considerably by site and from day to day. Typical coefficients of variation ranged from around 0.1 at Site 1 to 0.2 at Site 3.

In addition, in every case in which the typical point of flow breakdown could be seen, it was somewhere other than at the merge or diverge point (i.e., the right lane just upstream of the offramp or just downstream of the onramp), which would have been identified as critical by Chapter 5 of the HCM. In the three cases in which the merge point would have been expected to be critical, the actual point of flow breakdown was upstream. In the other case (Site 3), flow breakdown occurred both upstream and downstream of the apparently critical diverge point. In all cases, this situation occurred even though the merge or diverge rates in question were far in excess of the merge and diverge capacities reported in the HCM.

At all sites, flow breakdown was observed to be associated with unstable speeds in large and dense platoons of vehicles that formed in high-volume, noncongested flow. At the three sites involving upgrades, the critical platoons were normally associated with the presence of slow-moving heavy vehicles.

The results of all the case studies supported the two-capacity hypothesis, but the degree to which they supported it varied. In general, support for the hypothesis was strongest at Site 1 and weakest at Site 4. For instance, Figure 4 shows a comparison of the relative frequency distributions of 30-sec counts for the left lane of each site, before and after queue formation, and Figure 5 shows similar information for flow averaged across all lanes. The tendency for high counts to occur more frequently in the critical lane before queue formation is obvious at Site 1 and questionable at Site 4.

TABLE 2 COMPARISON OF RESULTS OF CASE STUDIES

Response to Flow Breakdown	Number of Occurrences/Total Days			
	Site 1	Site 2	Site 3	Site 4
Mean Flow Decreases (Decreases/Total Days)				
Left Lane	9/9	12/14	11/14	10/17
All Lanes	8/9	7/14	8/14	9/17
Average Flow Change (Percent, All Days)				
Left Lane	-10.5%	-4.0%	-4.8%	-0.6%
All Lanes	-3.2%	+0.7%	+0.1%	+1.2%
Count Distribution Shifts Down				
Left Lane	yes	yes	yes	yes
All Lanes	yes	yes	yes	yes
Regression Value Decreases (Decreases/ Total Days)	n.a.	13/14	13/14	n.a.
Relative Use of Left Lane Decreases (Decreases/Total Days)	9/9	13/14	13/14	13/17
Relative Use of Right Lane Increases (Increases/Total Days)	9/9	11/14	13/14	13/17
Variance Decreases (Decreases/Total Days)	7/9	11/14	8/14	16/17

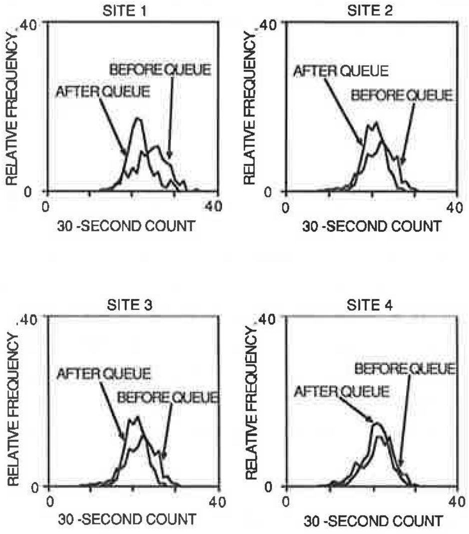


FIGURE 4 Comparison of frequency polygons for left-lane counts.

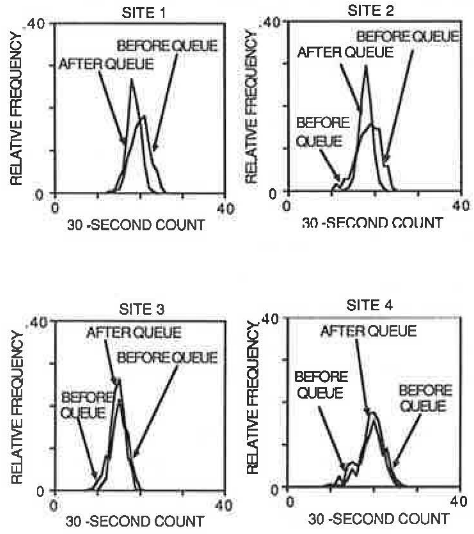


FIGURE 5 Comparison of frequency polygons, mean count, all lanes.

To some extent, this finding may have been the result of using too long an averaging interval at Site 4, where queues sometimes cleared up spontaneously in less than 12 min. Reduction of the averaging interval from 12 to 6 min at Site 4 increased the number of cases in which flow in the left lane decreased when flow broke down from 59 percent (10 out of 17) to 76 percent (13 out of 17). Still, the strength of the two-capacity phenomenon clearly depends on site characteristics.

Although flows averaged across all lanes decreased at least half the time at all sites, they decreased consistently (when averaged over 12 min before and after queue formation) only at Site 1. Thus, even though the existence of the two-capacity phenomenon was confirmed, it is unlikely that the hypothesis can be of value as a basis for metering at more than a small fraction of all bottlenecks.

IMPLICATIONS FOR METERING

There is need for caution even in cases in which the two-capacity phenomenon might be exploited as a basis for metering. Athol and Bullen (*6*) have discussed the relationship between the two-capacity phenomenon and metering strategy. A modification of their analysis was presented previously (*1*) as a conceptual basis for evaluating metering strategies intended to exploit this relationship. Following is a further development of that model, including the results of sensitivity analyses. From these analyses, certain features of the metering strategy were identified as being of special concern in attempts to exploit the two-capacity phenomenon.

Athol and Bullen (*6*) assumed that the probability of flow breakdown during some short time interval is a function of flow. They then considered the probability of flow breakdown over a number of successive time intervals and found the expected time to flow breakdown (after the beginning of the peak) to be a declining function of prequeue flow. They selected an optimum flow (to be produced by metering) to maximize the expected value of uncongested flow for the peak period.

An alternative objective of minimizing total delay in the system has been proposed (*1*). Figure 6 shows a queueing diagram that compares delay for metered and unmetered conditions. In the diagram, the cumulative demand function represents the cumulative flow that would arrive at the bottleneck in the absence of queues or metering. It is assumed that no traffic is diverted, so this diagram is the same regardless of the metering rate.

Any shift in cumulative discharge to the right of this line represents delay, either in a main-line queue or in the queues at the ramp meters. It is assumed that the discharge rate decreases when the queue is formed and that the time it takes the queue to form is a function of the prequeue flow rate. Hence, limitation of the prequeue flow to the metering rate delays queue formation from t_1 to t_2. Area A represents delay experienced because the metering rate is initially less than the unmetered flow; Area B represents delay experienced without metering because the queue discharge rate is less than the metering rate, and metering delays queue formation. Thus, neglecting the effects both of ramp queues and of main-line queues on vehicles not passing through the bottleneck, the expected benefit of metering (i.e., total time saved) is Area B minus Area A.

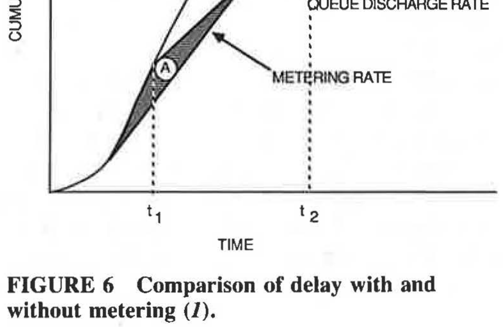


FIGURE 6 Comparison of delay with and without metering (*1*).

Figure 6 indicates that the total time saved by metering should be sensitive to the metering rate, the difference between the maximum unmetered free flow rate and the queue discharge rate, the relationship between flow and time of breakdown, and the difference between the time metering starts and the time unmetered flow would normally break down (referred to as anticipation time). In this model, the metering rate is actually the composite of metering rates at several ramps, and the time metering starts is actually the time the first metered vehicle reaches the bottleneck. It is assumed that the beginning of metering at various ramps is so offset that the first metered vehicles from each ramp reach the bottleneck at approximately the same time.

Calculation of the areas in Figure 6 can be simplified by assuming that the arrival function is linear where it bounds either Area A or Area B. This assumption is conservative; normally it would be assumed that the arrival rate is increasing just before breakdown and decreasing at the time the queue clears. Thus, Area A is overstated and Area B is slightly understated. With this simplification, the expressions for the areas (as functions of the time to flow breakdown), the duration of the peak, the anticipation time, and the arrival, queue discharge, and metering rates are straightforward (because they are combinations of triangles and trapezoids) but somewhat messy. They do not lend themselves to sensitivity analysis by inspection and hence are omitted here; however, they do lend themselves to rather simple numerical analysis.

To carry out these sensitivity analyses, assumptions were made about reasonable ranges for the values of the various parameters of the model and about the nature of the flow-breakdown-time relationship. Important parameters are the ratios of the various flow rates to one another and the ratio of the anticipation time to the duration of the peak, where the peak was defined as the anticipated time that a queue would be present without metering.

Let q_0 be the maximum nonmetered free flow rate (assumed to exist just before flow breakdown in the unmetered case), q_c be the queue discharge rate, q_d be the arrival rate at the end of the peak (which is assumed to be less than q_c because the queue eventually vanishes), and q be the metering rate. With the possible exception of Site 4, it was not possible to observe the values of q_0 directly; however, on the basis of the case study results, the ratio of q_0/q_c would not be expected to exceed about 1.05. Thus, a range of values of 1.00 to 1.05 was assumed for q_0/q_c. Values of q_d/q_c were assumed to range from 0.85 to 0.95, and anticipation times were assumed to range from 0.05 to 0.20 of the duration of the peak.

The flow-breakdown-time relationship posed further difficulties because the form of the relationship is unknown. For Site 1, mean prequeue flow rates were plotted against the time of queue formation to see whether a clear pattern would emerge, but none did. The only conclusion that could be drawn was that the time of flow breakdown varied widely (when compared with the duration of the peak) over a fairly narrow range of flows.

Given this uncertainty, a relationship was assumed that was simple, somewhat flexible, and followed Athol and Bullen's assumption (*6*) that the expected time to flow breakdown is a declining function of flow. The model actually used in the sensitivity analysis involves an instantaneous failure rate of

$$Z(t) = k[(q_0 - q_c)/(q_0 - q)]^m \qquad (1)$$

where q_0, q_c, and q are as defined previously and k and m are scale and shape parameters, respectively. For $k = -T/\ln(0.5)$, where T is the normal duration of queueing in the unmetered case, this equation leads to an expected time to flow breakdown of

$$E(t) = T[(q_0 - q)/(q_0 - q_c)]^m \qquad (2)$$

which varies from 0 at $q = q_0$ to T at $q = q_c$. The shape parameter m expresses the relationship between $E(t)$ and q, with $m = 1$ corresponding to a linear relationship between q and $E(t)$ and $m > 1$ to various relationships that are convex to the origin. Figure 7 shows the effects of parameter m. Again, this particular relationship has no support other than that it is plausible, meets the criteria outlined previously, and is obviously analogous to the failure function resulting from the Weibull distribution.

Given the uncertainty surrounding the flow-breakdown-time relationship and the high probability that there would be considerable scatter in breakdown times in any case, the sensitivity analysis was conducted in two stages. In the first stage, a number of combinations of flow parameters and anticipation times were assumed; for each, the expected benefit was calculated as a function of metering rate, assuming the flow-breakdown–time relationship given by Equation 2 and values of m ranging from 1 to 3. Figure 8 shows an example of the resulting plots of benefit versus metering rate. In the second stage, the expected benefit was calculated as a function of breakdown time for a variety of metering rates. Figure 9 shows an example of the plots resulting from this calculation. Thus, the first stage allows identification of an optimum metering rate, given assumptions about the flow-breakdown–time relationship, whereas the second stage shows the sensitivity of the expected benefit to scatter in the breakdown time for a given metering rate.

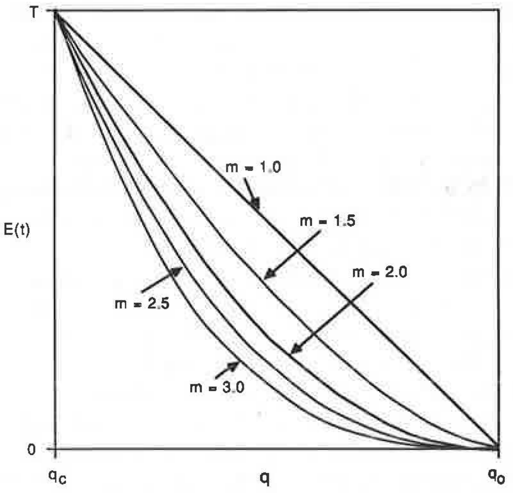


FIGURE 7 Effect of shape parameter m on expected time to flow breakdown.

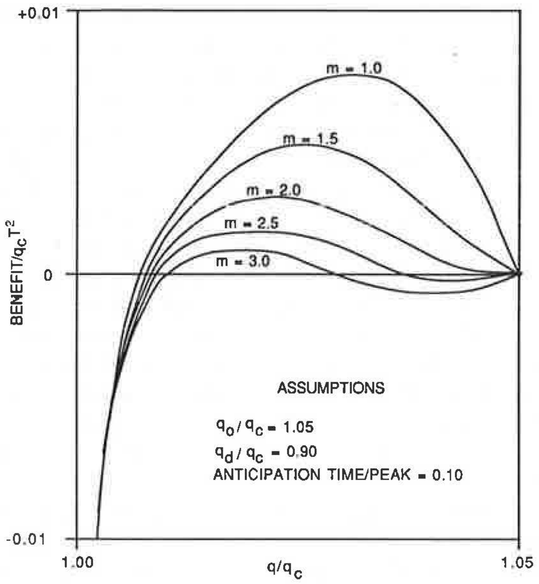


FIGURE 8 Benefit of metering versus metering rate.

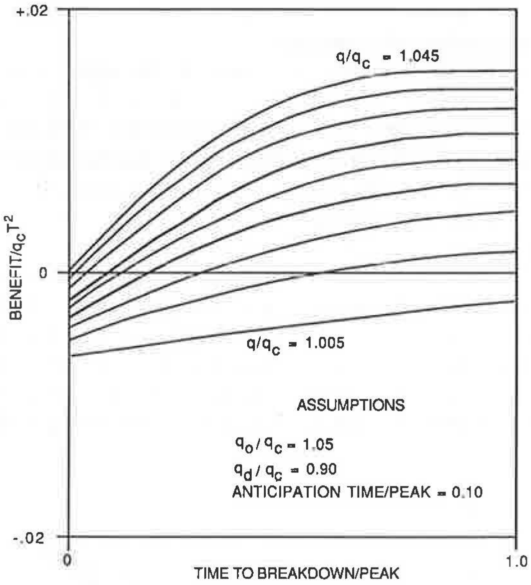


FIGURE 9 Benefit of metering versus time to flow breakdown.

Overall, the sensitivity analyses showed that the expected benefit is a function of the square of the duration of the peak and the volume passing through the bottleneck during the peak (represented here by the queue discharge rate because other flows were expressed as fractions of this). The expected benefit is sensitive to the ratio of q_0/q_c and the duration of the peak, which limit the possible benefit of metering; the anticipation time, which is responsible for the time loss represented by Area A in Figure 6; the metering rate; and shape parameter m. Optimum metering rates are primarily sensitive to shape parameter m. For $m = 1$ (assuming no scatter around the expected breakdown time), the rates are remarkably stable at about $q_c + (2/3)(q_0 - q_c)$. For larger values of m, they decline and are also sensitive to the anticipation time, increasing as it increases. Optimum metering rates appear to be relatively insensitive to scatter in breakdown times, but time savings will be overstated if this scatter is neglected.

From these sensitivity analyses, it appears that, in most cases in which flow across all lanes decreases when queues form, the optimum strategy is to delay metering as long as possible and then meter at a fairly high rate. The major practical dangers appear to lie in beginning metering too soon or setting too low a metering rate because these errors can lead to situations in which metering is counterproductive.

Key practical issues are whether it is possible to (a) anticipate the onset of queueing in the absence of metering well enough to begin metering at a number of different locations just in time to prevent flow breakdown, but not a great deal before, and (b) thereafter hold arrivals at the bottleneck within the fairly narrow range between q_0 and q_c.

The existing metering system in San Diego probably cannot provide the necessary precision. Certainly, the data indicated that there was considerable variation in metered flow arriving at the bottlenecks studied before flow breakdown. This variation was true at Site 1, where flows are extensively metered, as well as at the other sites, where upstream metering was less extensive or nonexistent.

To some extent, this finding may be the result of the details of the current metering strategy; consequently, it may be possible to improve the performance of the system. Work is currently under way to try to improve the precision of the metering provided by the San Diego system by considering the effect of metering rates at upstream locations on subsequent downstream flow. It is expected that one result of this study will be a better understanding of the extent to which target flows can actually be met at bottlenecks. At present, however, whether the system can provide the necessary precision by any conceivable control strategy is not clear.

CONCLUSIONS

The issue of whether ramp metering can increase maximum flow rates through freeway bottlenecks by preventing (or delaying) flow breakdown on the freeway main line has been considered. A test of the hypothesis that bottleneck capacities decrease when flow breaks down has been conducted.

On the basis of evidence from four bottlenecks in San Diego, it was concluded that the two-capacity hypothesis is confirmed, provided it applies to the capacities of individual lanes. The decrease in capacity is not great (at most, about 10 percent for left lanes and 3 percent for flow averaged across all lanes); in three out of the four cases studied, there was no consistent decrease in flow averaged across all lanes when data were aggregated over 12 min before and after flow breakdown. Hence, the two-capacity phenomenon is unlikely to provide a basis for a metering strategy at more than a few locations.

It also appears (on the basis of sensitivity analyses) that, even when there are decreases in total flow, there is substantial risk that metering will be counterproductive unless it is precise, and there is reason to doubt that the necessary precision is possible.

These somewhat pessimistic conclusions about the potential of ramp metering to exploit the two-capacity phenomenon do not rule out the possibility that metering can increase the capacity of freeway bottlenecks in other ways. One of the more startling findings is that merge conflicts were almost never the direct cause of flow breakdown, despite merge rates at three of the sites that far exceeded the supposed capacity of the merge point. Merge conflicts are more likely to lead to general flow breakdown at unmetered sites, especially those at which ramp vehicles arrive at the merge point in bunches. If such is the case, metering may increase bottleneck capacity by eliminating merge conflicts as a cause of flow breakdown. To settle this question, it would be desirable either to study the flow breakdown process at a bottleneck before and after the initiation of metering or to compare breakdown processes at metered and unmetered sites.

ACKNOWLEDGMENTS

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

REFERENCES

1. J. H. Banks. Flow Processes at a Freeway Bottleneck. Presented at 69th Annual Meeting of the Transportation Research Board, Washington, D.C., 1990.
2. C. Pinnell, D. R. Drew, W. R. McCasland, and J. A. Wattleworth. Evaluation of Entrance Ramp Control on a Six-Mile Freeway Section. In *Highway Research Record 157*, HRB, National Research Council, Washington, D.C., 1967, pp. 22–70.
3. L. Newman, A. M. Dunnet, and G. J. Meis. An Evaluation of Ramp Control on the Harbor Freeway in Los Angeles. In *Highway Research Record 303*, HRB, National Research Council, Washington, D.C., 1970, pp. 44–55.
4. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
5. *Traffic Systems Control Handbook*. FHWA, U.S. Department of Transportation, 1976.
6. P. J. Athol and A. G. R. Bullen. Multiple Ramp Control for a Freeway Bottleneck. In *Highway Research Record 456*, HRB, National Research Council, Washington, D.C., 1973, pp. 50–54.
7. L. C. Edie. Car Following and Steady-State Theory for Non-Congested Traffic. *Operations Research*, Vol. 9, No. 1, 1961, pp. 61–76.
8. J. Drake, J. Shofer, and A. May. A Statistical Analysis of Speed-Density Hypotheses. In *Highway Research Record 154*, HRB, National Research Council, Washington, D.C., 1967, pp. 53–87.
9. H. J. Payne. Discontinuity in Equilibrium Freeway Traffic Flow. In *Transportation Research Record 971*, TRB, National Research Council, Washington, D.C., 1984, pp. 140–146.
10. F. L. Hall, B. L. Allen, and M. A. Gunter. Empirical Analysis of Freeway Flow-Density Relationships. *Transportation Research*, Vol. 20A, No. 3, 1986, pp. 197–210.
11. F. L. Hall. An Interpretation of Speed-Flow-Concentration Relationships Using Catastrophe Theory. *Transportation Research*, Vol. 21A, No. 3, 1987, pp. 191–201.
12. J. H. Banks. Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 53–60.
13. V. F. Hurdle and P. K. Datta. Speeds and Flows on an Urban Freeway: Some Measurements and a Hypothesis. In *Transpor-*

tation Research Record 905, TRB, National Research Council, Washington, D.C., 1983, pp. 127–137.

14. B. N. Persaud. *Study of a Freeway Bottleneck To Explore Some Unresolved Traffic Flow Issues*. Dissertation. Department of Civil Engineering, University of Toronto, Ontario, Canada, 1986.
15. B. L. Allen, F. L. Hall, and M. A. Gunter. Another Look at Identifying Speed-Flow Relationships on Freeways. In *Transportation Research Record 1005*, TRB, National Research Council, Washington, D.C., 1985, pp. 54–64.
16. J. H. Banks. *Evaluation of the Two-Capacity Phenomenon as a Basis for Ramp Metering*. Civil Engineering Report Series 9002. San Diego State University, San Diego, Calif., 1990.

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

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

Freeway Capacity Drop and the Definition of Capacity

FRED L. HALL AND KWAKU AGYEMANG-DUAH

Two aspects of the definition of freeway capacity are considered here. The first is whether there is a reduction of maximum flow rates when a queue forms. There appears to be roughly a 6 percent reduction in maximum flow rates after the onset of congestion, but not of the type discussed in the current *Highway Capacity Manual* (HCM). An indirect issue arises during this analysis pertaining to the question of where capacity can properly be measured. The answer, not surprisingly, is that it can only be measured in a bottleneck, and not in a queue. However, the HCM discussion identifies the potential capacity drop on the basis of operations in a queue. The analysis contained here explains why that is inappropriate. The second aspect considered is the distribution over time of maximum flows at a single location. The distribution approximates a normal one reasonably well, with a mean of 6,071 and a standard deviation of 262 vehicles per hour. The unanswered question from the analysis is what portion, or percentile, of the distribution is appropriate to use to meet the HCM definition of capacity, which calls for a value that can "reasonably to expected" to be achieved.

Despite the existence of an explicit definition of capacity in the *Highway Capacity Manual* (HCM) (*1*) and a value for it of 2,000 passenger cars per hour per lane (pcphpl) (under ideal conditions), there remains some contention about the concept, not to mention about the numerical value for it. This paper focuses on two of the critical issues underlying the ongoing debate: the possible existence of a capacity drop after a queue has formed and the distribution of maximum flow rates at a single location over time. Both are addressed with new data. During the investigation of the possibility of a capacity drop, the question is also considered of where it is appropriate to seek such a drop. The answers found in this paper will not resolve the troublesome problems relating to capacity on uninterrupted-flow facilities, but may help to focus the debate.

In the first section of the paper, the background is provided, drawing largely on the HCM and related discussion. In the second section, the data that form the focus for the analysis are discussed. There then follow two sections of analysis, the first dealing with the capacity drop issue, the second with the distribution over time of capacity flows. The conclusions of the paper are presented in the final section.

BACKGROUND

There are three important elements to raise from the background material for a discussion of capacity, all of them arising from the treatment of the concept of capacity in the HCM. The first is the definition itself and two of the key ideas within it. The second is the numerical value given for capacity. The third is the discussion within the HCM that leads to the suggestion of a capacity reduction under congested conditions.

The 1985 HCM defines capacity as "the maximum hourly rate at which persons or vehicles can reasonably be expected to traverse a point or uniform section of a lane or roadway during a given time period under prevailing roadway, traffic, and control conditions" (*1*, p. 1–3). The 1965 HCM definition differed in only a few respects: "the maximum number of vehicles which has a reasonable expectation of passing over a given section of a lane or a roadway in one direction during a given time period under prevailing roadway and traffic conditions" (*2*, p. 5). It went on to say, "In the absence of a time modifier, capacity is an hourly volume." Both definitions refer to maximum rate, reasonable expectation, and prevailing conditions. The major differences are the removal of the presumption of hourly volume and the addition of control to the list of conditions. The presumption of hourly volume was in fact explicitly altered in a subsequent chapter of the 1985 HCM in which freeway capacity is defined as "the maximum sustained (15-min) rate of flow" (*1*, p. 3–3).

Perhaps the most critical idea within these definitions is that of "reasonable expectation." McShane and Roess offer a useful clarification of this idea:

> A rate of flow that can be repeatedly achieved during every peak period for which sufficient demand exists and that can be achieved on any facility with similar characteristics anywhere within North America. It is not the absolute maximum rate of flow ever observed on such a facility. . . . The defined capacity of a facility is that maximum rate of flow which the traffic engineer may be reasonably assured of being able to achieve day in and day out anywhere in North America. (*3*, pp. 192–193)

Two points come from this discussion, the first very useful, the second arguable.

The useful point is that capacity is not the absolute maximum flow observed. Rather, it is a rate of flow that can be repeatedly achieved, day in and day out. This brings to mind the notion of a distribution of such flows. Capacity is not the extreme value of this distribution, but it is not clear what portion of the distribution should represent capacity. Elsewhere, the HCM states that the manual utilizes "national average" traffic characteristics, and that "the recommended value of 2000 pcphpl represents a national average" (p. 2–2). But if the distribution is normal, the average is not reached roughly half of the time. It will be valuable to look at the distribution, which will be done later in this paper.

Departments of Geography and Civil Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4K1.

The arguable point is McShane and Roess's assertion that the rate of flow must be achievable on any similar facility in North America. This would suggest that it is not useful to talk about the capacity of a specific facility, that is, of any particular highway. In contrast, we suggest that it is useful to distinguish between the capacity and a *type* of facility (e.g., a six-lane level freeway) and the capacity of a *specific* facility. In this paper, we intend to analyze data from one specific facility and to draw conclusions about that facility. Whether these conclusions apply to other similar facilities is always an open question, but we will argue that in several important respects, they do. Whether the numerical value of capacity applies to any similar facility in North America is a different question. Nevertheless, we suggest that it is more practical to operate this particular facility in light of its capacity than to use some continent-wide capacity value.

The numerical value of capacity has already been cited as 2,000 pcphpl. The 1965 HCM stated that the "largest number of vehicles that can pass a point . . . averages between 1,900 and 2,200 passenger vehicles per hour" (*2*, p. 75). Having stated this, it went on to say "the capacity of a multi-lane highway under ideal conditions is considered to be 2,000 passenger vehicles per lane per hour" (p. 76). The 1985 Manual, as quoted earlier, justifies the same number as being in some sense a "national average." Maintaining the same numerical value implies that there has been no net effect of the downsizing of vehicles over the last generation or of increasing driver experience with freeway driving. The numerical value is of particular interest at this time because there is a proposal to change the HCM capacity for multilane rural roads to 2,300 pcphpl (*4*), which would suggest that such a facility can handle more traffic than a freeway, an idea that is counterintuitive. Other recently published work (*5*) suggests sustainable freeway flows similar to the 2,300 value, albeit in only one location.

The final important element in the HCM regarding freeway capacity is the suggestion of a capacity reduction for congested operations. This is phrased as follows:

> Some researchers have fit continuous curves through density-flow data, yielding a single maximum flow rate. Others have projected discontinuous curves through data, with one curve treating stable flow points, and another unstable or forced flow points. In these cases two maxima are achieved, one for each curve. All such models indicate that the maximum flow rate for the stable flow curve is considerably higher than that for the unstable flow curve, perhaps as much as 200 vph higher. (*1*, p. 2–23)

This discussion is clearly based on density-flow behaviour *within* a queue, because that is the only place where "forced flow points" can be observed. It is our contention that this discussion, although based on a number of published studies [for example, those of Drake et al. (*6*) and Ceder and May (*1*)], is based on a mistaken premise about where the data are collected. Reduced capacity within a queue says nothing about capacity in the bottleneck downstream of the queue. The hypothesis to be tested in this paper is that there is a capacity drop in the bottleneck flow at the time a queue forms upstream.

DATA

Three issues about the data will be discussed: the location at which to collect data, the time intervals to use for analysis, and the estimation of truck percentages. The location selected is a part of the Freeway Traffic Management System (FTMS) on the Queen Elizabeth Way (QEW) in Mississauga, west of Toronto, Ontario, just downstream of the Cawthra Road interchange (Figure 1). The rationale for selecting this location was explained by Hall and Hall (*5*) following ideas by Hurdle and Datta (*8*). Cawthra Road is the final entrance ramp in a series and experiences daily congestion immediately upstream of it. Downstream of the ramp, there is a straight, level section of roadway for 2 km, after which there is a downhill section. There is an additional entrance ramp 40 m downstream of Station 25, but the road widens to four lanes with the addition of this ramp, so no queue occurs here. The presence of a queue upstream of Cawthra Road for an extended period confirms that there is sufficient demand for capacity operations in the downstream section. Indeed, the queue is present because there is more demand than the system can accommodate in this downstream section. This location in the bottleneck is clearly the only place to look for capacity operations.

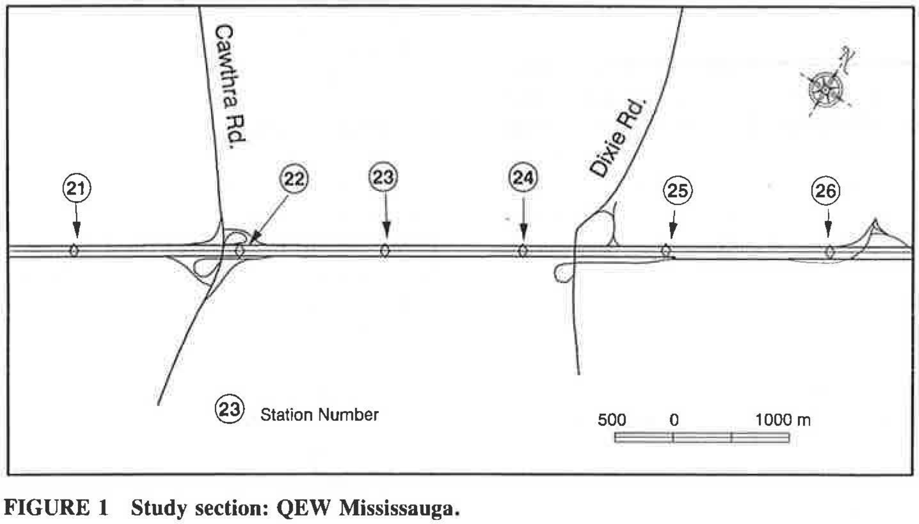


FIGURE 1 Study section: QEW Mississauga.

The analysis by Hall and Hall used the only station downstream of Cawthra Road at that time, which was located less than 200 m from the end of the on ramp. Since that time, the FTMS has been extended eastward so that there are now four additional stations, two of them within the section of interest. (After the second, the freeway widens to four lanes.) We have concentrated our bottleneck analysis on Station 25 (Figure 1), roughly 1.5 km downstream of the end of the entrance ramp. (For comparison with the earlier analysis, note that the system has since been renumbered, so that what was Station 22 in the paper by Hall and Hall is now Station 23.) Station 23 is very close to the end of the ramp, and Station 24 is only a single-loop station, so it does not provide speeds.

The data are available for 30-sec intervals 24 hr a day. Data for the morning peak period were used. The 1985 HCM refers to 15-min intervals in its definition of freeway capacity, but that is not a useful interval for the present analysis. Instead, three levels of detail are used: 30-sec values, 5-min values, and the full peak period. The 30-sec data were necessary for determining the beginning and end of a queue, and for a closer examination of the flow process, following Banks's approach (*9*). For the discussion of the distribution of capacity flows over different days, the unit of analysis is the peak period, which extends for 2 to 3 hr.

The focus of this study is freeway operations during good weather conditions. The study has been limited to normal weekdays, during which regular commuters use the facility, so the driver population is one that is quite familiar with the road and the usual traffic patterns. Data were available for the period April 25, 1990, through May 30, 1990. On May 7, peak-period flows were exceptionally low (5,400 vehicles/hr), suggesting that something was happening upstream that was not detected despite the use of closed-circuit television. This day was therefore left out of the analysis. May 21 was a holiday, so was omitted. There was rain during the peak period on May 16, 17, and 29, so these days were also omitted. For the remaining days, there were no incidents recorded by the operators that would affect this section. In addition, a cursory inspection of the Station 26 data indicated that it was consistently operating at high speeds, meaning that there was no downstream congestion that might not have been mentioned in the operators' log. In total, then, 20 days of data are used in these analyses, representing nearly ideal conditions.

The Mississauga FTMS collects flow data, but does not separately count truck volumes. To obtain these numbers, manual counts are needed. One such count was obtained on May 29, 1990. Although this was a rainy day, and not included in the analysis, the truck percentages can be taken to be representative of the other days, because the rain does not likely change the total volumes, but merely affects operating conditions. Truck percentages for the period 6:20 to 8:45 a.m. averaged 6 percent. There was no noticeable trend in the truck percentage over time.

CAPACITY DROP ISSUE

Two versions of the capacity drop issue are discussed. The first looks at the possibility of a capacity drop as described in the HCM and discussed above. The second looks at the idea of a capacity drop within the bottleneck flow.

Capacity Drop per HCM

The HCM quotation about the capacity drop arises from data like that shown in Figure 2, which is from Station 22 (see Figure 1). There are both a congested branch and an uncongested branch of the curve. There is even a visible gap between the two. According to the HCM discussion, the "stable" (i.e., uncongested) flow has a somewhat higher maximum than the "unstable" (i.e., congested) flow. But this station is not operating at capacity during "unstable" flow. The station is immediately upstream of a major entrance ramp.

Figure 3 presents data from Station 25 in the bottleneck for the same time period as that shown in Figure 2, namely, 5:30 a.m. to 10:00 a.m., covering the brief buildup to capacity (pre-queue data in the figure) over 2 hr of queue discharge flow and the flows after the morning rush has dissipated (post-queue). Because Station 25 is never in queue, there are no "unstable" data (i.e., high values of occupancy), and hence what would be the right-hand side of the flow-occupancy curve simply does not occur. Consequently there can be no drop in capacity described by a discontinuity of the curve. Yet this is the section of the freeway that is operating at capacity, and therefore the only place where one should be able to see if there is a drop in capacity.

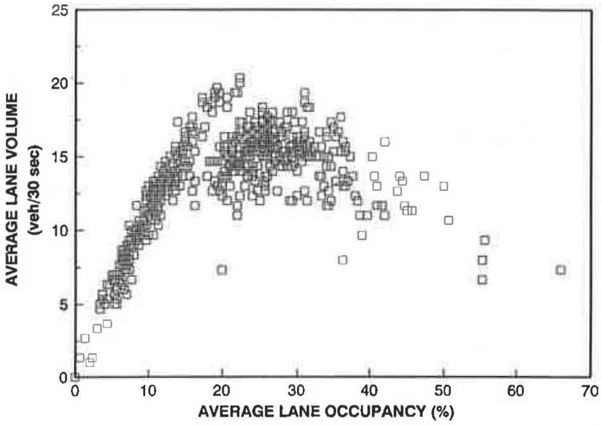


FIGURE 2 Flow versus occupancy: Station 22, May 11, 1990, 5:30–10:00 a.m., 30-sec data.

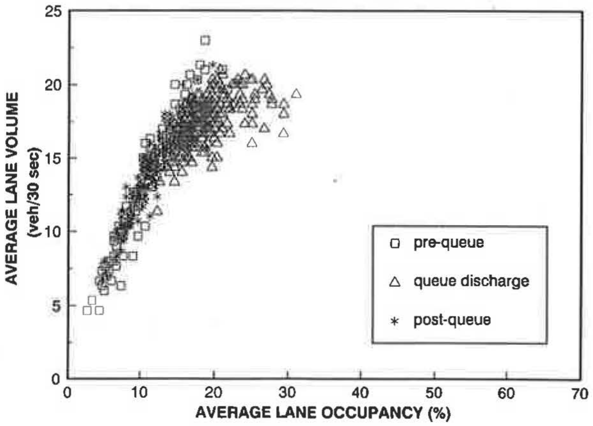


FIGURE 3 Flow versus occupancy: Station 25, May 11, 1990, 5:30–10:00 a.m., 30-sec data.

Closer inspection of the data for the three time periods within Figure 3 suggests that the pre-queue flows certainly occur at lower occupancies than do comparable queue discharge flows, and might reach a slightly higher maximum. The post-queue flows line up more closely with the pre-queue flows, and also occur at lower occupancies than comparable queue discharge flows. This occupancy difference is of course sensible if the queue discharge flows occur at lower speeds, which they do, for reasons explained by Persaud and Hurdle (*10*).

The data behind the HCM passage quoted earlier were collected in the queue upstream of a bottleneck, not in the bottleneck itself. The observed discontinuous curve simply reflects, for example, the volume of traffic entering an entrance ramp, which has caused the queue in the first place. The only place to look for a capacity drop (as opposed to simply a reduction in flow) is in the bottleneck itself. Banks did this (*9*), and found a minimal drop, but a statistically significant one, given that it occurred in eight out of nine cases. His method, comparing mean flows before and after queue formation, has been adopted here.

Capacity Drop Within the Bottleneck

Banks averaged the 30-sec flow rates before the onset of congestion with those for a similar interval immediately after congestion began. He used a "speed drop associated with the formation of the upstream queue" (*9*, p. 12) to identify the beginning of congestion. For the present analysis, queue presence was identified from the detector data at Station 22, where the queue first forms (see Figure 1). This station provides only volume and occupancy, not speed. Because it is known that the estimate of speed from volume and occupancy is not an unbiased one (*11*), and videotapes to calculate speeds were not available, the presence of the queue had to be determined directly from the volume and occupancy data.

Averages of flow and occupancy across the three lanes were used. The two tended to vary together in the period before congestion and to diverge during the congested period. Determining the exact beginning and end of congestion was, however, difficult from these numbers, so the ratio of occupancy to flow was used. Three values of the ratio were tested for the threshold level: 1.0, 1.1, and 1.2. A ratio of 1.0 gives a longer duration of bottleneck flows, some of which were very low, suggesting that demand was below capacity. A ratio of 1.2 excludes sustained periods (10 min) of high flows (5,800 vehicles/hr or more). A ratio of 1.1 or above persisting for 3 min was selected as the criterion for the identification of the start of a queue.

To consider capacity drop due to queue formation, it is necessary to restrict inspection of pre-queue flows to that period when demand is equal to capacity. If earlier, lower flows were included, the mean pre-queue flow rate would be reduced, biasing the results. The criterion for identifying the start of the queue is intentionally conservative. There is a considerable period before the start of the queue when the occupancy-flow ratio approaches congested conditions but does not maintain it. This is the period when the highest flow rates are being achieved, and therefore the ones to be kept distinct from queue discharge flow. In the following discussion, this portion of the pre-queue period is referred to as the transition period, that is, the period before the formation of a queue when demand is very high, resulting in high flows.

The ending time for this transition period of pre-queue flow is easy to identify in that it coincides with the start time for the queue. Identification of the start time for the transition period is more difficult. The occupancy-flow ratio for Station 22 could not be used because there was very little variation in the ratio below 1.1. Hence Station 25 data were used. In order to coordinate times when two stations were being used, it was necessary to take into account the travel time between them. The distance between the two stations is 2250 m. Calculation of the mean speeds during the period of high flows at Station 25 before the queue start time indicates that 1.5 min should be added to the start time of the queue at Station 22 to give the time of arrival of queue discharge flow at Station 25. The start time for the transition period was determined as follows, using the flow data at Station 25 and working backward in time from the interval when the queue discharge flow reached Station 25. When relatively low flows were observed in four consecutive 30-sec intervals, the latest time of such flows was taken as the tentative beginning of the transition period. Where two periods of comparatively high flows were separated by a sharp drop in flows for several consecutive 30-sec intervals, the *t*-test, or an approximation of it (depending on whether the variances in the two time periods were equal or not), was computed. When the difference in the means of the flow rates for the two periods was found to be significant, the tentative start time was accepted as the start of the transition. When the difference was found not to be significant, the two time periods were combined and were taken as one period.

The time at which the queue ended was identified in a similar fashion to the start of the transition period. A tentative ending time was identified on the basis of the first time that the occupancy-flow ratio at Station 22 fell below 1.0. Then attention was turned to Station 25. If low flows occurred there before the corresponding time, the same *t*-test approach was used to test average flows before and after the tentative end time. If the flows were equal, the queue end time was moved earlier accordingly.

With the pre-queue and queue discharge time periods identified, the task was then to determine whether the mean flow rates differed between the two periods. The two test statistics applicable in this case are the *t*-test for samples with equal variance and the approximation of *t*, which tests for the significance of difference between two means for samples with unequal variances. *F*-test results showed that at the 1 percent level, the variances for only 7 days were significantly different. At the 5 percent level, 10 of the 20 days have significantly different variances.

One-tailed tests were run for both *t* and its approximation, with emphasis on the "correct" *t* for each day, given its *F*-test result at the 5 percent level. There was a significant difference in the mean flow rates, pre-queue to queue discharge, for 12 days at the 5 percent level using the estimate for *t* and for 16 days using the *t*-test (Table 1). The "correct" test for each day (identified by an asterisk in Table 1) shows 12 days with a significant difference at the 5 percent level. At the 1 percent level, the difference in mean flows is significant for 10 days for the approximation of *t*, for 13 days using the *t*-test, and for 10 days using the "correct" test. There was only one day where queue discharge flow was higher than

TABLE 1 ONE-TAILED *t*-ESTIMATE AND *t*-TEST TO COMPARE MEAN VALUES OF 30-SEC RATES FOR PRE-QUEUE AND QUEUE DISCHARGE FLOWS

PRE-QUEUE DURATION			QDF DURATION		t	SIGNIFICANCE		t	SIGNIFICANCE	
Date	flowrate	minutes	flowrate	minutes	estimate	1%	5%	test	1%	5%
900425	6423	9.5	6173	156.0	1.48	no	no*	1.82	no	yes
900426	6498	13.5	5980	188.0	3.56	yes	yes	4.20	yes	yes*
900427	6400	9.0	6127	132.5	1.29	no	no*	2.14	no	yes
900430	6526	10.5	5919	180.0	4.40	yes	yes	4.78	yes	yes*
900501	6631	13.5	5995	183.0	3.59	yes	yes*	5.12	yes	yes
900502	6202	25.0	5920	179.5	2.77	yes	yes*	3.09	yes	yes
900503	6388	19.0	6097	161.0	2.42	no	yes*	2.95	yes	yes
900504	5750	6.0	6196	164.5	-1.42	no	no*	-2.54	no	no
900508	6669	16.5	5978	187.0	6.35	yes	yes	6.20	yes	yes*
900509	6471	13.0	6137	164.0	2.01	no	no*	2.74	yes	yes
900511	6505	17.0	6195	142.0	2.14	no	yes*	2.74	yes	yes
900514	6415	17.5	5959	176.5	3.75	yes	yes	3.77	yes	yes*
900515	6669	31.5	6199	118.0	4.98	yes	yes	5.20	yes	yes*
900518	6298	15.5	6218	132.5	0.62	no	no	0.69	no	no*
900522	6538	14.5	5902	170.0	5.31	yes	yes	5.10	yes	yes*
900523	6380	9.0	6214	148.0	0.94	no	no	1.11	no	no*
900524	6413	18.0	6313	128.5	0.64	no	no*	0.83	no	yes
900525	6424	25.5	6173	153.0	2.71	yes	yes	2.82	yes	yes*
900528	6484	14.5	5991	171.5	3.55	yes	yes*	4.36	yes	yes
900530	6123	17.5	6066	164.0	0.54	no	no	0.60	no	no*
Weighted mean	6432		6075							

* correct t calculation given the F-test results at the 5% level.

that in the transition period. Because this is an even split, it was deemed appropriate to consider the trend in the results, as Banks did. If the two means are equal, the probability of obtaining only one higher value for queue discharge flow in 20 observations is only 0.00001. Hence these results appear to support Banks's conclusion that there is a drop in achievable maximum flow rates when the queue forms. A comparison of the overall weighted mean flows in the two time periods, taken over the 20 days, shows a difference of 357 vehicles/hr, or 5.8 percent.

Two aspects of this result warrant discussion. First, there is the question of how these results relate to those reported by Hall and Hall (*5*). Second, the drop in flows when the queue forms highlights a problem associated with the concept of sustainable flows. Hall and Hall selected the maximum 40-min flow rates as representing capacity, which gave a value of about 6,500 vehicles/hr as the capacity for the same bottleneck. This value is roughly consistent with the 6,400 vehicles/hr found here as the average for pre-queue flows, but the duration of the pre-queue flows is not as long as the 40 min found in their paper. There are two possible reasons why the high flows did not last so long. The first is simply that Hall and Hall did not have as good information with which to ascertain the time of queue formation. The second is based on the idea that flows were not quite so heavy in the data they used (late 1987 and one day in July 1988), with the result that these high maximum flow rates were sustainable for a longer period before queue formation. This explanation is consistent with the HCM statement that Level of Service E (i.e., capacity operation) is "unstable." It will last longer on some days than on others.

However, given the information now available, it seems likely that the 40-min values that Hall and Hall identified as capacity represent the pre-queue flows. First, the current data show that it can take up to 30 min for a queue to stabilize. Second, a comparison of the observed speeds and reconsideration of the distances involved support the interpretation that the speed drop in their data represented the onset of queue discharge flow, not of operations in a queue as they suggested. For the current Station 25 data, the observed mean speed was 85 km/hr in the transition period and 73 km/hr for the queue discharge flows. Hall and Hall stated that their station was 800 m downstream of the entrance ramp, but in fact it is 650 m downstream from the beginning of the merging lane and less than 200 m downstream of the end of that lane. Hence their lower speeds are consistent with queue discharge flow, as defined by Persaud and Hurdle's expectations for speed recovery downstream of a queue (*10*).

But this correction of the Hall and Hall result serves only to highlight the difficulties with the HCM definition of capacity as "sustainable" flows. In Chapter 3, "Basic Freeway Segments," the HCM states that "freeway capacity is the maximum sustained (15-min) rate of flow" (*1*, p. 3–3). Clearly, the pre-queue flows meet this definition. In the Hall and Hall paper, such flows lasted for 40 min; in this paper they last from 6 min to over 30 min and 10 of the 20 exceed the 15-min requirement. (Two more fall short by only 30 sec). These 10 cases had average flows of between 6,123 and 6,660 vehicles/hr. According to the HCM definition of a 15-min sustained flow, these values represent capacity, and the 6,500 value put forward by Hall and Hall is valid. Hurdle and Datta (*8*), however, have suggested that capacity should be defined as queue discharge flow. On the basis of the start and end times, there was only one instance in which the queue discharge flow lasted less than 2 hr (on May 15, when it fell 2 min short of 2 hr). In four instances, it extended for 3 hr or more. There is no question that queue discharge flow is sustainable. If both types of flow are "sustainable," which one represents capacity? We return to this question later.

DISTRIBUTION OF QUEUE DISCHARGE FLOWS OVER TIME

For analysis of the behavior of queue discharge flows over time, data on the 5-min flow rates from all 20 available days were taken together. For convenience and comparability, the 5-min intervals for each day were begun on the hour. The

5-min interval that contained the start of queue discharge flow was not included, nor was the interval containing the end of queue discharge flow. All intervals between these two were included in the analysis. The 5-min counts for each of the three lanes were summed and multiplied by 12 to get the hourly flow rate across all three lanes. As stated earlier, no truck correction was made in these values. Table 2 contains examples of such data from 6 days.

The mean, standard deviation, and duration of queue discharge flow for the 20 days based on these 5-min data appear in Table 3. The values differ slightly from those in Table 1 because of the omission of a small number of 30-sec intervals. The daily mean queue discharge flows ranged from 5,891 to 6,307 vehicles/hr across the section. There were eight cases of mean flows under 6,000 vehicles/hr. Three of these had relatively high standard deviations, over 300 vehicles/hr. The high standard deviations seem to be a consequence of very low flow rates: the two with the highest standard deviations recorded flow rates below 5,000 vehicles/hr for two consecutive 5-min intervals.

Given this distribution of mean queue discharge flow over 20 days, what should one say is capacity? For eight days there were average flow rates below the HCM values; for 12 days there were averages above the HCM values. Twenty observations do not make for a very useful histogram, so all 620 of the 5-min volume counts have been used to create one (Figure 4) in which the data have been grouped in cells of 100-vehicle/hr width and frequencies plotted at the midpoint of the cells. The modal value (the peak of the diagram) is just above 6,000, as are the median (6,072) and mean (6,071). This frequency distribution represents a near-normal distribution, with a Pearson coefficient of skewness of only −0.011. It is reasonable to assume that a distribution of capacities for a larger number of days would show somewhat the same shape, although it might be a bit narrower. This frequency distribution is for queue discharge flows, but a similar diagram could be constructed for pre-queue flows. What is the proper segment of either distribution to select as "capacity," given the HCM wording of "reasonable expectation"? The expected value is, of course, the mean. In that case, these data suggest a discharge flow rate of 6,071 vehicles/hr. However, selecting the mean, with a normal distribution, implies that capacity will not be achieved half of the time. Does that suggest that a lower percentile of the distribution should be selected? The answer is not obvious, but certainly it will be valuable to identify explicitly the fact that there is a distribution of achievable maximum sustained flows at any given location, whether capacity is queue discharge flow or pre-queue flow. The same kind of analysis holds over space as well as over time.

TABLE 2 EXAMPLES OF QUEUE DISCHARGE FLOWS AT STATION 25: 5-MIN VALUES

End time	900427 flow/h	900501 flow/h	900508 flow/h	900509 flow/h	900515 flow/h	900524 flow/h
6:30	6636					
6:35	6372	6228		5664		
6:40	6324	6204	6000	6060		
6:45	6012	6132	6060	6084		7092
6:50	6300	5760	6060	5832		6252
6:55	5964	6228	6168	5988		6204
7:00	6120	5868	5868	6276	6156	6216
7:05	6216	5976	5880	6048	5820	6084
7:10	6228	5856	5892	6144	6084	6168
7:15	6180	6240	6156	6240	6468	6468
7:20	6072	5868	5676	6132	6264	6528
7:25	5880	5988	5928	6048	6264	6408
7:30	6024	5868	6084	6288	6252	6420
7:35	6276	6348	5640	6204	6384	6300
7:40	6156	5676	5772	6264	6708	6216
7:45	5964	5748	6012	6336	6168	6252
7:50	6432	6132	5988	6444	5712	6252
7:55	6192	6036	6012	6480	6468	6168
8:00	6084	5628	6336	6120	6264	6420
8:05	6084	5760	6056	6024	6072	5772
8:10	5844	5844	5892	6036	6288	6324
8:15	6072	5988	6180	6064	6156	6600
8:20	6144	5748	6012	6108	6024	6432
8:25	6276	5976	6384	6216	5952	6144
8:30	5952	5904	6240	5976	6132	6480
8:35	5988	6384	5952	5976	6300	6360
8:40	6000	5976	5868	6132	6036	6048
8:45		6012	6348	6336	6324	6060
8:50		6276	5796	6060	6048	
8:55		6156	6288	6192		
9:00		6288	5880	6420		
9:05		6264	5892	6048		
9:10		5940	6084	6060		
9:15		5580	6360			
9:20		5904	5700			
9:25		5844	5040			
9:30						
9:35						
9:40						
Mean	6121	5989	5985	6134	6189	6307
Std. Dev.	153	215	258	172	218	244
Duration(min)	130	175	170	160	115	125

TABLE 3 MEANS, STANDARD DEVIATION, AND DURATION OF QUEUE DISCHARGE FLOWS AT STATION 25: 5-MIN VALUES

Date	Mean flowrate	Standard Dev.	Duration minutes
900425	6170	181	150
900426	5976	301	185
900427	6121	153	130
900430	5913	173	175
900501	5989	215	175
900502	5921	235	180
900503	6094	184	155
900504	6205	182	160
900508	5985	258	170
900509	6134	172	160
900511	6198	232	140
900514	5959	397	175
900515	6189	218	115
900518	6227	248	125
900522	5891	365	165
900523	6208	191	145
900524	6307	244	125
900525	6169	182	150
900528	5970	160	165
900530	6062	188	160

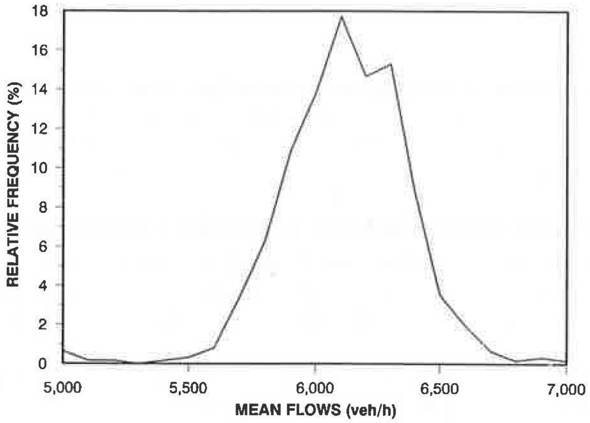


FIGURE 4 Frequency histogram for 5-min queue discharge flows, Station 25.

This discussion of queue discharge flows also raises the concept of stable flows, as used in the HCM: operations in Level-of-Service E are said to be "extremely unstable," in the sense that any disruption "establishes a disruption wave which propagates through the upstream traffic flow" (*1*, p. 3–10). It is indeed upstream that the effects are felt, where the queue forms. Downstream, in the bottleneck (such as Station 25 in this paper), flow rates remain high (although not quite as high as before), and speeds drop somewhat, with the actual value depending on how far downstream of the queue the measurement is taken. Downstream, where capacity flows are actually occurring, operations are in fact extremely stable, not unstable, as shown by their 2- to 3-hr duration. It therefore seems inappropriate for the Level-of-Service E description to assert that operations at or near capacity are unstable.

Note that the HCM also states that "average travel speeds at capacity are approximately 30 mph" (about 50 km/hr) (p. 3–10). These data contradict that assertion also. At capacity during pre-queue operations, speeds are reduced only slightly from free-flow values. In the bottleneck during queue discharge flow, speeds depend on the distance downstream. In the queue, speeds may well be roughly 50 km/hr (or less), but the queue is not operating at capacity. There is a restriction, such as an entrance ramp with high volume, that is forcing it to operate at lower flow rates.

CONCLUSIONS

There are four conclusions from this analysis, one relating to the notion of a two-regime flow-occupancy diagram, the other three relating to the definition of capacity. This section closes with some comments about future research.

For a flow-occupancy diagram to be able to show two regimes, the data must have been obtained within a queue (to get the right-hand portion of the curve). Figure 3, showing the flow-occupancy data for station 25, shows only the left-hand side of such a curve, reaching an occupancy of perhaps 28 percent. Similar figures from stations in a queue, such as Figure 2, show a lower maximum for the congested data than for the uncongested precisely because of the bottleneck effect: a queue forms at the location because something else (usually an entrance ramp) has taken up a portion of the available capacity, leaving a reduced flow possible within the queue. The only legitimate place to look for a capacity drop is in the bottleneck. The notion of two capacities may be right, but it has been raised for the wrong reasons.

The first conclusion about capacity is that there is a capacity drop in the bottleneck. Once a queue has formed upstream, the bottleneck location does not handle as many vehicles as it did before the queue formation. The demand is clearly there, as shown by the presence of the queue, so the reduced flow is a consequence of the way drivers accelerate away from the queue. From these data, the reduction in flow appears to be about 5 to 6 percent, from just over 6,400 vehicles/hr to just under 6,100, on average.

The second conclusion is about the duality of capacity. One capacity is the pre-queue flow, which can last for 30 min and more. Queue discharge flows can last for 2 to 3 hr. Both then meet the HCM requirement for sustainable flow. The important issue is the practical one: once congestion occurs on a facility, it is the queue discharge flow rate that will govern the time to recovery. If queues can be delayed during high flows, there is a 5 to 6 percent "bonus" flow available.

The final conclusion about capacity should be its numerical value, given the previous conclusion. However, the analysis of the distribution of queue discharge flows over time shows that there is indeed a distribution, not simply a single value achieved on all similar days. If the mean of that distribution is taken, then the value for queue discharge flow for this location is 6,071 vehicles/hr. Applying the truck percentage found by manual counting on the one day (6 percent) and a truck equivalent of 1.7 (*1*, p. 3–13) for a general segment of level terrain, capacity is just under 6,700 passenger cars/hr. Using the mean value from Table 1 for the pre-queue flows gives 6,432 vehicles/hr, or 7,088 passenger cars/hr. This corresponds favorably with the value of 2,300 pcplph for rural multilane highways (*4*). However, it is not obvious that the mean is the proper value to take from the distribution. More consideration is needed.

More research is needed on several of the topics addressed in this paper. The most important is probably the validation at other locations of the conclusion found with these data that there is a drop in capacity when a queue forms. As well, it would be useful to have equally well-specified information on the numerical value of capacity (both pre-queue and during queue discharge) at other locations. On a more theoretical note, Athol and Bullen (*12*) suggested 18 years ago that the probability of transition from uncongested to congested flow depends on the value of the flow. The discussion in this paper offers a different interpretation, namely, that operations at quite high flows fed by a queue can be sustained almost indefinitely (up to 3 hr in these data) without breaking down further. Instead, breakdown occurs at the entrance ramp where total flow exceeds capacity. It would be worthwhile to investigate in more detail which of these models is more reliable for freeway operations.

ACKNOWLEDGMENTS

The work reported in this paper represents the early stages of the analysis for Agyemang-Duah's master's thesis. Since presentation of this paper, but before its publication, he has completed his thesis (*13*), and a paper based on the full 52 days of data has been presented at the Karlsruhe conference (*14*). This work was supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada. The data were provided by the Freeway Traffic Management Section, Ontario Ministry of Transportation.

REFERENCES

1. *Special Report 209: Highway Capacity Manual 1985*. TRB, National Research Council, Washington, D.C., 1985.
2. *Special Report 87: Highway Capacity Manual 1965*. HRB, National Research Council, Washington, D.C., 1965.
3. W. R. McShane and R. P. Roess. *Traffic Engineering*. Prentice Hall, Englewood Cliffs, N.J., 1990.
4. W. R. Reilly, D. W. Harwood, J. M. Schoen, R. O. Kuehl, K. Bauer, and A. D. St. John. *Capacity and Level of Service Pro-*

cedures for Multilane Rural and Suburban Highways. Preliminary draft final report. NCHRP, TRB, National Research Council, Washington, D.C., 1988.

5. F. L. Hall and L. M. Hall. Capacity and Speed Flow Analysis of the QEW in Ontario. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 108–118.
6. J. S. Drake, J. L. Schofer, and A. D. May. A Statistical Analysis of Speed Density Hypotheses. In *Highway Research Record 154*, TRB, National Research Council, Washington, D.C., 1967, pp. 53–87.
7. A. Ceder and A. D. May. Further Evaluation of Single- and Two-Regime Traffic Flow Models. In *Transportation Research Record 567*, TRB, National Research Council, Washington, D.C., 1976, pp. 1–15.
8. V. F. Hurdle and P. K. Datta. Speeds and Flows on an Urban Freeway: Some Measurements and a Hypothesis. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 127–137.
9. J. H. Banks. Flow Processes at a Freeway Bottleneck. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 20–28.
10. B. N. Persaud and V. F. Hurdle. Some New Data That Challenge Some Old Ideas About Speed-Flow Relationships. In *Transportation Research Record 1194*, TRB, National Research Council, Washington, D.C., 1988, pp. 191–198.
11. F. L. Hall and B. N. Persaud. An Evaluation of Speed Estimates Made with Single-Detector Data from Freeway Traffic Management Systems. In *Transportation Research Record 1232*, TRB, National Research Council, Washington, D.C., 1989, pp. 9–16.
12. P. J. Athol and A. G. R. Bullen. Multiple Ramp Control for a Freeway Bottleneck. In *Highway Research Record 456*, TRB, National Research Council, Washington, D.C., 1973, pp. 50–54.
13. K. Agyemang-Duah. Investigation of Some Unresolved Issues in Freeway Capacity. M.A. thesis. Department of Geography, McMaster University, Hamilton, Ontario, Canada, 1991.
14. K. Agyemang-Duah and F. L. Hall. Some Issues Regarding the Numerical Value of Capacity. In *Highway Capacity and Level of Service: Proceedings of the International Symposium on Highway Capacity, Karlsruhe* (U. Brennolte, ed.), Balkema, Rotterdam, 1991, pp. 1–15.

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

Proposed Analytical Technique for Estimating Capacity and Level of Service of Major Freeway Weaving Sections

MICHAEL J. CASSIDY AND ADOLF D. MAY

Weaving typically occurs where merging traffic movements cross over diverging movements. The prevalence of weaving areas on freeways warrants the need for analytical techniques that can reliably analyze or design these critical freeway components. However, previous research at the University of California suggested that existing analytical procedures may not predict weaving operation with a sufficient degree of reliability. Consequently, a more reliable procedure for evaluating weaving performance was developed. Unlike most existing procedures, the proposed technique evaluates traffic flow behavior in individual lanes of the weaving section. The procedure is applicable to major weaving areas, a subset of all weaving configurations. The proposed procedure predicts vehicle flow rates in critical regions within the weaving section as a function of prevailing traffic flow and geometric conditions. Predicted flows are then used to assess the capacity sufficiency and level of service of a subject weaving area. The model itself was developed using large amounts of empirical and simulation data, collected from a number of sites throughout California. Some of the more interesting traffic flow characteristics empirically observed on a single weaving site are highlighted. Moreover, the basic format of the proposed procedure is presented.

The 1985 Highway Capacity Manual (HCM) (*1*) defines weaving as "the crossing of two or more traffic streams traveling in the same general direction along a significant length of highway, without the aid of traffic control devices." For freeway facilities, a weaving section is typically formed "when a merge area is closely followed by a diverge area, or when an on-ramp is closely followed by an off-ramp and the two are joined by an auxiliary lane." Four types of traffic movements will generally travel on a freeway weaving section:

- Freeway-to-freeway traffic (a nonweaving movement),
- Freeway-to-off-ramp traffic (a weaving movement),
- On-ramp-to-freeway traffic (a weaving movement), and
- On-ramp to off-ramp traffic (a nonweaving movement).

These four traffic movements are shown in Figure 1.

Considerable lane-changing activity typically occurs on weaving sections as motorists access lanes appropriate for their destinations. Vehicular conflicts occur as weaving traffic movements are forced to cross one another and merge into nonweaving traffic streams. These intense lane-changing maneuvers often result in operational problems within the weaving area.

RESEARCH SCOPE

An improved analytical technique for analyzing and designing freeway weaving areas has been developed. Major weaving sections were emphasized. Such weaving configurations are formed where at least three of the weaving areas' entrance and exit legs have two or more lanes. (Figure 1 shows one possible geometric configuration for a major weaving section.) This type of weaving configuration commonly occurs at freeway-to-freeway interchanges.

The proposed weaving technique is presented. In calibrating the proposed procedure, a great deal of empirical (and simulation) data were collected from numerous weaving sites in California. Space constraints prohibit presentation of all data and analysis results. Thus, in the interest of brevity, empirical observations were collected from a single weaving site. These data reveal some of the more interesting traffic flow characteristics observed in the overall study; Cassidy et al. (*2*) present greater detail.

Existing Weaving Procedures

Given the prevalence of weaving sections on freeways, and the turbulence problems often associated with weaving operation, the need exists for analytical techniques that reliably model weaving performance. A number of existing analytical procedures are available for analyzing and designing freeway weaving areas (*3–8*). These procedures typically attempt to predict the average travel speed of vehicles operating on the subject weaving section as a function of certain prevailing traffic flow and geometric conditions. However, data collected and analyzed by the Institute of Transportation Studies (ITS) at the University of California (*9–12*) and the California Department of Transportation (Caltrans) (*13*) indicate that the existing procedures cannot reliably predict operation observed on weaving areas in California. Much of these findings have been outlined by Cassidy et al. (*9*). There are several

M. J. Cassidy, Department of Civil Engineering, Purdue University, West Lafayette, Ind. 47907. A. D. May, Department of Civil Engineering, University of California, 109 McLaughlin Hall, Berkeley, Calif. 94720.

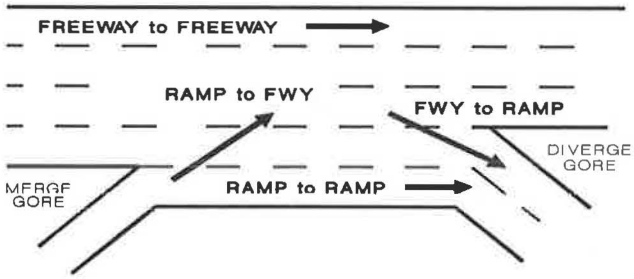


FIGURE 1 Weaving area traffic movements.

possible explanations for the deficiencies associated with the existing analytical techniques.

The procedures use aggregate traffic volumes to predict performance. That is to say, information reflecting total flow conditions operating on the subject weaving section is used to predict average travel speeds. No attempts are made to model the lane use distributions of vehicles or the lane-changing patterns of traffic within the subject weaving area.

However, the operation of freeway weaving sections may be largely influenced by what is occurring in individual lanes. Thus, operational performance can perhaps best be modeled on a lane-by-lane basis. The distribution of vehicles across available lanes and the conflicts created by weaving vehicles should be considered. Only in this way can the complex interactions occurring between weaving and nonweaving traffic streams be reliably modeled.

Input requirements for the existing procedures generally include variables that represent basic geometric design conditions of a subject weaving area. Given the influence of geometrics on traffic flow behavior, weaving procedures should perhaps consider geometrics more explicitly.

Perhaps the most significant problem associated with the existing procedures is the measure of effectiveness used. It appears that average vehicle travel speed does not reliably reflect operational quality occurring in a weaving section. Figure 2 is a scatterplot of average traffic flow per lane (*V/N*) versus average composite vehicle travel speed. Note that this composite speed represents a weighted average observed among weaving and nonweaving traffic movements. Data points shown in Figure 2 represent 5-min empirical observations collected from eight major freeway weaving sites in California (*14*).

Referring to Figure 2, speed appears to be rather insensitive to flow, up to *V/N* values of about 1,600 passenger cars per hour (pcph). Thus, average speed does not decrease with decreasing level of service (LOS) (i.e., with operational quality). Moreover, the high observed variance exhibited in Figure 2 suggests that speed is not easily predicted given aggregate flow information.

Research Approach

As a result of research efforts at ITS, a proposed procedure has been developed for analyzing and designing major freeway weaving sections. In an effort to remedy potential weaknesses in the existing weaving techniques, the proposed procedure consists of a different analytical framework. Rather than predicting the average speeds of vehicles traveling in the weaving area, the proposed technique predicts the distribution of vehicles at any location within the right-most lanes of a major weave. In other words, the proportions of each traffic movement occupying the right-hand lanes of the weaving area are estimated and used to measure operational performance. In this way, the proposed technique evaluates performance by examining traffic flow behavior in individual lanes of a subject weaving section.

By predicting the number of vehicles in individual lanes, the proposed approach models lane use. Moreover, changes in vehicle distributions (within individual lanes) over some interval of length reflect lane-changing activity. Both lane utilization and lane-changing activity can be used to reflect capacity of a given weaving area. One of the primary objectives behind this work was to identify weaving area capacity and to develop a technique to determine under which geometric and traffic flow conditions capacity is exceeded.

Lane utilization and lane changing activity also reflect motorists' perceptions concerning service quality. The proposed measure of effectiveness used in assessing weaving area level of service is discussed in a later section.

The analytical approach presented was originally developed by Moskowitz (*15*). The procedure is included in the 1965 HCM (*7*). However, the Moskowitz procedure was developed using data observed on so-called ramp weave sections. Such weaving areas are formed where a one-lane on-ramp is followed by a one-lane off-ramp and the two are joined by a

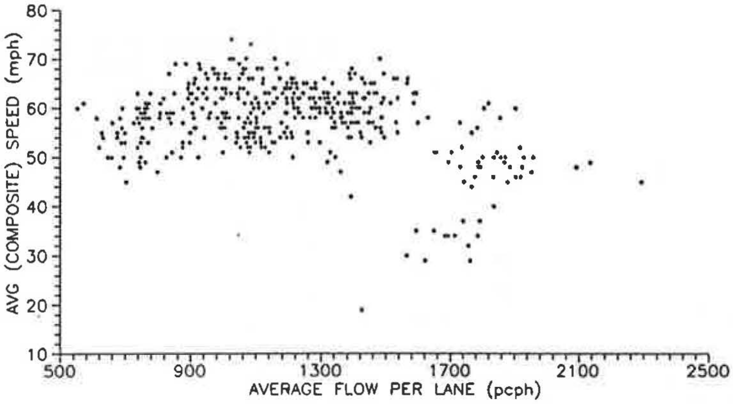


FIGURE 2 Average (composite) travel speed versus average flow per lane.

continuous auxiliary lane. (The geometric configuration of such a weaving section does not vary.)

Moreover, the Moskowitz procedure does not predict operation as a function of traffic flow conditions. The procedure was originally developed to evaluate high-flow (i.e., near-capacity) conditions. The procedure does not account for the effects of varied contributions from the four traffic movements (as shown in Figure 1).

In contrast, the proposed technique was developed to evaluate weaving operation as a function of geometric design and varying traffic flow conditions. The effects of geometric design and traffic flows are further examined in the following section.

EMPIRICAL OBSERVATIONS

To develop the proposed procedure, more than 30 hr of empirical data were collected from 9 major weaving sites in California and analyzed in considerable detail. In the interest of brevity, empirical observations from only one weaving location are presented. Data from this weaving site reflect some of the more interesting operational patterns observed on all weaving test beds.

The Weaving Test Site

Operational data presented were collected on the westbound San Bernardino Freeway, just east of Los Angeles, California. The weaving section itself is formed by a one-lane on-ramp (Garvey Avenue) followed by a two-lane off-ramp (Freeway 605 Interchange). The distance between the painted merge and diverge gore points (i.e., the ends of the painted merge and diverge areas) is 1,460 ft. The weaving site is shown in Figure 3.

Data Collection

In this research, operational data were collected using video recording. Videotaping traffic provides a permanent record of the data that can later be analyzed at various levels of detail or rechecked as necessary. A Panasonic camera model WV-3250, with a 12:1 lens, was mounted on a tripod and stationed at overcrossings immediately upstream or downstream of the subject weaving areas. The camera was positioned such that the operation of the entire weaving section could be videotaped. Once captured on video, appropriate operational data were extracted directly from the video tapes. Required data were processed through repeated viewing on a large monitor.

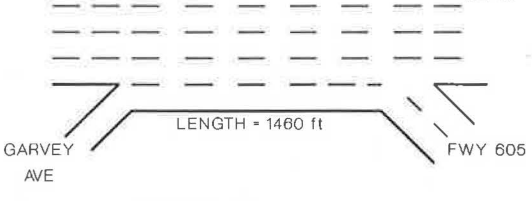


FIGURE 3 Weaving site.

All empirical data were collected by 5-min totals. Just under 5 hr of operational data (i.e., fifty-nine 5-min data points) were collected from the weaving site. These observations were collected during a.m. and p.m. peak periods and the noon off-peak period.

As previously stated, the proposed technique predicts traffic flow behavior in the rightmost portion of a subject weaving area. It is this portion of the weaving section that typically experiences the greatest turbulence because of merging and diverging traffic streams. For this reason, spatial distribution data were collected only in the weaving site's auxiliary lane (Lane 1) and the rightmost freeway lane (Lane 2).

In collecting spatial distribution data, the weaving site was divided into reference points located at fixed intervals within the weaving section's right-hand lanes. Flows from each of the four traffic movements were measured at these points. The first reference points are located 0 ft downstream of the painted merge gore point (i.e., the entrance of the weaving section). The second and third sets of reference points are located 250 and 500 ft downstream of the painted merge point, respectively. Remaining reference points are sequentially located at 500-ft intervals until the end of the weaving area (i.e., the painted diverge gore point) is reached.

Data Analysis

Freeway-to-Ramp Traffic

In an effort to evaluate the traffic flow behavior of freeway-to-ramp vehicles (in individual lanes) as a function of the location within the weaving section, scatterplots were constructed illustrating the observed percent spatial distributions of freeway-to-ramp vehicles versus the location within the weave. Figure 4 shows these spatial distributions in Lane 1 for the a.m., noon, and p.m. data. Each data point represents a 5-min empirical observation.

Figure 4 shows that these spatial distributions follow a logical pattern. For all time periods, roughly 5 percent of the freeway-to-ramp traffic travel in Lane 1 at a location 0 ft

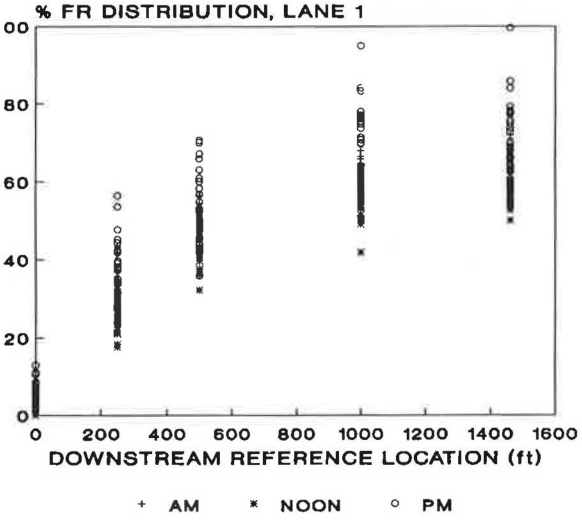


FIGURE 4 Freeway-to-ramp distributions, Lane 1.

downstream of the painted merge gore point. These vehicles perform somewhat premature lane changing by actually crossing into the painted merge gore area on executing their maneuver.

At a point 250 ft downstream of the merge gore, approximately 25 to 35 percent of the freeway-to-ramp vehicles are traveling in Lane 1. At this 250-ft location, the remaining freeway-to-ramp vehicles are traveling in Lane 2 or in median lanes.

At a point 500 ft downstream of the merge gore, about 45 to 55 percent of freeway-to-ramp traffic is traveling in Lane 1. The proportion of freeway-to-ramp vehicles traveling in Lane 1 continues to increase (slightly) at further downstream locations. Figure 4 shows how freeway-to-ramp vehicles diverge to the right as they move downstream within the weaving area.

Figure 5 shows freeway-to-ramp spatial distributions in Lane 2 as a function of the downstream location for a.m., noon, and p.m. data. The pattern appears compatible with Lane 1 flows. The proportion of freeway-to-ramp traffic in Lane 2 declines as vehicles move downstream within the weaving section.

Figures 4 and 5 indicate that the majority of lane changes made by freeway-to-ramp vehicles occur within the first 500 ft of the weaving area. Figures 4 and 5 suggest that a fair amount of scatter does exist among freeway-to-ramp observations. Much of this variance is caused by varying traffic flow conditions occurring during data collection periods.

Analysis of variance (ANOVA) tests were performed on the spatial distributions of each time period (i.e., a.m., noon, p.m.) at each reference location. At the 0.05 significance level, the null hypothesis (μa.m. = μnoon = μp.m.) could not be accepted at all but one reference location. Thus statistically significant differences exist between the distribution means of the a.m., noon, and p.m. observations.

Thus, some traffic flow parameters, or related externalities, may be contributing to differences in observed traffic flow patterns across all three time periods, as well as the variances exhibited within individual time periods. The traffic flow parameters that influence the operation of major weaving areas can therefore be determined by identifying the parameters causing scatter among the observed traffic flow patterns.

The traffic flow parameter that appears to be most clearly influencing the behavior of freeway-to-ramp vehicles on the weaving site is weaving flow rate (i.e., the sum of the freeway-to-ramp and ramp-to-freeway flow rates). Figure 6 is a scatterplot of the percent of freeway-to-ramp traffic traveling in Lane 2, at 250 ft downstream of the painted merge gore versus weaving flow. The figure indicates a downward-sloping relationship between these two variables. It suggests that as weaving flows increase, freeway-to-ramp motorists become more anxious to change lanes over shorter traveled distances (i.e., lane changes occur closer to the merge gore). Such a relationship may logically reflect motorist desires. Where weaving intensity increases, motorists may sense greater pressure in performing driving tasks. This increased feeling of pressure may encourage motorists to perform lane-change maneuvers as soon as possible. On the subject weaving site, ramp-to-freeway and ramp-to-ramp traffic operate at relatively low flow levels. Thus, these movements exhibit a preponderance of available gaps. Freeway-to-ramp motorists, wishing to perform required maneuvers as soon as possible, accept the first suitable gaps in conflicting traffic streams and are therefore able to change lanes over shorter traveled distances.

The ability of freeway-to-ramp motorists to perform lane changes in shorter distances under observed high-weaving flows is linked to the fact that freeway-to-ramp vehicles represent the predominant weaving movement on the weaving site. It was observed that on other weaving sites, where freeway-to-ramp vehicles comprise the smaller weaving movement, freeway-to-ramp motorists require greater distances to perform desired lane changes as weaving flow rates increase. Overall findings suggest that lane-changing characteristics are a function of gap availability in conflicting traffic streams.

As shown in Figure 6, the proportions of freeway-to-ramp traffic traveling in Lane 2, at 250 ft, were divided into three partitions:

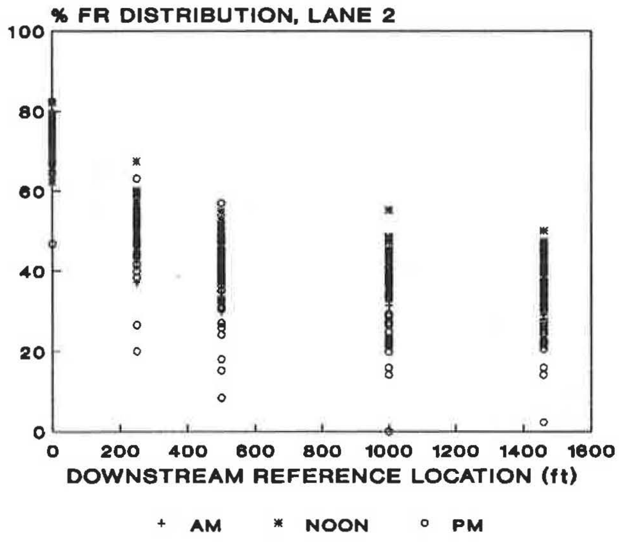


FIGURE 5 Freeway-to-ramp distributions, Lane 2.

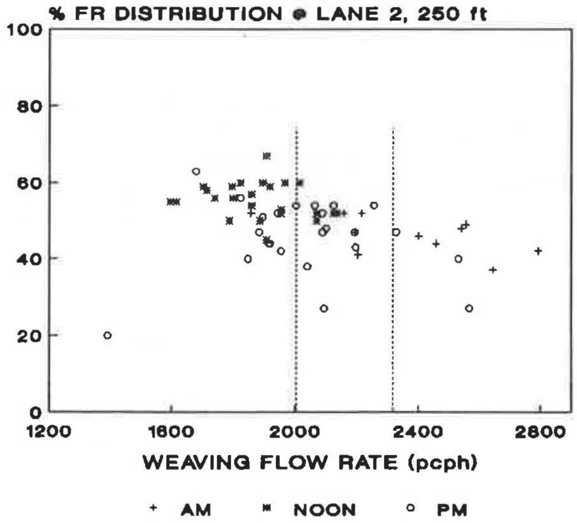


FIGURE 6 Freeway-to-ramp distributions, Lane 2, at 250 ft, versus weaving flow rate.

1. Observations where total weaving flow rates are less than 2,000 passenger cars per hour (pcph).
2. Observations where total weaving flow rates are between 2,000 and 2,300 pcph.
3. Observations where total weaving flow rates exceed 2,300 pcph.

ANOVA and difference of means (DOM) tests indicate that a.m., noon, and p.m. observations within these three partitions do not have statistically significant differences in their distribution means. Thus, three flow regimes have been identified. By partitioning the data, observations from all three time periods can be directly combined (within partitions). Similar findings occurred at other reference points within the subject weaving site.

Ramp to Freeway Traffic

Figures 7 and 8 show the spatial distributions of ramp-to-freeway traffic traveling in Lanes 1 and 2 for a.m., noon, and p.m. observations. Figure 7 shows that no ramp-to-freeway vehicles remain in Lane 1 at the downstream end of the weaving site (i.e., at 1,460 ft). This is logical given that Lane 1 of the subject weaving site is a mandatory off-ramp lane. Ramp-to-freeway vehicles must exit Lane 1 within the length of the weaving section.

Figure 8 shows that less than 100 percent of the ramp-to-freeway vehicles typically travel in Lane 2 at the downstream end of the weaving section. This merely illustrates that a portion of the ramp-to-freeway vehicles have merged from Lane 2 into the median lanes within the length of the weaving area.

The data shown in Figures 7 and 8 exhibit a high degree of scatter. These higher-observed variances are caused by two factors:

1. Ramp-to-freeway vehicles comprise the smaller weaving movement on the subject site.
2. Ramp-to-freeway vehicles are forced to change lanes using available gaps in conflicting traffic streams (i.e., freeway-to-ramp and freeway-to-freeway movements) that operate under high-flow conditions.

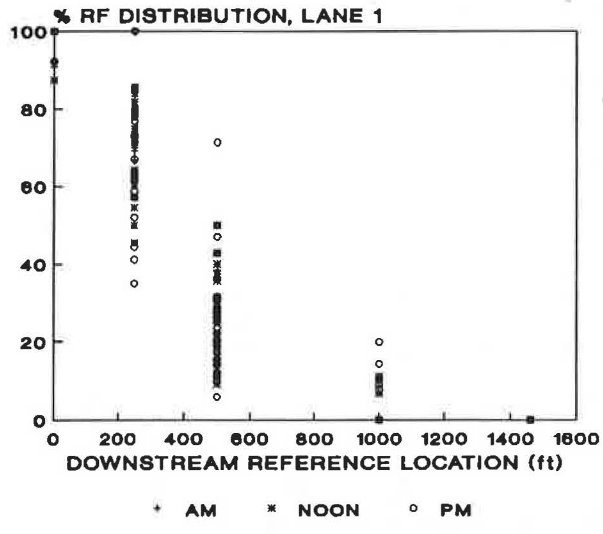


FIGURE 7 Ramp-to-freeway distributions, Lane 1.

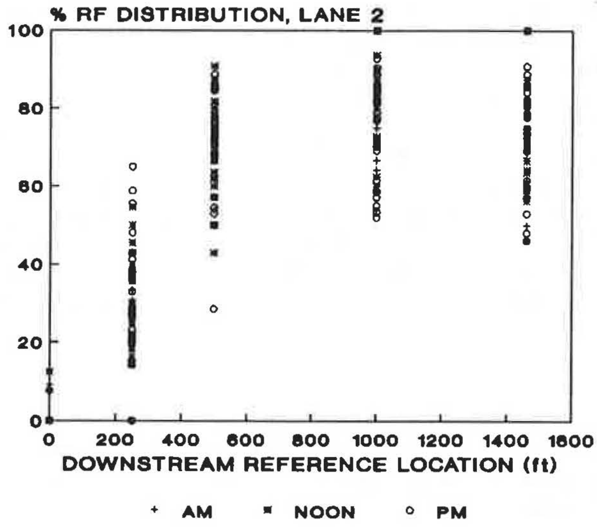


FIGURE 8 Ramp-to-freeway distributions, Lane 2.

The explanation of how these two factors contribute to higher observed variances is fairly straightforward. The lane-changing behavior of individual vehicles can be approximated as a binomial process. For example, a ramp-to-freeway vehicle traveling in Lane 1 of a weaving section will either (a) cross over a reference point located downstream in Lane 1, or (b) avoid crossing over the reference point by changing lanes. This process can be represented by the indicator function:

$$x = \begin{cases} 1 \text{ if vehicle crosses reference point} \\ 0 \text{ if vehicle does not cross reference point} \end{cases}$$

Let x_i represent a single binomial event. The number of vehicles (in a particular traffic movement) actually crossing a given reference point, Σ_n, can be represented as

$$\Sigma_n = x_1 + x_2 + \ldots x_n$$

Represented as a proportion,

$$\Sigma_n/n = (x_1 + x_2 + \ldots x_n)/n$$

Because each binomial event x_i can be approximated as an independent identically distributed random variable,

$$\text{VAR } (\Sigma_n/n) = (1/n^2) [n * \text{VAR } (x)] = (1/n) \text{ VAR } (x)$$

Each sample mean shown in Figures 7 and 8 can be thought of as independently identically distributed random variables. Therefore, the observed variance among sample means is

$$\text{VAR (5-min observations)} = N * \text{VAR } (\Sigma_n/n)$$

where N is the number of 5-min observations.

Thus, the variance among the 5-min data points is inversely proportional to n, the number of vehicles observed during the individual 5-min periods. Ramp-to-freeway flow is significantly lower than freeway-to-ramp flow on the subject weaving site. This partially explains the higher variances exhibited in Figures 7 and 8.

The scatter among ramp-to-freeway observations on the subject site is also influenced by conflicting traffic movements. The term VAR (x) represents the variations among individual lane-changing movements. These variations are clearly a function of the gaps available to individual lane-changing vehicles.

P is defined as the probability that a lane-changing vehicle accepts a gap in the conflicting traffic streams. Recalling that lane-changing behavior can be approximated as a binomial process,

$$\text{VAR}\ (x) = Pq$$

where q is defined as the probability that a lane-changing vehicle will not accept a gap in conflicting traffic streams.

Therefore,

$$\text{VAR}\ (x) = P(1 - P)$$

Thus, VAR (x) is minimal where conflicting traffic flow rates are low ($P \rightarrow 1.0$) and where conflicting flow rates are high ($\text{P} \rightarrow 0$).

The influence of sample size and conflicting flow rates is exhibited by the freeway-to-ramp vehicles operating on the subject site. Freeway-to-ramp traffic represents the higher weaving flow traveling on the subject site. Thus, the sample size of freeway-to-ramp data is relatively large. Moreover, freeway-to-ramp vehicles perform lane changes in conflicting traffic streams of relatively low flow. The result is less variance among freeway-to-ramp observations when compared with ramp-to-freeway data.

ANOVA tests indicate that statistically significant differences do not exist between mean values of the ramp-to-freeway spatial distributions observed during the a.m., noon, and p.m. periods.

Freeway-to-Freeway Traffic

Figure 9 shows the proportional distributions of freeway-to-freeway traffic traveling in Lane 2 for a.m., noon, and p.m. observations. Freeway-to-freeway vehicles will not travel in Lane 1 as it is a mandatory exit lane.

No observed traffic flow variables appear to influence the distributions of freeway-to-freeway vehicles in Lane 2.

Ramp-to-Ramp Traffic

Figures 10 and 11 show the proportions of ramp-to-ramp traffic traveling in Lanes 1 and 2 for the a.m., noon, and p.m. data. Some lane changes are executed among ramp-to-ramp traffic, despite the fact that lane changing is not required to perform the maneuver.

Similarly to freeway-to-freeway traffic, ramp-to-ramp distributions do not appear to be influenced by variations in observed traffic flow conditions.

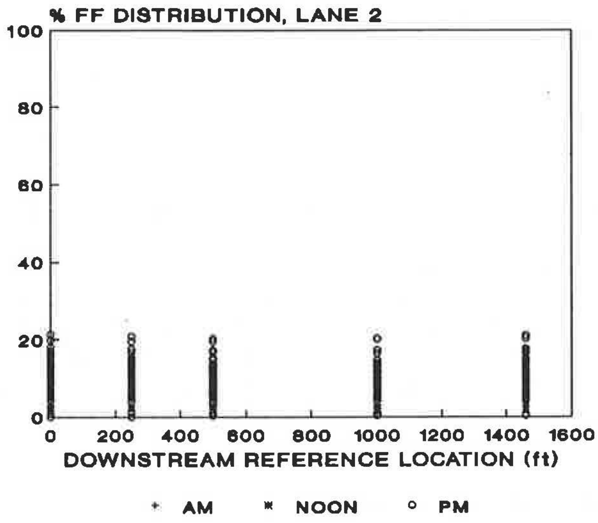


FIGURE 9 Freeway-to-freeway distributions, Lane 2.

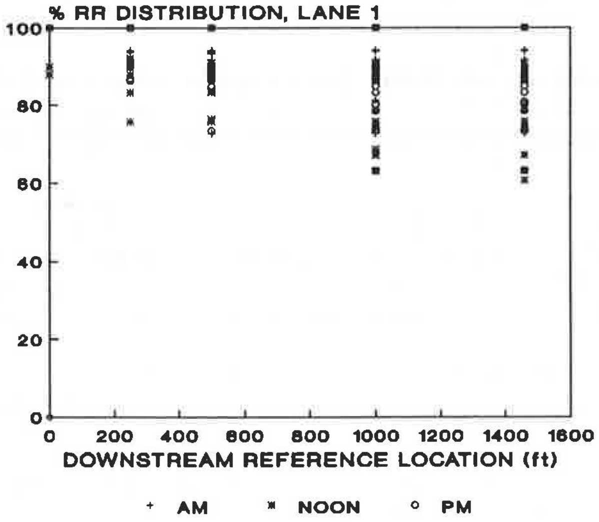


FIGURE 10 Ramp-to-ramp distributions, Lane 1.

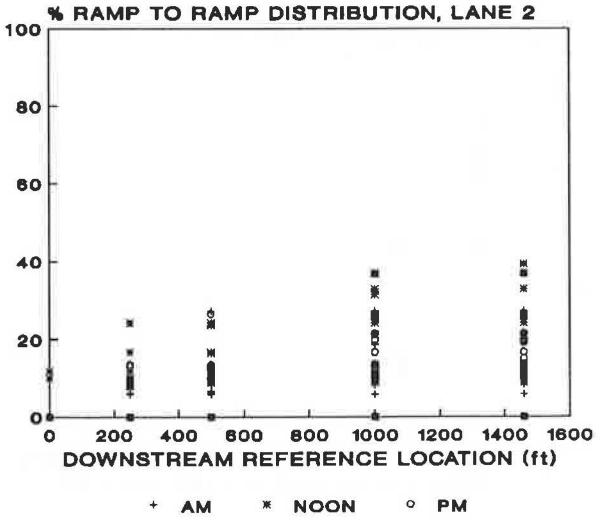


FIGURE 11 Ramp-to-ramp distributions, Lane 2.

PROPOSED WEAVING PROCEDURE

The ultimate objective of this research has been to develop an improved analytical technique for modeling traffic flow characteristics in major freeway weaving areas. The proposed procedure predicts the spatial distributions of all traffic movements within a subject weaving area as a function of the prevailing geometric design and traffic flow characteristics. Specifically, the procedure computes flow rates at user-specified reference points within the weave. Lane-changing activity is reflected in changes in flow rates of each traffic movement between sequential reference points. Flow rate and lane-changing information is then used for designing or evaluating performance of a weaving section.

The proposed technique is graphical in form. The procedure consists of a family of curves that are used to predict the spatial distributions of each traffic movement. The curves, which were developed both from empirical and from simulation data, reflect a wide range of traffic flow and geometric conditions. Curves vary in shape on the basis of factors such as the weaving area's geometric configuration, section length, and traffic flow rates. The curves represent interpolated data between mean values for sets of specific observations. The cubic spline function was used for constructing the curves.

Space constraints again dictate that only those curves calibrated from observations occurring on the single selected weaving site (shown in Figure 3) are presented. These curves are presented in Figures 12–17. Although applicable to only one specific geometric configuration, these curves show procedural format. Other sets of curves are available for other weaving configurations (*2*).

Procedural Steps

In performing the proposed procedure, the user first selects the curves appropriate for the weaving area geometric configuration. Separate graphs correspond to individual traffic movements. In many cases, appropriate curves are also selected on the basis of relevant traffic flow conditions.

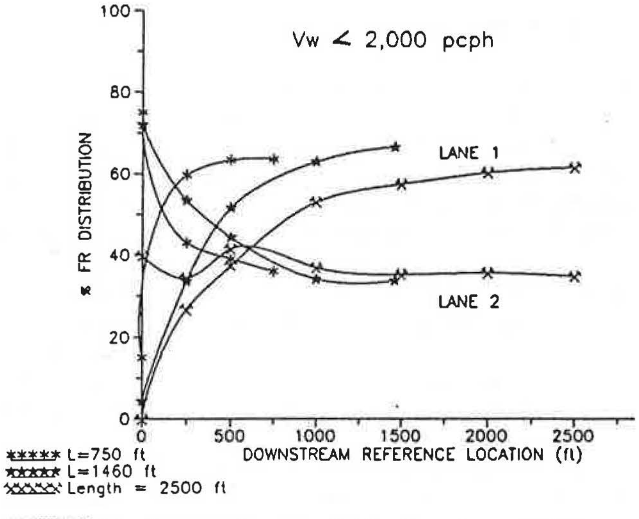


FIGURE 12 Freeway-to-ramp distributions, low weaving flow rates.

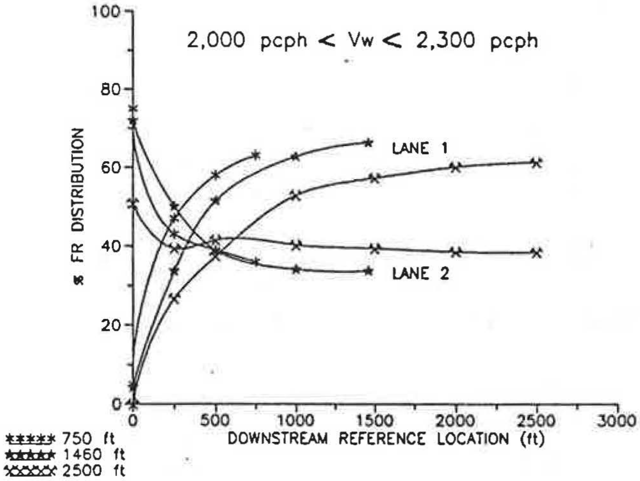


FIGURE 13 Freeway-to-ramp distributions, moderate weaving flow rates.

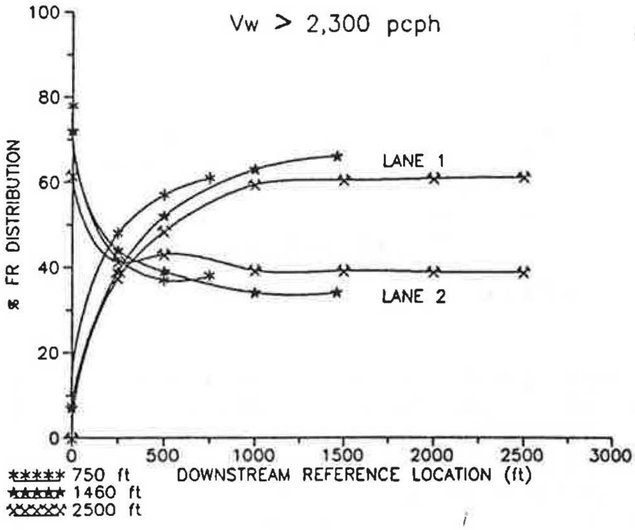


FIGURE 14 Freeway-to-ramp distributions, heavy weaving flow rates.

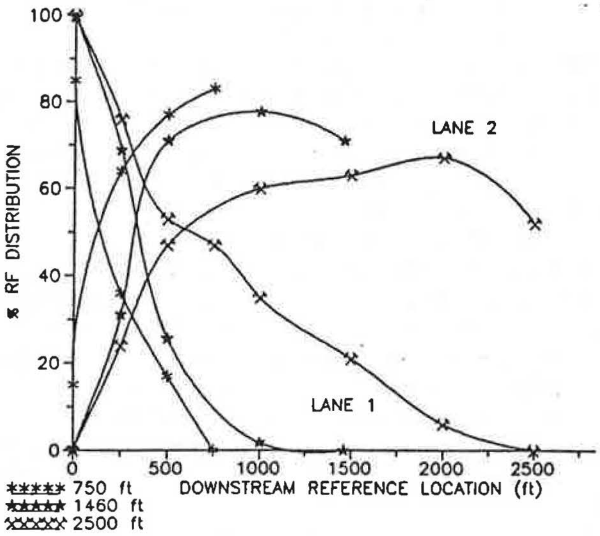


FIGURE 15 Ramp-to-freeway distributions.

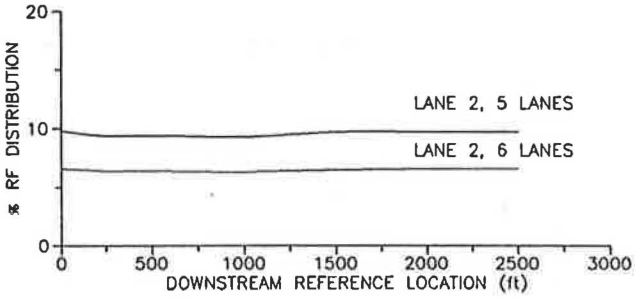


FIGURE 16 Freeway-to-freeway distributions.

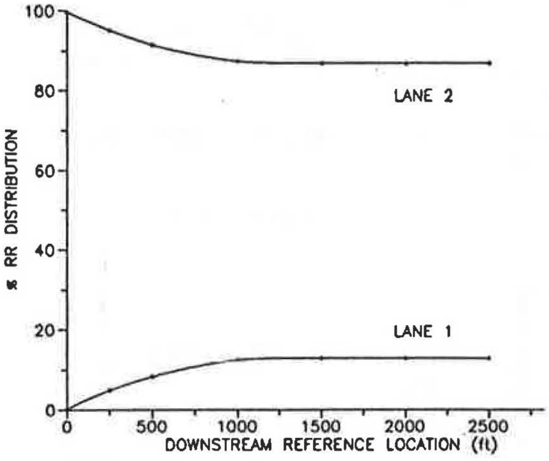


FIGURE 17 Ramp-to-ramp distributions.

Once applicable curves are selected, the user determines the reference points within the weaving area where operational performance is to be evaluated. A conservative approach would be to examine points across the entire length of a weaving section to ensure that all critical points are evaluated. Typical distance intervals between selected reference points would range from 250 to 500 ft. Reference points are located on the horizontal axis of each graph.

The user projects vertically upward from the horizontal axis until intersecting the appropriate curve on the graph. Individual curves typically represent predicted spatial distributions occurring in specific lanes on a weaving section of specific length. The user should refer to the legend in the lower left corner of the graph to determine lengths represented by each curve. Interpolation will generally be required for selecting appropriate curves.

From the intersection of the user's vertical line and the appropriate curve, the analyst projects horizontally to the left and determines the proportion of the specific movement traveling at the selected reference location. This predicted proportion is converted to a flow rate by simply multiplying by the total flow rate of the particular movement.

These tasks should be repeated for each of the four traffic movements. In this way, total flows are used in assessing LOS and in evaluating capacity sufficiency of the subject weaving area.

In Figures 12–15, the curves reflecting traffic flow behavior on 1,460-ft weaving sections were constructed using empirical data extracted from the subject weaving test site (Figure 3). Curves reflecting operation on 750- and 2,500-ft sites were constructed using calibrated simulations. Curves depicted in Figures 16 and 17 were developed using both empirical and simulation data.

Limitations of Proposed Model

The proposed model can be used for the analysis or design of simple, major freeway weaving areas. In this research, every effort was made to develop a procedure that is applicable to a wide range of weaving scenarios. Empirical and simulation data used to calibrate the proposed model reflect a wide range of traffic flow conditions. In extreme cases, however, traffic flow conditions at a subject weaving site may fall beyond the range of the calibration data set. In such cases, the reliability of the proposed model may deteriorate.

The model can be applied to a number of different weaving area geometric designs. However, the curves incorporated in the proposed model do not reflect all possible weaving site geometries. Where a weaving design not accommodated by the model is to be evaluated, traffic flow patterns may often be approximated using the model's curves that most closely match this actual weaving design.

The calibration test sites used in this research do not vary significantly in their number of lanes. Thus, the proposed model is most applicable for major weaving sections having five or six lanes within the weaving area.

This research has adopted an important simplifying assumption. Where on- and off-ramps have two lanes, each of these ramp lanes often leads to, or extends from, different directional flows on connecting facilities. Thus, the origin of on-ramp vehicles and the destination of off-ramp vehicles may, to some extent, influence presegregation and lane-changing within the weaving section. However, such considerations are beyond the scope of the proposed research. (Data collection alone precludes such considerations.) Therefore, it is assumed that lane changing influenced by such origin and destination considerations occurs on the ramp connection facility and not within the weaving section itself.

Weaving Area Capacity

A significant problem associated with weaving sections is that of determining capacity. Typically, capacity is thought of as the maximum flow of vehicles that can reasonably be expected to travel through a given facility. However, with weaving sections, the maximum number of vehicles that can traverse the facility will vary with different contributions from the four traffic movements. For example, lane-changing activity resulting from high weaving flow rates will restrict throughput capabilities within a given weaving section. Conversely, where weaving flow rates are low (i.e., little lane changing exists), weaving capacity will be relatively high. For this reason, the existing speed-prediction procedures do not directly predict weaving area capacity.

The proposed procedure defines capacity somewhat microscopically. Capacity is defined as

1. The maximum flow of vehicles that can travel at any point (within a lane) of roadway within a subject weaving area, and

2. The maximum rate of lane changing (between two adjacent lanes) that can occur over any 250-ft lane segment (i.e., lane stripe) within the weaving area.

These capacity values were determined largely through simulation. In this research, an updated version of the INTRAS (INtegrated TRAffic Simulation) microscopic simulation model (*16–18*) was used to augment empirical data. Weaving area capacity was determined by performing repeated simulations of high traffic flow conditions. Capacity threshold values were assumed to have been exceeded when simulated conditions became so congested that a small number of weaving vehicles (i.e., 1 or 2 percent) were unable to perform desired movements.

By successfully calibrating the INTRAS model so that it reliably replicated observed conditions (including near-capacity conditions), it is believed that capacity values derived from simulation are reliable. Using simulation to extrapolate observed flow conditions simply represented the best means available for determining capacity.

The results of extensive simulation modeling indicate that capacity flow values are 2,200 pcph at any point within a weaving section. Lane-changing capacity is found to range from 1,100 to 1,200 lane changes per hour (across a single lane-line) over any 250-ft segment within the weaving area. As these capacity values represent microscopic measures, they are consistent regardless of geometric and traffic flow conditions prevailing at the weaving site.

Where application of the proposed weaving technique yields flow and lane changing values that exceed the capacity thresholds, operational breakdown is expected and the user should consider geometric modifications.

Weaving Area LOS Criteria

Another aspect of the research has been to determine an appropriate measure and criteria by which the LOS of a given weaving section can be assessed. LOS values typically reflect one of two possible perspectives:

1. The viewpoint of the system operator—in which case LOS may be expressed in measures such as productivity or volume-capacity ratio.
2. The viewpoint of the highway user—in which case LOS may be expressed in measures such as average travel speed, delay, or density.

In an effort to stay consistent with the basic philosophy behind the HCM, the measure of effectiveness for this research must be a parameter directly perceivable by users (i.e., motorists and passengers). Therefore, flow or volume-to-capacity ratio would not be suitable measures of effectiveness. Although perceivable by the user, average travel speed would also appear to be an inappropriate measure, largely because of its insensitivity to changing traffic volumes (Figure 2).

Two remaining measures perceivable by the user are

1. The rate of lane-changing activity over a given segment within a weaving area, and
2. The average density within a small segment of a weaving section.

The level of lane-changing activity does characterize the overall operational performance of a weaving area. The negative aspect of such a measure is that the number of lane changes cannot be easily field-measured. The analyst trying to quantify performance by field observation would be charged with a difficult task.

Like lane-changing activity, density might initially seem to be a difficult parameter to field-measure. Figure 18 shows that density behaves predictably relative to flow conditions. Data points shown in Figure 18 represent over 400 5-min observations collected from eight major freeway weaving areas located in California. These data, which reflect uncongested flow conditions, were collected as part of the initial ITS weaving research (*5*).

The relatively low scatter among the observations in Figure 18 suggests that density can be reliably estimated given average flow information. Thus, to field-measure density, the analyst need only count observed flow rates at a single point.

The x-axis in Figure 18 reflects average flow per lane. However, one can assume that using the curve depicted in Figure 18 to estimate density on the basis of a (predicted) point flow in a single lane represents a reasonable approximation. Thus, the curve can be used to predict density as a function of point flow rate. The resulting best-fit curve is a polynomial with three degrees of freedom (also shown in Figure 18).

The equation for this curve is

$$D = 0.42 + 0.02(V/N) - 1.19 * 10^{-5}(V/N)^2 + 5.36 * 10^{-9}(V/N)^3$$

where

D = density, passenger cars per mile per lane; and
V/N = flow per lane, pcph.

Data used to develop Figure 18 indicate that, under high-flow conditions (i.e., $V/N > 1{,}600$ pcph), average travel speed for uncongested conditions is approximately 48 mph. Data appearing to be in the congested regime (i.e., speeds below 35 mph) were not included in determining average speed under high-flow conditions. As an approximation, it was assumed that travel speed takes on a constant speed over the range of high-flow conditions (i.e., $V/N > 1{,}600$ pcph).

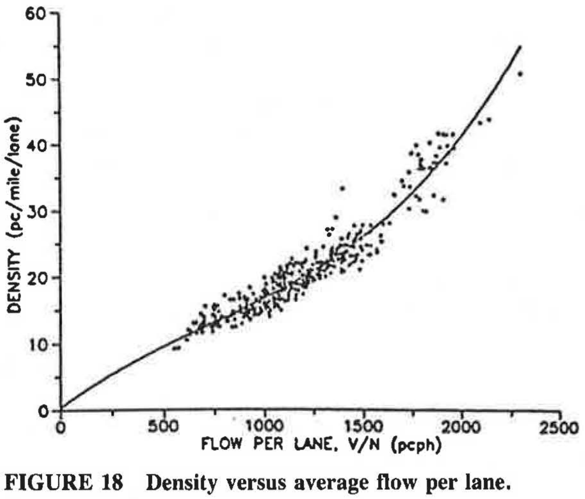


FIGURE 18 Density versus average flow per lane.

Because capacity flow conditions have been determined to be 2,200 pcph, the value of density at capacity is roughly 46 passenger cars, per mile, per lane (2,200 pcph divided by 48 mph). This value of optimum density corresponds to data shown in Figure 18.

m The 1985 HCM defines optimum density for basic freeway segments to be 67 passenger cars, per mile, per lane. It is expected that optimum densities for weaving areas would be less than those observed on straight-pipe freeway sections. Turbulence created by weaving traffic streams reduce optimum densities. Moreover, the relatively high freeway density values documented in the HCM are likely the result of overly conservative estimates of average travel speeds.

Once an optimum density value was determined, boundary conditions for the remaining LOS designations were divided evenly.

In part because of the simplicity associated with the proposed density estimating technique, average density is the suggested measure of effectiveness for assessing weaving area performance.

The proposed procedure does not automatically suggest that the lowest LOS governs for the entire weaving section. Such an approach may often lead to overly conservative LOS predictions. The proposed technique evaluates operation within relatively small lane segments of the weaving area. Traffic flows (and service qualities) vary from one segment to the next. Thus, application of the proposed technique will typically result in one or two reference points operating at the lowest relative LOS (i.e., highest relative densities) with surrounding reference points often operating at LOS values that are one or two designations higher. In estimating the overall LOS values for a given weaving area, the user is encouraged to evaluate the LOS values occurring in lane segments throughout the weaving section and determine if overall operation is acceptable. The overall LOS becomes a somewhat subjective average of the lowest LOS values occurring in a region of a weaving section. User judgment is required.

Space constraints prohibit the inclusion of example applications of the proposed technique. Example applications were provided by Cassidy et al. (*2*).

Model Validation

As part of this research, a small amount of data was collected and reserved exclusively for model validation. Thirty minutes of high-flow validation data were collected from the subject weaving site presented. These validation data represent flow conditions exceeding most or all of the calibration data. (All validation data points reflect weaving flow rates greater than 2,300 pcph.)

Figures 19 and 20 show validation results among freeway-to-ramp data. The solid curves in Figures 19 and 20 depict freeway-to-ramp spatial distributions (in Lanes 1 and 2) predicted by the proposed technique. The data points in Figures 19 and 20 represent empirical observations.

Figures 19 and 20 suggest that the proposed model predicts freeway-to-ramp flow patterns reasonably well. The same general results were observed among the remaining three vehicle movements.

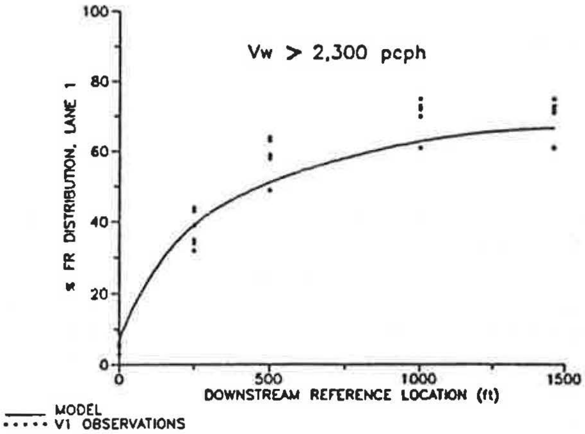


FIGURE 19 Model validation—freeway-to-ramp distributions, Lane 1.

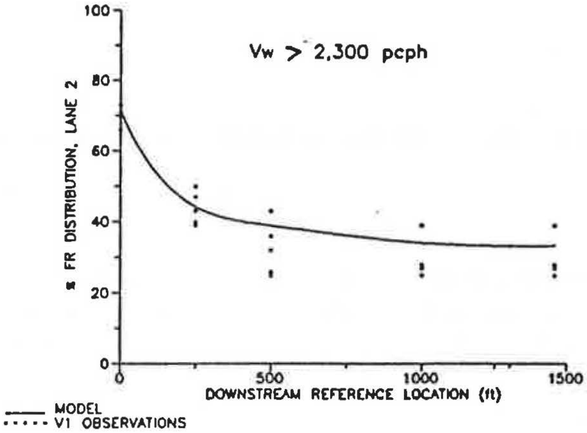


FIGURE 20 Model validation—freeway-to-ramp distributions, Lane 2.

CONCLUSIONS

An analytical technique has been developed for the design and analysis of simple, major freeway weaving sections. The research project examined traffic operation on a number of different weaving sites in California. Observations occurring on one of the weaving test locations were summarized.

Initial research (*12–14*) suggested that more reliable weaving area analysis and design procedures were needed. Currently available techniques for predicting weaving performance were found to be unreliable. The analytical technique detailed in this report is used to predict the spatial distributions of individual traffic movements at user-specified points within the weaving area. Such a technique implicitly estimates vehicle presegregation and lane-changing activity within the weave. In this way, the proposed model considers the interactions of vehicles traveling in individual lanes and operation is thereby evaluated on a more microscopic level.

Data analyses performed in this research suggest that evaluating operation in individual lanes represents a preferred approach for modeling weaving performance. Knowing key traffic flow and geometric conditions, spatial distributions of individual vehicle movements can be predicted reasonably well. Lane utilization and lane-changing activity can be used as a measure of weaving area capacity. Moreover, these measures reflect performance of weaving areas. Thus, the proposed technique enables engineers and planners to more reliably design and analyze weaving sections.

Finally, the proposed model represents a rational and easily understood evaluation method. In simplest terms, the procedure estimates changing lane volumes as vehicles travel downstream within the weaving section. Such changing traffic flow patterns should appear intuitive to the user. In contrast, many of the (regression-based) speed prediction techniques have received significant criticism from the user community for not appearing rational.

ACKNOWLEDGMENTS

This research was sponsored by the California Department of Transportation and FHWA. Many professionals have contributed to the success of this research. The project staff wish to thank Caltrans for its involvement throughout the project. In particular, Fred Rooney, Len Newman, and Howard Fong are acknowledged for their insights and assistance.

Several researchers at the Institute of Transportation Studies contributed significantly to the project's development. Specifically, we thank Patrick Chan, Bruce Robinson, Alex Skabardonis, and Edward Chung for their work.

We also thank Steven Cohen, FHWA, for his advice and assistance in simulation modeling.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., Sept. 1985.
2. M. Cassidy, P. Chan, B. Robinson, and A. D. May. *A Proposed Analytical Technique for Designing and Analyzing Major Freeway Weaving Areas.* Institute of Transportation Studies, University of California, Berkeley, UCB-ITS-RR-90-Draft, 1990.
3. J. Fazio and N. Raiphail. Freeway Weaving Sections: Comparison of Refinement of Design Analysis Procedures. Presented at the 65th Annual TRB Meeting, Washington, D.C., July 1987.
4. J. E. Leisch et al. *Procedure for Analysis and Design of Weaving Sections, User Guide,* Vol. 2, Report FHWA/RD-85/083, FHWA, U.S. Department of Transportation, Oct. 1985.
5. W. Reilly, H. Kell, and P. Johnson. *Weaving Analysis Procedures for the New Highway Capacity Manual.* JHK and Associates, Aug. 1984.
6. L. Pignataro et al. *NCHRP Report 195: Weaving Areas—Design and Analysis.* TRB, National Research Council, Washington, D.C., 1975.
7. *Special Report 87: Highway Capacity Manual.* HRB, National Research Council, Washington, D.C., 1965.
8. J. Hess. Traffic Operations in Urban Weaving Areas. 1963 BPR Weaving Area Study Data Base. Bureau of Public Roads, Washington, D.C., 1963.
9. M. Cassidy, A. Skabardonis, and A. D. May. *Operation of Major Freeway Weaving Sections: Recent Empirical Evidence.* In *Transportation Research Record 1225,* TRB, National Research Council, Washington, D.C., 1989, pp. 61–72.
10. A. Skabardonis, M. Cassidy, and A. D. May. *Operation of Major Freeway Weaving Areas: Findings from the Application of Simulation Modeling.* Presented at 68th Annual TRB Meeting, Washington, D.C., Jan. 1989.
11. A. Skabardonis, M. Cassidy, and A. D. May. *Operation of Major Freeway Weaving Areas: Findings from the Application of Existing Analytical Methods and Simulation Modeling,* UCB-ITS-WB-88-12, Institute of Transportation Studies, University of California, Berkeley, 1988.
12. A. Skabardonis, M. Cassidy, and A. D. May. Evaluation of Existing Methods for the Design and Analysis of Freeway Weaving Sections. Presented at the 67th Annual TRB Meeting, Washington, D.C., Jan. 1988.
13. H. K. Fong and F. D. Rooney. *Weaving Data.* Interim Reports 1–9. State of California, Department of Transportation, 1987.
14. M. Cassidy, A. Skabardonis, and A. D. May. *Operation of Major Freeway Weaving Areas: Recent Empirical Evidence,* UCB-ITS-WP-88-11. Institute of Transportation Studies, University of California, Berkeley, 1988.
15. K. Moskowitz and L. Newman. Notes on Freeway Capacity, In *Highway Research Record 27*: HRB, National Research Council, Washington, D.C., 1963.
16. S. L. Cohen, and J. Clark. Analysis of Freeway Reconstruction Alternatives Using Traffic Simulation. Presented at 66th Annual TRB Meeting, Washington, D.C., Jan. 1987.
17. D. A. Wicks and E. B. Lieberman. *Development and Testing of INTRAS, A Microscopic Freeway Simulation Model, Vol. 1, Program Design, Parameters Calibration and Freeway Dynamics Component Development.* Final Report, FHWA/RD/-80/106. FHWA, U.S. Department of Transportation, Oct. 1980.
18. D. A. Wicks and E. B. Lieberman. *Development and Testing of INTRAS: A Microscopic Freeway Simulation Model, Vol. 2, User's Manual.* Final Report, FHWA/RD-80/107. FHWA, U.S. Department of Transportation, Oct. 1980.

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

Evaluation of High-Volume Urban Texas Freeways

THOMAS URBANIK II, WANDA HINSHAW, AND KIRK BARNES

The current value of capacity of basic freeway segments has been called into question with increasing numbers of observations of flow rates in excess of 2,000 passenger cars per hour per lane. The results of studies at seven high-volume sites in Texas are discussed. In addition, data are presented with variations in flow rates by lane. Finally, data are presented illustrating differences in flow rates before and after the onset of upstream congestion.

The issue of freeway capacity has been slowly increasing in interest as more freeway segments regularly experience volumes in excess of 2,000 vehicles per hour (vph). Recent research (*1*) on multilane highways has suggested that multilane highway capacity should be in excess of 2,000 passenger cars per hour per lane (pcphpl). High-volume basic freeway sections in Texas are addressed in the following paragraphs.

The *Highway Capacity Manual* (HCM) (*2*) gives the following general definition of capacity: "the maximum rate of flow at which persons or vehicles can reasonably be expected to traverse a point or uniform section of a lane or roadway during a specified time period under prevailing roadway traffic and control conditions." In Chapter 3 of the HCM, freeway capacity is defined as "the maximum sustained (15 min.) rate of flow at which traffic can pass a point or uniform segment of freeway under prevailing roadway and traffic conditions." Basic freeway segments are described as being sections of freeway that are unaffected by merging, diverging, or weaving movements (i.e., located 500 ft upstream or 2,500 ft downstream of an onramp, 2,500 ft upstream or 500 ft downstream of an offramp, or 500 ft from a weaving section). Furthermore, the HCM states that the capacity of a basic freeway section that is straight and level, under ideal conditions, is 2,000 pcphpl.

Difficult questions about freeway capacity still exist, including how to tell when capacity flows exist and if there actually are two capacities. The two-capacity hypothesis states that the capacity under free flow conditions is higher than the capacity discharging from a queue. Banks (*3*) provides a recent summary of the limited studies on freeway capacity over the past 40 years. He addresses the important problem of the wide range of conditions under which most data were collected. Banks presents data that are consistent with a two-capacity hypothesis; however, the difference between the two capacities was only 3 percent.

Hall and Hall (*4*) present data indicating flow rates of 2,300 pcphpl on the Queen Elizabeth's Way in Ontario, on the basis of an assumed truck–passenger car equivalent (PCE) of 2. The Hall and Hall data exhibit no drop in capacity because of congestion. However, Hall and Hall used 5-min averages, whereas Banks used 30-sec averages. Data collected in four Texas cities that demonstrate bottleneck discharge capacities in excess of 2,000 pcphpl are presented. Some limited data that appear consistent with the two-capacity hypothesis are also provided.

METHODOLOGY

The data were collected using video cameras and were reduced lane by lane using computer-assisted data reduction. This method of data collection provides for individual vehicle headways and classifications by lane. The information corresponding to each vehicle is identified by time of observation. This methodology allows summarization of vehicle headways according to the classification of the vehicle as well as that of the preceding vehicle. The availability of this type of data provides the opportunity to investigate many aspects of highway capacity and the limitations to it.

Reliability analysis of the count data suggests that the data reduction errors are less than 0.5 percent (2,400 ± 2 vehicles). The limiting factor in the video data is the lack of information on vehicle speeds. Additional equipment is being obtained to automate the data collection process; the methodology includes the calculation of speed. The ability to measure the headway and speed and to classify individual vehicles provides the necessary data base to better understand capacity flows.

RESULTS

Data were collected at seven locations in the four largest Texas cities. Most locations are downstream of bottlenecks, although some locations have intervening onramps. Data collection locations were limited by the ability to find a vantage point for videotaping. This limitation is an important one and is discussed in a subsequent section. All sites had virtually no grade.

Several types of estimators were calculated to determine capacity limitations at the various sites. Vehicle counts at 1-min intervals were computed from individual vehicle data for each lane of traffic. From these counts, flow rates based on intervals of different durations were easily computed and examined. To be consistent with existing convention, 15-min flow rates were employed for analysis. The peak 15-min flow is the highest flow observed for 15 consecutive minutes, independent of actual clock time. Total volumes for the peak hour of flow were also computed. The reported truck percentages

Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843.

reflect the truck traffic in this peak hour. No attempt was made to adjust for truck PCEs at any location. Data summaries are presented in Tables 1 and 2. Table 1 presents the hourly volume summaries, and Table 2 presents the corresponding 15-min summaries. A more detailed discussion of the data, including site characteristics, follows.

The I-35–US-67 study site is located in southern Dallas. The site is just downstream of the intersection of a three-lane radial Interstate highway and a two-lane U.S. highway, as shown in Figure 1. Although not a typical bottleneck, the study section exhibits the characteristics of typical onramp bottlenecks. Table 1 presents data from three studies at this location. Peak 15-min flow rates averaged across the four northbound lanes were 2,213, 2,298, and 2,278 vehicles per hour per lane (vphpl). Truck percentages corresponding to the peak 15-min flows were 1.2, 1.2, and 2.0 percent, respectively.

The second Dallas location, I-635 at Coit, is located on the northern circumferential Interstate highway. As shown in Figure 2, the study section is downstream of a heavy-volume a.m. peak onramp, on the westbound lanes of this eight-lane freeway. The facility routinely operates with demands in excess of capacity. The peak 15-min flow rate averaged across the three observed lanes, as presented in Table 1, was 2,249 vphpl with 1.9 percent trucks.

The Dallas North Tollway is a six-lane, radial freeway of closely controlled access that extends from near the Dallas central business district (CBD) northward 15 mi. The study site is located on the southbound side. Even though the site is not a bottleneck, flows were known to be high. The site geometry is shown in Figure 3. The observed peak-hour flow rate averaged across the three observed lanes was 2,373 vphpl with 0.7 percent trucks.

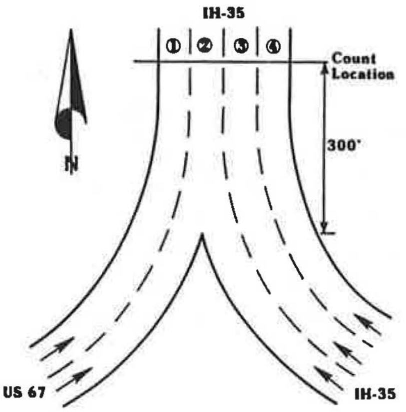


FIGURE 1 Schematic of I-35–US-67 study site.

The I-820 site is a four-lane circumferential Interstate on the east side of Ft. Worth. The study site was selected because

TABLE 1 STUDY SITE SUMMARIES AND VOLUMES

City	Highway	Number of Lanes	Observation	Peak Hour Volume Average per Lane (Vehicles/hour)	% Trucks in Peak Hour
Dallas	IH 35/US 67	4	1	2125	1.7
			2	2211	1.4
			3	2194	1.9
	ΓH 635 @ Coit	4	1	2117	2.1
	North Tollway	3	1	2026	0.5
Ft. Worth	IH 820	2	1	1887	1.7
			2	1882	3.5
	US 183	3	1	2222	2.5
			2	2308	2.4
			3	2197	2.4
Houston	US 290	3	1	2143	3.0
San Antonio	IH 410	3	1	2122	1.4

TABLE 2 PEAK 15-min FLOW RATE SUMMARIES

Highway	Observation	Peak 15-Minute Flow Rate by Lanes					% Trucks in Peak 15 Minutes
		1	2	3	4	Average	
IH 35/US 67	1	2472	2528	2084	1768	2213	1.2
	2	2392	2532	2180	2088	2298	1.1
	3	2332	2456	2164	2160	2278	2.0
IH 635 @ Coit	1	2380	2244	2220	2152	2249	1.9
North Tollway	1	2468	2344	2308	----	2373	0.7
IH 820	1	2176	1788	----	----	1982	0.8
	2	2272	1808	----	----	2040	3.0
US 183	1	2548	2244	2044	----	2279	3.2
	2	2608	2248	2388	----	2415	2.5
	3	2468	2116	2200	----	2261	2.6
US 290	1	2644	2388	2016	----	2349	2.8
IH 410	1	2348	2076	2120	----	2181	0.9

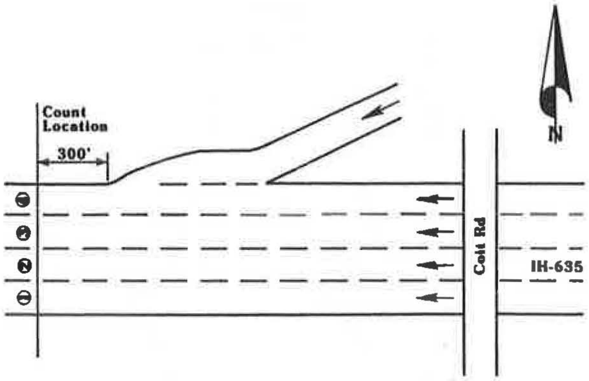


FIGURE 2 Schematic of I-635 study site.

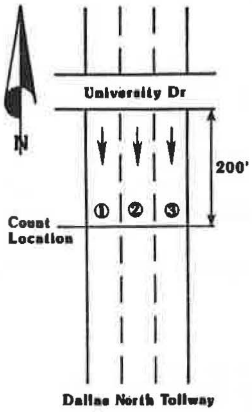


FIGURE 3 Schematic of Dallas North Tollway study site.

it had a regular southbound bottleneck and only two directional lanes, as shown in Figure 4. Data from other sites suggest that the outside lane generally operates at lower volumes than the other lanes. Given that most freeways are only four lanes, it is important to determine how cross section affects capacity. The observed peak 15-min flow rates averaged across the two lanes for the two observations were 1,982 and 2,040 vphpl with 0.8 and 3.0 percent trucks, respectively.

The US-183 site is a six-lane freeway in eastern Ft. Worth, located one interchange downstream of a freeway-to-freeway merge, as shown in Figure 5. The onramp at Central Drive does not add sufficient volume to create a bottleneck immediately upstream of the study site. Peak 15-min flow rates averaged across the three westbound lanes for three different

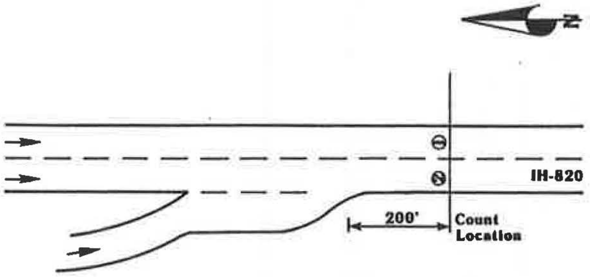


FIGURE 4 Schematic of I-820 study site.

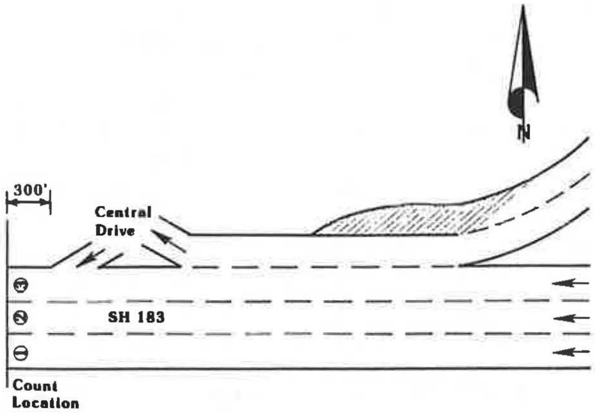


FIGURE 5 Schematic of US-183 study site.

observations were 2,279, 2,415, and 2,261 vphpl with 3.2, 2.5, and 2.6 percent trucks, respectively.

The US-290 site is on a six-lane radial freeway originating from a circumferential Interstate in Houston. The study site is approximately 1 mi downstream from a lane drop and exit ramp (see Figure 6). The median of this roadway also contains a concrete median barrier (CMB) separated authorized vehicle lane (AVL) and thereby allows for practically no inside shoulder and lane widths of only 11 ft. The observed peak 15-min flow rate averaged across the three westbound lanes was 2,349 vphpl with 2.8 percent trucks.

The I-410 site is a six-lane circumferential freeway in San Antonio. The study site is located on the eastbound side, just downstream of a high-volume entrance ramp and just upstream of an exit and fly-over entrance ramp (see Figure 7).

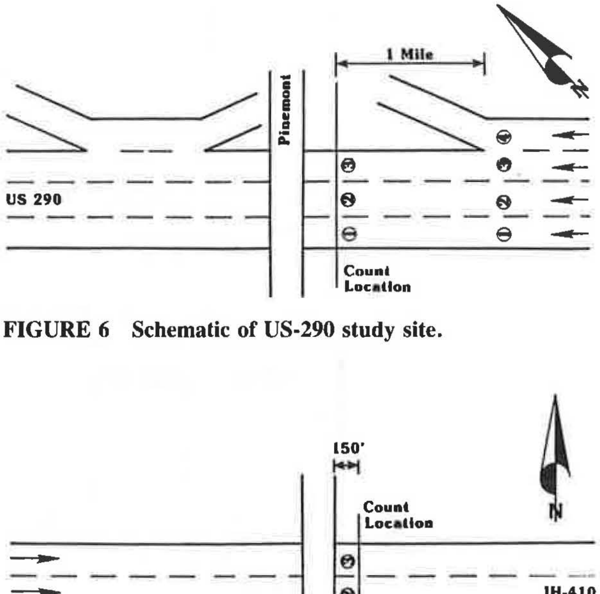


FIGURE 6 Schematic of US-290 study site.

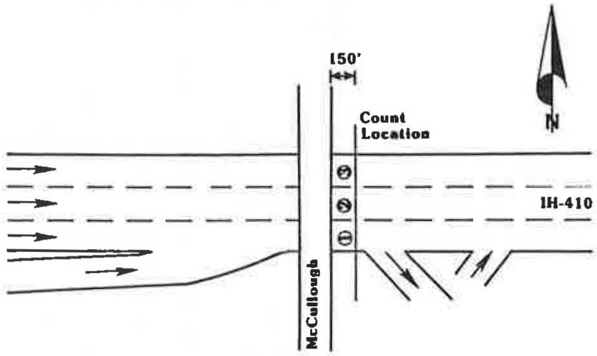


FIGURE 7 Schematic of Interstate 410 study site.

The peak 15-min flow rate averaged over the three lanes was 2,181 vphpl with 0.9 percent trucks.

TWO-CAPACITY HYPOTHESIS

The two-capacity hypothesis has important implications in collecting and interpreting data. Some examples illustrate the problems. If two capacities exist, it is necessary to know precisely when free-flow and forced-flow conditions occur because averages that do not segregate the data are corrupted by the two flow regimes. In addition, if two capacities exist, some locations can actually operate at flows higher than queue discharge capacity, yet below free-flow capacity. In other words, a bottleneck creates a metered flow less than free-flow capacity. If the ensuing ramp is an onramp and adds traffic that produces rates that do not exceed the free-flow capacity, flow rates are created that are higher than bottleneck discharge rates. The problem associated with this second observed flow is one of interpretation. If the flow regime is not understood, incorrect interpretations concerning bottleneck discharge might be made. This flow rate might not be sustainable at higher demand levels.

For highway design purposes, bottleneck discharge capacities represent the limiting case for operating freeways without using more sophisticated control systems. Clearly, free flow capacity, if significantly higher than bottleneck discharge capacity, is an important consideration in the operation of freeway systems.

The existence of forced-flow conditions upstream of a bottleneck is one indication of knowing that demand exceeds capacity. The following analysis was based on the existence of this type of flow condition (bottleneck discharge). Because it is necessary to know when a queue discharge exists to examine the two-capacity hypothesis, one study site was selected that provided the most advantageous viewing angle to observe the traffic flow. The I-35–US-67 site allowed a view from downstream of the bottleneck to several hundred feet upstream of the bottleneck. Figures 8–18 summarize the flow data collected at the I-35–US-67 study site.

Figure 8 shows the moving 15-min flow rates across all four lanes of traffic. The values plotted are flow rates calculated for every 15-min interval (moving 1 min at a time) throughout the 2-hr observation period. The flow rates are computed from the sum of 1-min counts, which represent the smallest interval examined. This method of presenting the data, which is equivalent to the concept of a moving average, smooths the data to allow trends to be seen. These 15-min flow rates can be contrasted with the corresponding 1-min flow rates shown in Figure 9. From this graph, it is much more difficult to discern the pattern of traffic flow. A stalled vehicle is shown in the left lane at approximately 7:21, but it was removed by 7:23.

The vertical reference lines on the graphs indicate the beginning and ending of upstream queues (7:05 and 7:47, respectively.) Figure 8 indicates that the highest flow exists around the time of the onset of congestion—the peak 15-min flow occurring between 7:01 and 7:15, inclusive. The cyclic fluctuations in flow rate over time are also significant. Different averaging schemes illustrate various aspects of this pattern of fluctuation, as shown by Figure 10, which shows a different representation of the same data, with 5- rather than 15-min flow rates. In this figure, the shorter interval illustrates more extreme fluctuations in traffic flow. The cyclic nature can still be observed, but shorter, more extreme cycles are apparent.

Figures 11–14 indicate that the individual 15-min lane flows are not at all similar. Lane 1 is the left lane of US-67. This lane exhibits a similar pattern to the before-queue flows but

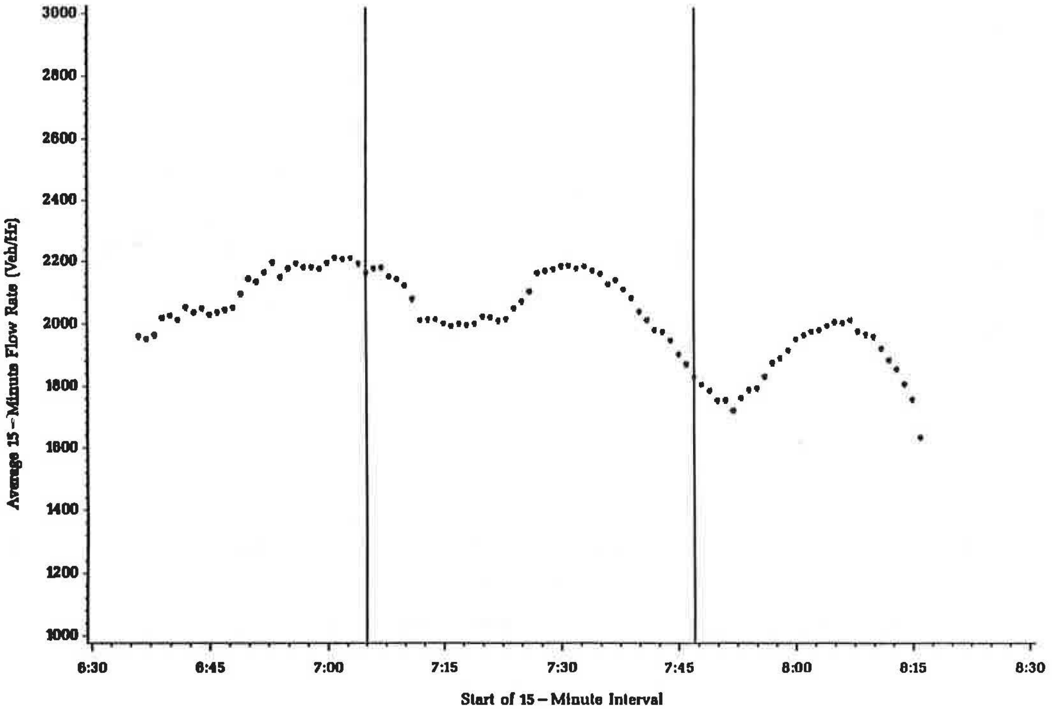


FIGURE 8 I-35–US-67 moving 15-min flow rates averaged across all lanes.

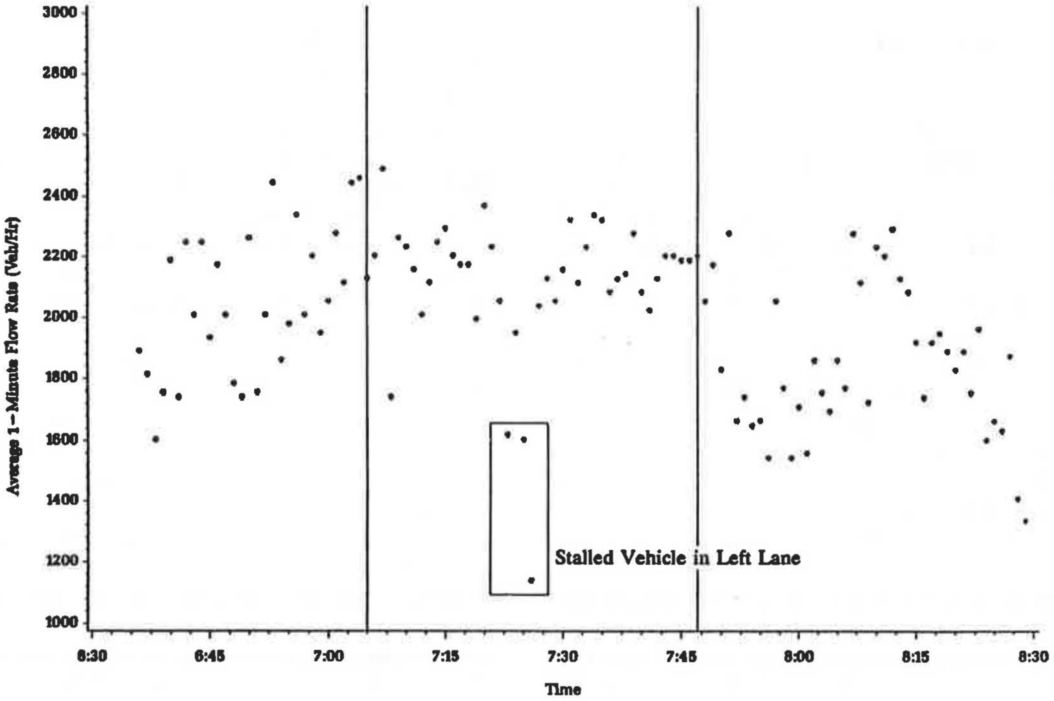


FIGURE 9 I-35–US-67 1-min flow rates averaged across all lanes.

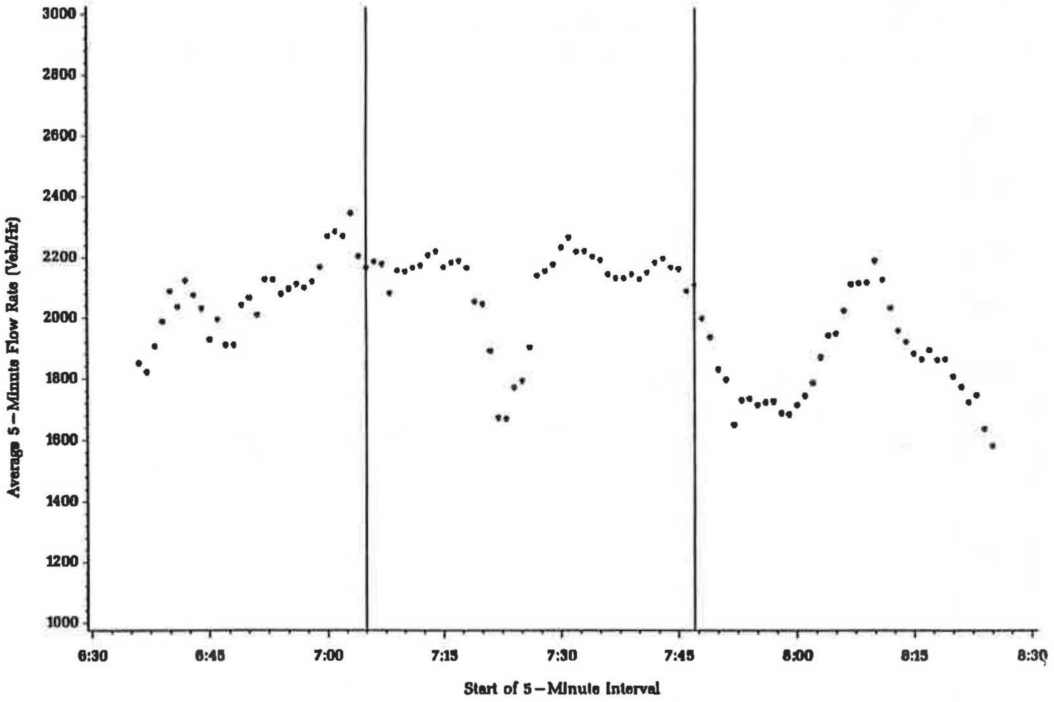


FIGURE 10 I-35–US-67 moving 5-min flow rates averaged across all lanes.

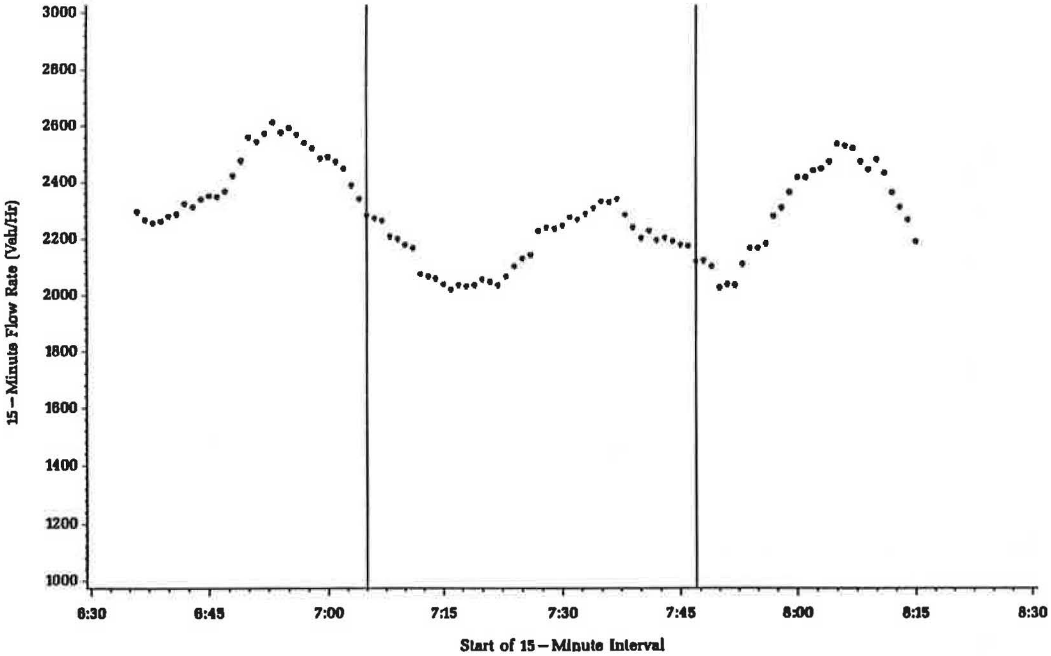


FIGURE 11 I-35–US-67 moving 15-min flow rates in Lane 1.

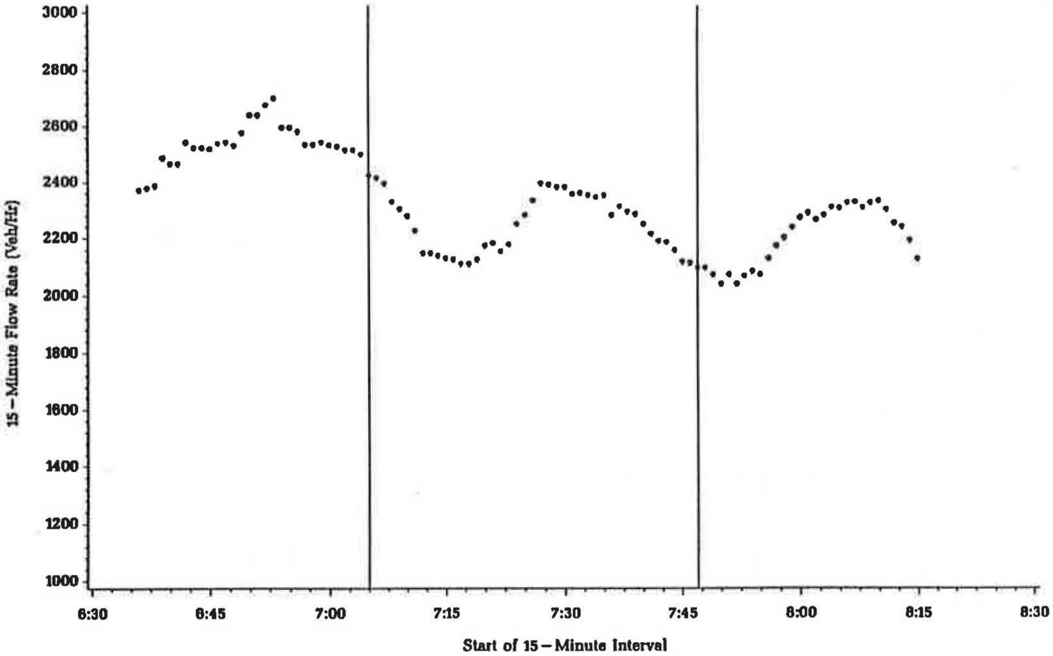


FIGURE 12 I-35–US-67 moving 15-min flow rates in Lane 2.

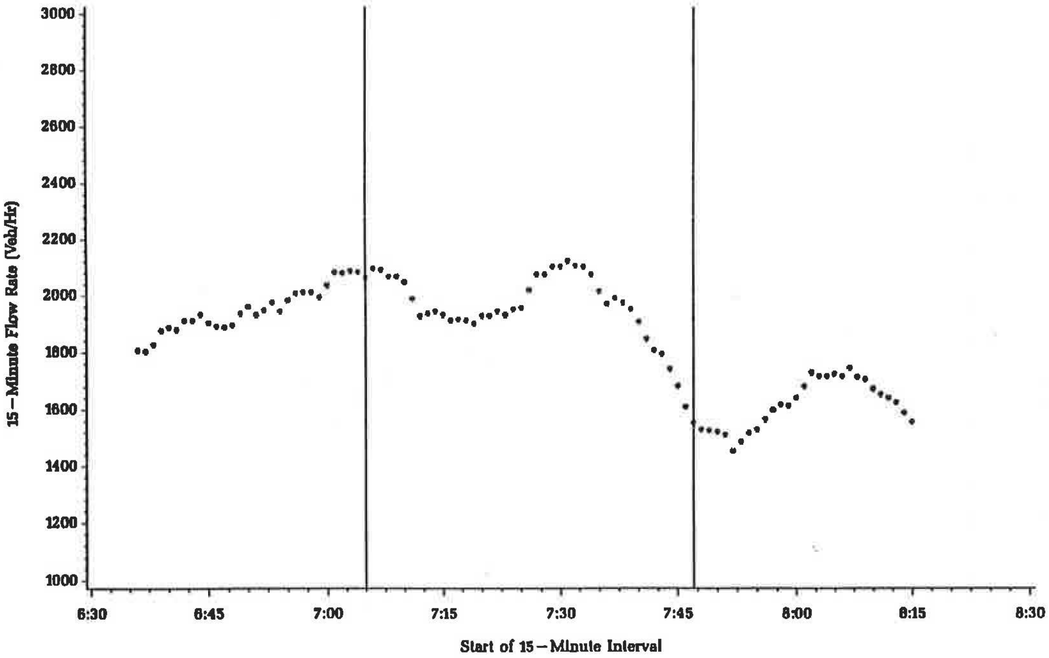


FIGURE 13 I-35-US-67 moving 15-min flow rates in Lane 3.

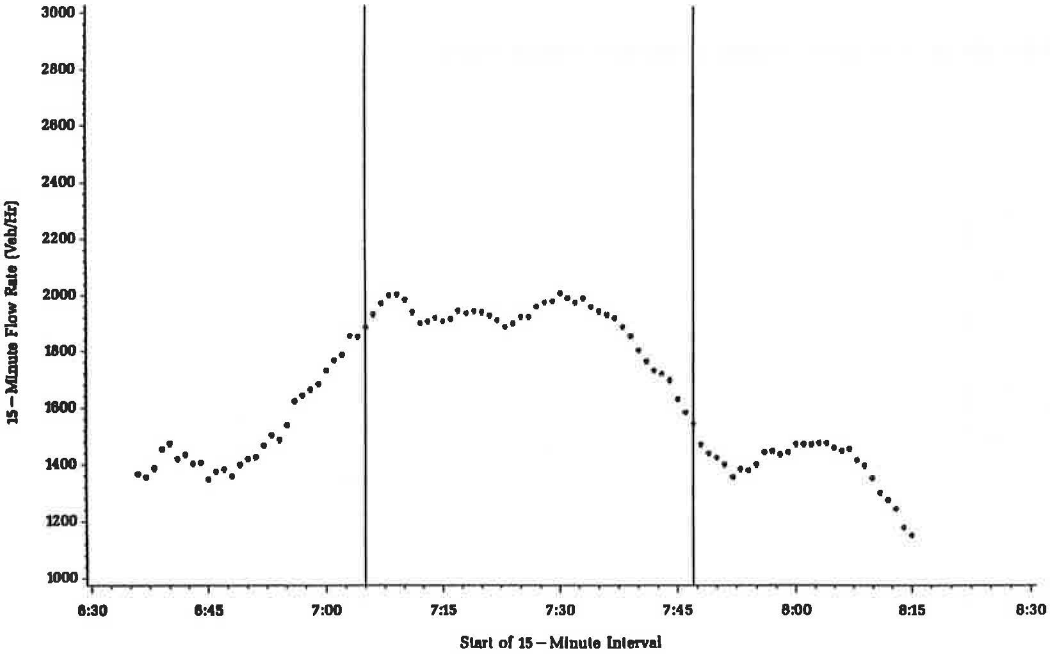


FIGURE 14 I-35-US-67 moving 15-min flow rates in Lane 4.

is higher than the after-queue flows. Lane 2 is the combination of thc US-67 right lanc and I-35 lcft-lanc flows. It exhibits higher flows before queueing and moderate flows after queueing. Lane 3, the middle lane on I-35, exhibits lower flows before and after the onset of queues. Lane 4, the outside lane on I-35, exhibits similar but more pronounced effects than those of Lane 3.

Figures 15-18 show the individual lane flows on the basis of 5-min intervals. These data better illustrate the effects shown previously. Plots of 1-min flow rates would show the extent of the scatter in the observed traffic flow, without illustrating the patterns that can be seen from the 5- and 15-min flow rates.

Some tentative observations are offered for further study. Capacity may be reached in only one or two lanes before a bottleneck forms. Bottleneck discharge rates are lower than prebottleneck flows in some lanes and higher than prebottleneck flows in other lanes.

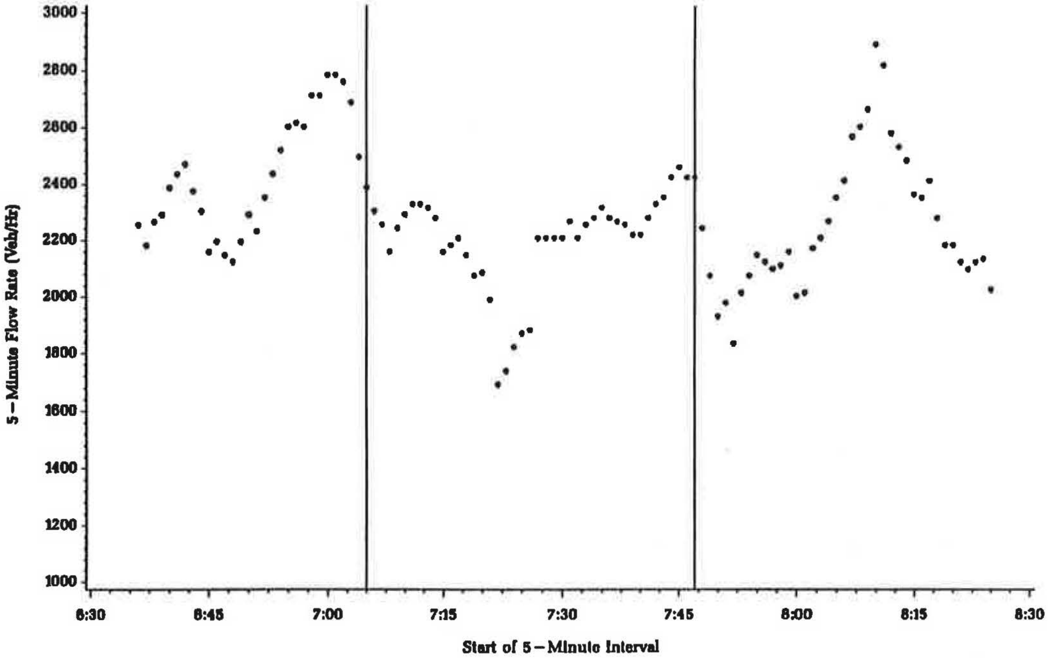


FIGURE 15 I-35-US-67 moving 5-min flow rates in Lane 1.

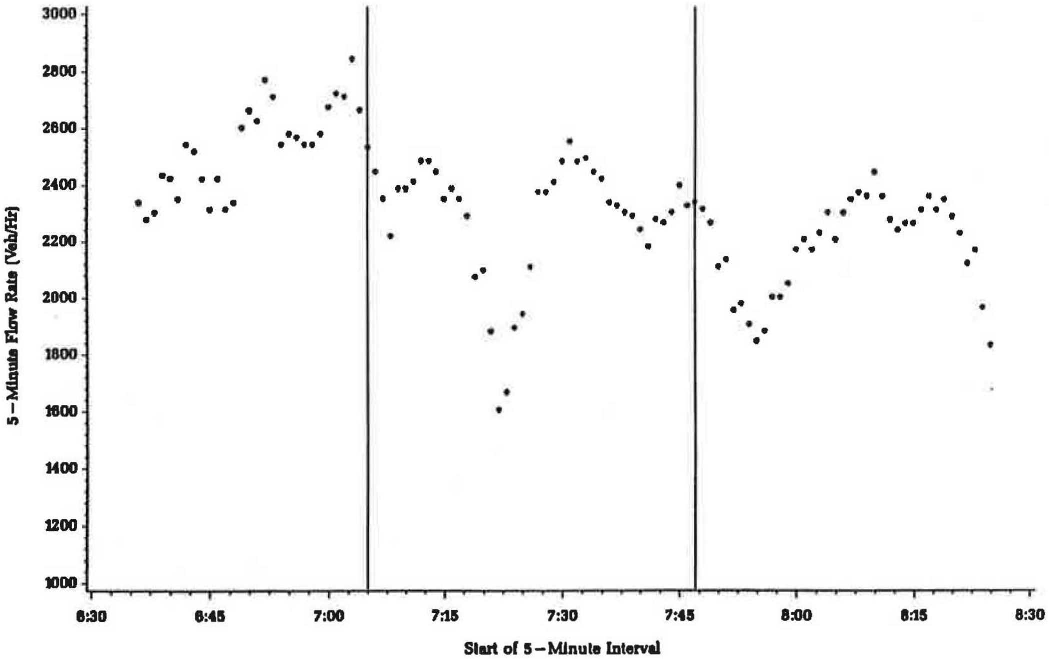


FIGURE 16 I-35-US-67 moving 5-min flow rates in Lane 2.

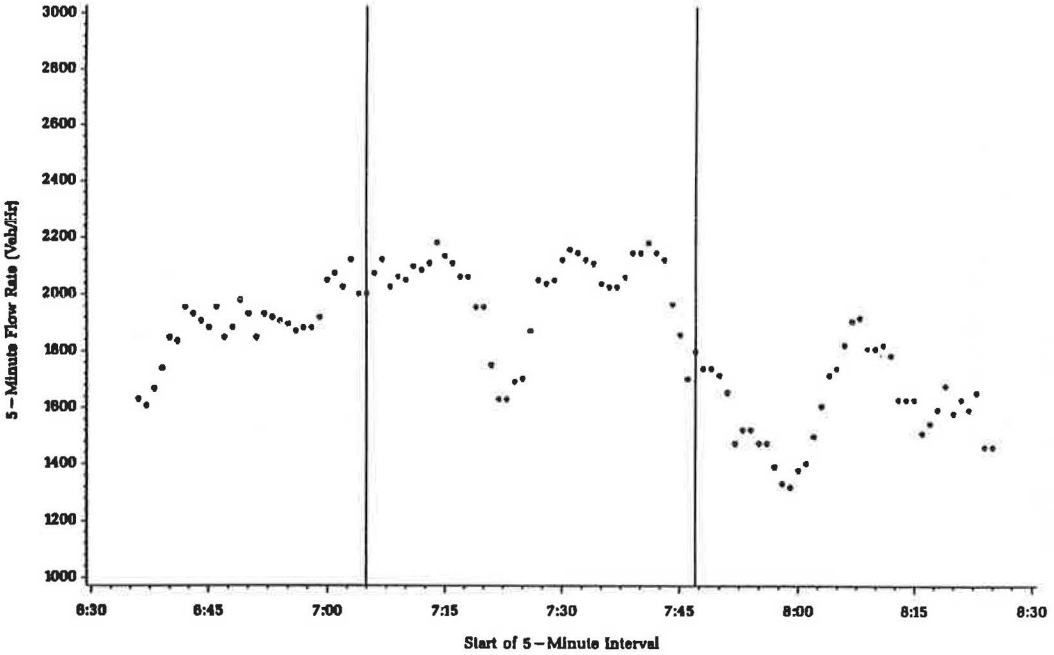


FIGURE 17 I-35–US-67 moving 5-min flow rates in Lane 3.

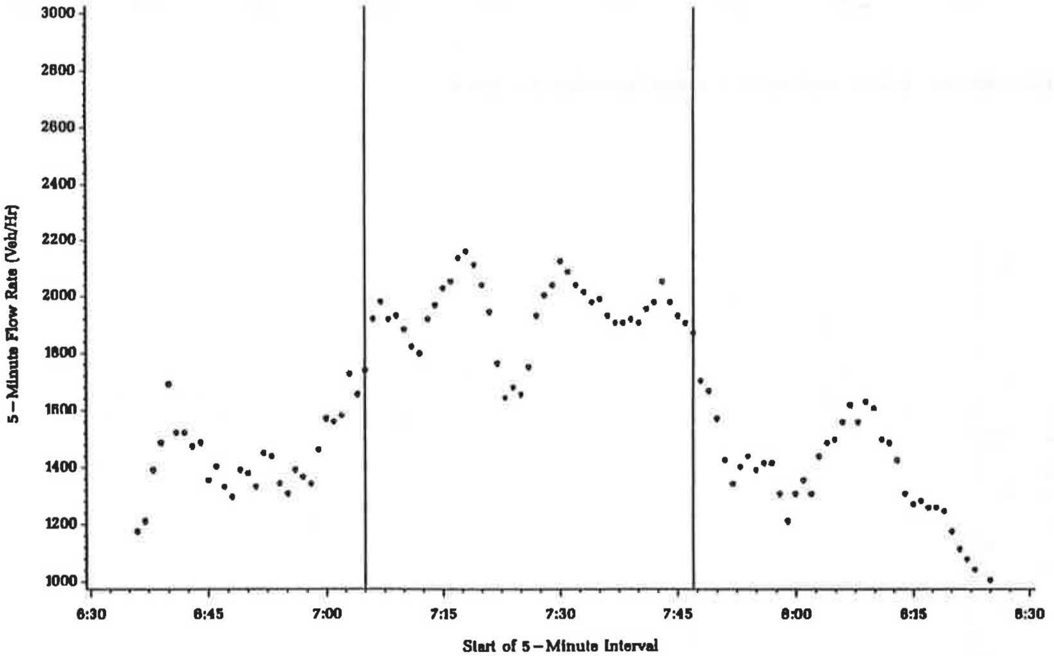


FIGURE 18 I-35–US-67 moving 5-min flow rates in Lane 4.

ACKNOWLEDGMENT

This study was sponsored by the Texas State Department of Highways and Public Transportation and FHWA, U.S. Department of Transportation.

REFERENCES

1. W. R. Reilly, D. W. Harwood, J. J. Shoen, R. O. Kuehl, K. Bauer, and A. D. St. John. *Capacity and Level of Service Procedures for Multilane Rural and Suburban Highways.* NCHRP Draft Final Report. TRB, National Research Council, Washington, D.C., May 1989.
2. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
3. F. L. Banks. Flow Process at a Freeway Bottleneck. Presented at 69th Annual Meeting of the Transportation Research Board, Washington, D.C., 1990.
4. F. L. Hall and L. M. Hall. Capacity and Speed Flow Analysis of the QEW in Ontario. Presented at 69th Annual Meeting of the Transportation Research Board, Washington, D.C., 1990.

The contents of this report reflect the views of the authors, who are responsible for the opinions, findings, and conclusions presented. The contents do not necessarily reflect the official views of FHWA or the Texas State Department of Highways and Public Transportation. This paper has not been reviewed by the sponsoring agency.

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

Toll Plaza Capacity and Level of Service

T. Hugh Woo and Lester A. Hoel

A methodology for evaluating the capacity of a toll plaza was developed, and level-of-service (LOS) criteria for toll area traffic were established. LOS for toll facilities should be quantified for several reasons. First, quantification of the LOS would enable designers to evaluate design alternatives using accepted standards. Second, it would provide a scientifically sound basis for comparing traffic operations of various facilities. Third, it would furnish a means to evaluate before-and-after conditions and thus determine the effectiveness of any improvement. Finally, it would give the general public and legislative representatives a readily understandable and yet scientifically established measure of overall performance. Traffic data at the four plazas of the Richmond-Petersburg Turnpike in Virginia were collected using synchronized video cameras. The capacity of the toll booths was found to range from 600 to 750 passenger cars per hour, depending on the type of toll collection. Average density, which is highly correlated with volume-to-capacity ratios, can be used as a criterion for defining LOS for toll plaza areas.

Toll financing has been used as a supplemental source of highway revenue since this nation was founded. The primary advantages of toll financing are (a) direct payment by each driver for services received, (b) an assured source of revenue that facilitates highway construction and maintenance, and (c) the ability to use tolls as a form of congestion pricing.

The construction of toll roads has experienced a resurgence in recent years. Numerous facilities are in varying stages of planning, design, and construction, not only across the United States but in many other countries as well. In the United States, 26 states have 4,700 mi of toll roads, bridges, and tunnels. Some 8500 km of toll highways has been built in France, Italy, and Spain (*1*).

Figures collected by the International Bridge, Tunnel, and Turnpike Association indicate that 16 states are considering $8.5 billion worth of proposals that would add 822 mi of toll structures (*2*, p. A3). One of the most innovative is an 80-km segment of the Denver metropolitan beltway in Colorado. Three Colorado counties and a small city formed the E-470 Highway Authority in 1986 to build the highway without federal or state funding (*1*). A feasibility study on a privately financed 400-mi toll road from Chicago to Kansas City is under way (*3*, p. 1). In a nationwide survey conducted in November 1988 by the *Urban Transportation Monitor*, 80 percent of the respondents indicated that their metropolitan area is either actively planning toll roads or will be doing so in the foreseeable future (*4*, p. 8). The respondents were transportation professionals involved in planning for their metropolitan area. The recent Transportation 2020 initiative by AASHTO recommended the use of toll collection as a financing option for highway construction (*5*, p. 72). Clearly, there is an increased emphasis on toll roads as a method of financing urban freeways.

The basic mechanism of toll collection has remained essentially unchanged since its inception: vehicles must stop to render payment at a collection booth. Stops at toll plazas, however, impede the smooth flow of traffic and, consequently, can reduce the level of service (LOS) provided. A toll plaza can be a bottleneck on a highway if its capacity is exceeded. The public accepts the notion of a fee to pay for roads but is unwilling to wait in traffic queues to render payment. In view of the time involved in each toll collection transaction, any improvement that can save even a fraction of a second will represent a substantial increase in efficiency. Efforts are now under way to develop new intelligent vehicle-highway system technologies that eliminate or reduce the delay at toll plazas. These include exact-change lanes, flash pass lanes, and automatic vehicle identification (AVI). Nonetheless, toll facilities continue to affect overall travel time and traffic flows by requiring each vehicle to stop. Surprisingly, the effects that toll plaza collection facilities have on the LOS and capacity have received little attention by researchers and highway design specialists.

Most roadway design features (such as freeway lanes, ramps, and intersections) have nationally accepted LOS standards, but the *Highway Capacity Manual* (HCM) (*6*) is silent on the subject of LOS standards at toll plazas, and no national design guidelines exist. The LOS for toll facilities should be quantified for several reasons. First, quantification of the LOS would enable designers to evaluate design alternatives using accepted standards. Second, it would provide a scientifically sound basis for comparing traffic operations of various facilities. Third, it would furnish a means to evaluate before-and-after conditions and thus determine the effectiveness of any improvement. Finally, it would give the general public and legislative representatives a readily understandable and yet scientifically established measure of overall performance.

Capacity is one of the factors of interest in design of a toll facility. Volume-to-capacity (v/c) ratios have been used in defining LOS for several highway facility types, such as basic freeway sections and rural highways. Nevertheless, the term "capacity" is not easily defined for toll facilities, and there appears to be no general agreement among traffic engineers as to its precise meaning.

There is essentially no literature describing nationally accepted design standards for toll facilities on the basis of LOS ratings. This void limits the application of a consistent nationwide design process. Neither the 1985 HCM nor its predecessor, the 1965 HCM, considers traffic characteristics at highway toll plazas. However, freeway toll plaza capacity and LOS were recognized as a research need by TRB in 1987 (*7*). Several reports have examined the issue of toll road financing, but few investigations of traffic characteristics at toll plazas

T. H. Woo, Department of Transportation Engineering and Management, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan 30049, Republic of China. L. A. Hoel, Department of Civil Engineering, University of Virginia, Charlottesville, Va. 22903.

are available. Toll plaza capacities have been established by performance observation or by assumption. Considerable variation was found among these individual results (*8–11*). A sound theoretical basis for capacity and LOS does not exist. The LOS for toll plaza performance has been used infrequently (*8,9,12*). A review of the literature indicates that there has not been any criterion proposed as a measure of effectiveness for determining the LOS for toll facilities.

The purpose of this research was to improve the understanding of traffic characteristics at toll plazas and to develop a methodology that could be used in the analysis of traffic operations at toll plazas.

MODEL DEVELOPMENT FOR CAPACITY AND LEVEL OF SERVICE

Capacity

A principal objective of capacity analysis is to estimate the maximum amount of traffic that can be accommodated by a given facility. Capacity analysis is also intended to estimate the service flow rate, which is the maximum amount of traffic that can be accommodated by a facility while maintaining prescribed operational qualities. The 1985 HCM (*6*) defines the capacity of a facility as the maximum hourly rate at which vehicles can reasonably be expected to traverse a point or uniform section of a roadway during a given period of time under prevailing roadway, traffic, and control conditions. Service flow rates are defined for each LOS rating.

A toll booth operates at full capacity when a queue is built up and the toll collector is busy at all times. When traffic is light, there is lag time between each toll collection, which means the toll collector is not fully occupied. The service time under a nonwaiting condition is sometimes intuitively mistaken to be shorter than it really is. Under light traffic, the toll collectors may actually consume more time than when pressured with a queue. When toll collectors are under greater pressure from a growing queue, they tend to process transactions faster.

The actual service time is affected by the number of coins that must be processed. It may also be influenced by such factors as the experience of the toll collectors, the physical dimensions of toll gates, the methods of toll collection, and the presence of drivers with exact change. Manual booths with heavy-truck traffic normally have lower service volumes than those that are primarily for automobiles. Traffic congestion levels also affect service time. When queues develop, motorists have time to search for needed change before the transaction.

If nonwaiting service time is determined as t_n seconds, the capacity in vehicles per hour of a toll booth occurs when the arrival rate equals $3{,}600/t_n$. Service time under waiting conditions, t_w, determines the processing rate in vehicles per hour as $3{,}600/t_w$. If it is assumed that $t_w \geq t_n$, then $3{,}600/t_n \geq 3{,}600/t_w$. However, when upstream traffic approaches $3{,}600/t_n$, queues start to occur. Thus, service time t_n is no longer valid, and t_w should be used. One of the many concerns of this study is the service time at traffic levels near capacity.

When traffic is at capacity and queues exist, an imaginary reference line can be located a short distance beyond the toll booth where the rear end of a vehicle passes when the following vehicle just stops to pay the toll. Once this reference line is defined, the service time can be obtained by observing traffic at a toll booth or by analyzing videotapes that contain toll booth traffic. For example, the service time for Vehicle C in Figure 1 starts when the vehicle stops to pay the toll and ends when the rear of Vehicle C passes the reference line, as shown in Figures 1B and 1C.

Service time can be determined for various toll collection types and geometric conditions. Assume that t_{ij} represents the service time in seconds for Vehicle Type i and Toll Collection Type j. Also assume that service time for trucks, trailers, or buses can be expressed in terms of service time for automobiles as follows:

$$t_{ij} = \alpha_i t_{1j} \tag{1}$$

where

t_{1j} = service time for an automobile (Type 1) and Toll Collection Type j, and
α_i = automobile equivalent at toll areas for Vehicle Type i.

It is apparent that vehicle length and the ability to accelerate are the two major factors that affect the service time. Nevertheless, longer vehicles generally have poorer acceleration, mostly because of the higher ratio of weight to power. Therefore, for simplicity, vehicles were divided into two categories: (a) automobiles and (b) trucks and buses. Thus, the denomination of α can be omitted, and the service time for trucks or buses under Toll Collection Type j is

$$t_{2j} = \alpha \cdot t_{1j} \tag{2}$$

Similarly, service time for different toll collection types can be converted to a common scale, such as manual collection, as follows:

$$t_{ij} = \beta_j t_{i1} \tag{3}$$

where

t_{i1} = service time for Vehicle Type i and manual toll collection (Type 1), and
β_j = conversion factor for Type j collection to manual toll collection.

Therefore, the capacity of a toll booth with Collection Type j is

$$c_j = \frac{3{,}600}{t_{1j}} \tag{4}$$

If there are n_j booths with Collection Type j, the capacity of a toll plaza would be

$$C = \sum_{j=1}^{J} n_j c_j = n_1 \frac{3{,}600}{t_{11}} + n_2 \frac{3{,}600}{t_{12}} + \dots + n_J \frac{3{,}600}{t_{1J}} = \sum_{j=1}^{J} n_j \frac{3{,}600}{t_{1j}} \tag{5}$$

Capacity c_j and C are both expressed in terms of passenger cars per hour.

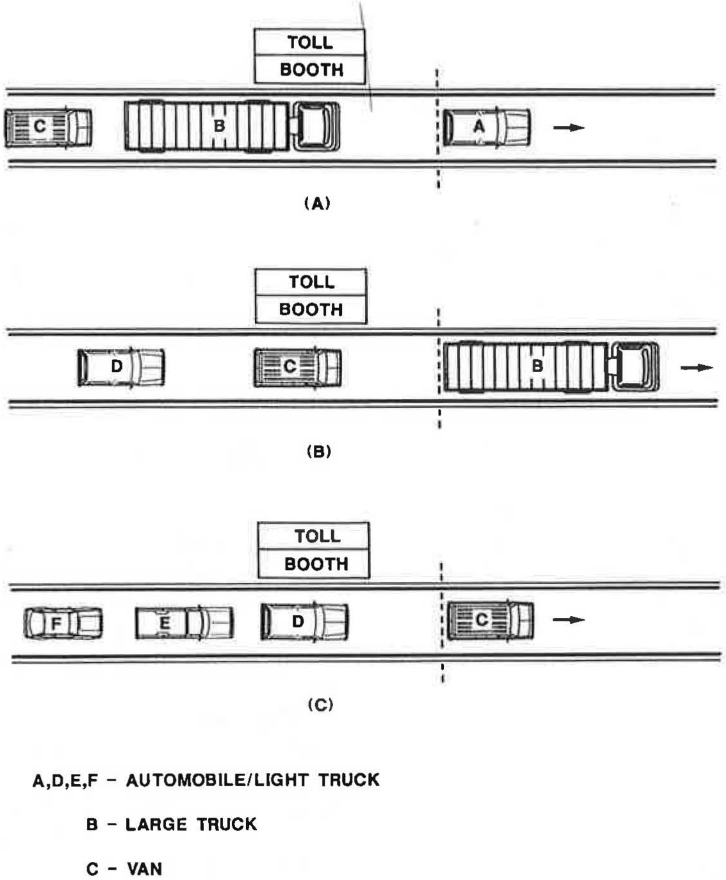


FIGURE 1 Vehicle movement at toll booth.

LOS

LOS is a qualitative measure describing operational conditions within a traffic stream and their perception by motorists. LOS for interrupted flow facilities, such as toll plazas, varies widely in terms of both the user's perception of service quality and the variables used to describe the operational conditions.

LOS analysis has been based on the use of one or two relatively simple measures that effectively evaluate the quality of traffic service. These measures are understood by the average motorist and are useful to the transportation analyst and to management in evaluating the relative need for specific improvements or for evaluating current operations.

The toll plaza is seen as a bottleneck in which traffic is regulated by the capacity of the toll plaza. Unless the toll plaza area is designed so that its downstream section has lower capacities than the toll plaza (which could happen if the number of available travel lanes downstream is less than that upstream), the number of vehicles passing a toll plaza would be a nondecreasing function of incoming traffic flow. This situation is shown in Figure 2. Before incoming traffic reaches the capacity of the toll plaza, toll booths are able to accommodate all traffic, as represented in Zone A of the figure. When traffic approximates the capacity of the toll plaza, delays occur. This condition is indicated as an unstable zone in the figure. If traffic exceeds capacity, the plaza can only serve at its capacity, as shown in Zone C.

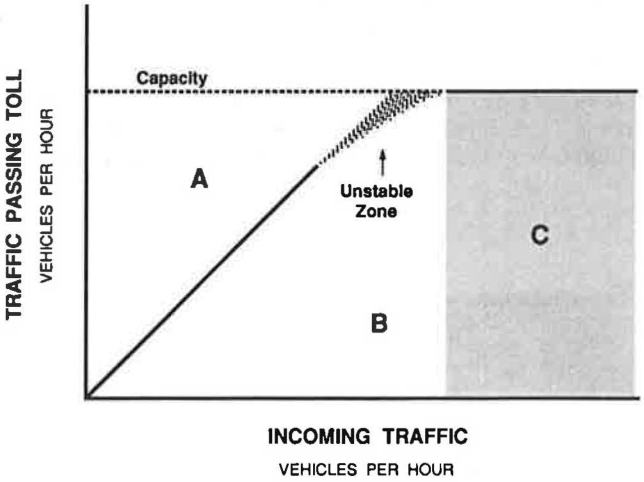


FIGURE 2 Traffic passing toll versus incoming traffic.

Traffic engineers often view overall speed as the most representative measure of traffic quality. Others believe that the true measure of service provided by a roadway is the volume of traffic it can handle and that a relationship between volume and capacity (in effect, between demand and supply) is a measure superior to that of speed, particularly because speed is a function of volume (*13*).

Traffic engineers have long been faced with the dilemma of relating capacity to LOS. Much of the difficulty exists because capacity is expressed in volume, whereas LOS is highly subjective in nature. Volume is a logical measure of efficiency from the point of view of the engineer, and density and the *v/c* ratio may be proper in addressing freedom to maneuver and proximity to other vehicles with respect to service quality. However, speed and the magnitude and frequency of speed changes are important measures of LOS from the point of view of the individual driver. Unlike freeway operating characteristics, which include a wide range of rates of flow over which speed is relatively constant, toll plaza traffic bears a wide range of speeds even at a constant rate of flow.

Although speed is a major concern of drivers with respect to service quality, freedom to maneuver and proximity to other vehicles are equally important parameters. These other qualities are directly related to the density of the traffic stream. Further, it has been found that rate of flow increases with increasing density throughout the full range of stable flows.

In traffic flow, density stands for number of vehicles per unit length of a travel lane. A toll plaza area is composed of two trapezoids (see Figure 3). The area is then

$$\begin{aligned} A &= A_1 + A_2 \\ &= 1/2(n_1 + n_2)L_1 + 1/2(n_2 + n_3)\mathrm{L}_2 \qquad (6) \end{aligned}$$

where

A = total area in length-lanes,
A_1 = area in length-lanes for convergence section,
A_2 = area in length-lanes for reconvergence section,
n_1 = number of arrival lanes,
n_2 = number of booths,
n_3 = number of departure lanes,
L_1 = length of convergence section, and
L_2 = length of reconvergence section.

If the average total time to travel through the toll plaza area is T, the number of vehicles appearing within this area should be the flow rate, Q, times T. If vehicles are placed in two categories, the density of a toll plaza area can be expressed as

$$K = \frac{\Sigma\, Q_iT_i}{A} = \frac{2(Q_aT_a + Q_tT_t)}{(n_1 + n_2)L_1 + (n_2 + n_3)L_2} \qquad (7)$$

where i represents vehicle type, and a and t denote vehicle types of automobiles and trucks, respectively.

Area includes not only the length but also the width of a toll plaza area. With flow within the limit (capacity), density increases when speed decreases, but the widening of the booth area absorbs some of the effects.

MEASUREMENT OF TRAFFIC FLOW AT TOLL PLAZAS

There are seven toll roads in Virginia. The majority of these facilities are located in the Richmond area, and the study sites selected are on the Richmond-Petersburg Turnpike. There are eight sites, representing a wide variety of traffic volumes and number of lanes (see Table 1).

VHS camcorders were used for data collection because they could transfer recorded information into broadcast-quality videotapes that could be analyzed at 1/30-sec precision. The camcorders were synchronously used in pairs, as shown in Figure 4. By setting a camcorder (Camcorder 1) on the top of a toll plaza pointing into the upstream traffic, data on volume, traffic mix, and delay could be collected. A second camcorder (Camcorder 2) pointed downstream to determine travel time, and the third camcorder (Camcorder 3) faced the toll booth from downstream to determine service time. Camcorder 2 was also set on top of the toll plaza, whereas Camcorder 3 was placed at a vantage point such as the top of a nearby toll office building or a platform on top of a van. Data were collected during the summer and fall of 1989.

The tapes collected in the field were transferred onto ¾-in. broadcast-quality videotapes before evaluation and pro-

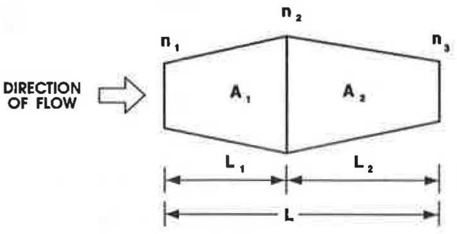


FIGURE 3 Area of toll plaza.

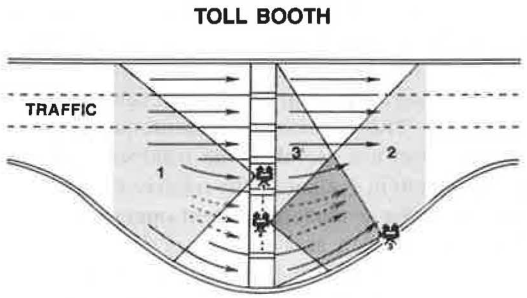


FIGURE 4 Layout for camcorders.

TABLE 1 STUDY SITES

Site	Area Distance (ft)	No. of Lanes n_1	n_2	n_3	1988 AADT
1	1313	3	8	3	89,070
2	1440	3	8	3	
3	1460	3	6	3	61,920
4	1250	3	6	3	
5	891	3	6	3	57,530
6	1274	3	6	3	
7	860	2	4	2	26,400
8	822	2	4	2	

cessing. The ¾-in. videotapes were then played back on equipment that included an edit controller, a timer screen, two video cassette recorders, and two monitors. Traffic counts were made for 15-sec intervals. The total number of vehicles observed at all sites was 15,746 (including 2,473 trucks).

EVALUATION OF CAPACITY AND LEVEL OF SERVICE

The location of the reference lines was determined by observing traffic on the videotapes. Only those vehicles that had an immediately following vehicle were recorded for the location at which their rear ends stood when that following vehicle stopped to pay the toll. The data are grouped by toll collection type, and automobiles immediately followed by another automobile were used to determine the reference lines. Table 2 presents the locations of the reference lines for the study sites. The reference line for a general toll booth lies about 55 ft from the point at which the toll attendant stands. For an exact-change booth, the reference line is between 1 and 7 ft closer to the plaza than it is for a general booth.

After the reference lines were located, the videotapes were viewed again to record the service time for as many vehicles as available by counting the time lag between the time at which a vehicle stopped to pay the toll and the time at which its rear end passed the defined reference line. Table 3 presents the results.

Comparisons of the service time for automobiles and trucks, automobiles at general booths, and automobiles at exact-change booths were carried out using the *t*-test for each study site. The results indicate that, at a 5 percent significance level, there is a significant difference in service time between automobiles and trucks (see Table 4). The automobile equivalents for trucks (α) range from 2.39 to 2.91 (see Table 3). The mean of α is 2.70.

At the time of data collection, there was an automatic lifting barrier at the exact-change booth at Site 2. The service time was therefore counted by measuring the time lapse after the barrier lifted between two consecutive automobiles.The *t*-test results in Table 5 indicate that the service time of this booth is significantly greater than that of general booths. Automobile drivers sometimes had to stop and wait after depositing change into the automatic collector before the barrier lifted. For all other sites, automobile service times at exact-change booths were either significantly shorter than or showed no difference than those at general booths. Nevertheless, the conversion factors (β_j) are close to 1.

The capacity of a toll booth and a toll plaza can be obtained from Equations 4 and 5. Table 6 presents the results. Capacities range from 650 to 705 automobiles per hour for a general toll booth and from 665 to 745 automobiles per hour for an exact-

TABLE 2 LOCATIONS OF REFERENCE LINES AT STUDY SITES, IN FEET

Site	General Booth	Exact-Change Booth
2	53.42	N/A*
3	54.28	N/A**
4	54.65	51.22
5	54.55	49.35
6	56.73	49.33
7	55.73	54.53
8	54.57	51.47

*Equipped with an automatic gate.
**No vantage location available for observation.

TABLE 3 SERVICE TIME FOR STUDY SITES, IN SECONDS

Site	Auto (t_{11})	Truck (αt_{11})	Exact-Change ($\beta_j t_{11}$)	α	β_j
2	5.47	14.40	6.53	2.63	1.19
3	5.17	14.43	N/A	2.79	—
4	5.44	14.77	5.21	2.72	0.96
5	5.21	14.23	5.33	2.73	1.02
6	5.11	14.88	4.83	2.91	0.95
7	5.32	14.58	5.25	2.74	0.99
8	5.39	12.87	5.41	2.39	1.00

TABLE 4 COMPARISON OF SERVICE TIME FOR AUTOMOBILES VERSUS TRUCKS

	Auto		Truck		*t*	
Site	*t**	*n***	*t**	*n***	Value	Remarks
2	5.47	661	14.40	30	–8.46	Significant
3	5.17	473	14.43	2	–2.75#	Not Significant
4	5.44	482	14.77	93	–18.21	Significant
5	5.21	293	14.23	77	–15.04	Significant
6	5.11	404	14.88	179	–22.58	Significant
7	5.32	248	14.58	143	–23.99	Significant
8	5.39	238	12.87	128	–20.83	Significant

# Due to small sample of trucks.
* Mean service time in seconds.
** Sample number.

TABLE 5 COMPARISON OF AUTOMOBILE SERVICE TIME FOR GENERAL BOOTHS VERSUS EXACT-CHANGE BOOTHS

	General		Exact-Change		t	
Site	t*	n**	t*	n**	Value	Remarks
2	5.47	661	6.53	230	−7.78	Significant
4	5.44	482	5.21	207	1.97	Significant
5	5.21	293	5.33	129	−0.98	Not Significant
6	5.11	404	4.83	246	2.78	Significant
7	5.32	248	5.25	160	0.59	Not Significant
8	5.39	238	5.41	90	−0.14	Not Significant

* Mean service time in seconds.
** Sample number.

TABLE 6 CAPACITY OF STUDY SITES

Site	General	Exact-Change	Plaza Total
2	650	591	5101
3	696	(691)*	4171
4	662	691	4001
5	691	675	4130
6	705	745	4270
7	677	686	2717
8	668	665	2669

*Based on the value of Site 4.

NOTE: Capacity measured in passenger cars per hour.

change booth. Because of the reaction delay of the lifting barrier, the capacity of the exact-change booth is 591 automobiles per hour, which is the lowest of any booth observed.

The traffic and travel time data were segregated into 1-, 3-, and 5-min periods. Once the capacity of each site was determined, the v/c ratio was obtained by

$$v/c = \frac{\text{Volume}}{\text{Capacity}}$$

where volume represents the flow rate in passenger cars per hour. Trucks were converted into equivalent numbers of passenger cars by α.

By playing back tapes taken by Camcorders 1 and 2, the time a vehicle entered and left the toll plaza area could be obtained from the readings on the timer. Total travel time, then, is the difference of the two observed values. Travel time was recorded for as many vehicles as could be identified, and the data were separated for automobiles and trucks (including buses). Travel times for 12,737 vehicles, which represented more than 80 percent of the observed traffic, were recorded from the eight study sites.

For a standard design and for all of the study sites, the number of arrival lanes equals the number of departure lanes, that is, $n_1 = n_3$. If $n_1 > n_3$ or $n_1 < n_3$, the one with fewer lanes must be overloaded or congested, or the other one must be a slack design under design flow. Therefore, the area of a toll plaza in length-lanes is as follows:

$$A = 1/2(n_1 + n_2)L_1 + 1/2(n_1 + n_2)L_2$$

$$= 1/2(n_1 + n_2)(L_1 + L_2) = \frac{L}{2}(n_1 + n_2) \qquad (8)$$

Equation 7 becomes

$$K = \frac{Q_aT_a + Q_tT_t}{A} = \frac{2(Q_aT_a + Q_tT_t)}{(n_1 + n_2)L} \qquad (9)$$

The units in Q_i and T_i in Equation 9 should be the same. Density was calculated using Equation 9 for all data points observed.

An analysis of the relationship between the v/c ratio and density was performed. Table 7 presents the regression results, and Figures 5–7 show the scatter diagrams for each data set. The values of R^2 and PROB > F indicate statistical significance between these two variables and a high degree of correlation with a quadratic model. Thus, density is a good predictor of the v/c ratio for toll plaza traffic.

The density criterion used for basic freeway segments and multilane rural highways is also correlated to the v/c ratio value with a quadratic model, as presented in Table 8. Figure 8 shows a plot of the models from Tables 7 and 8.

TABLE 7 v/c-DENSITY REGRESSION MODELS

Data Set	Model	R^2	PROB>F
1-min	$v/c = -0.000105K^2 + 0.0216K + 0.0243$	0.7900	0.0001
3-min	$v/c = -0.000141K^2 + 0.0246K - 0.0408$	0.7560	0.0001
5-min	$v/c = -0.000122K^2 + 0.0234K - 0.0293$	0.7417	0.0001

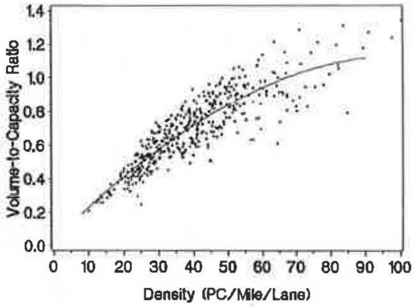


FIGURE 5 The v/c-density diagram for 1-min data set.

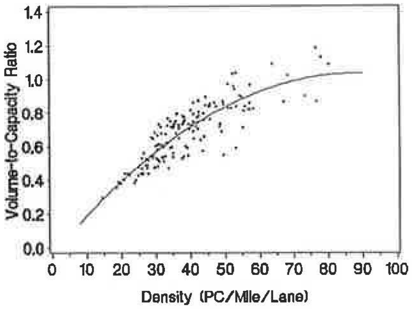


FIGURE 6 The *v/c*-density diagram for 3-min data set.

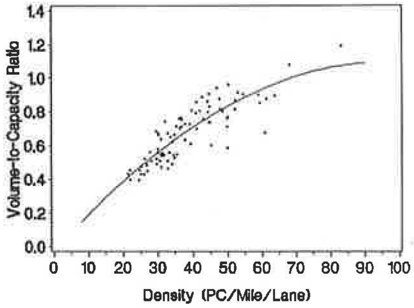


FIGURE 7 The *v/c*-density diagram for 5-min data set.

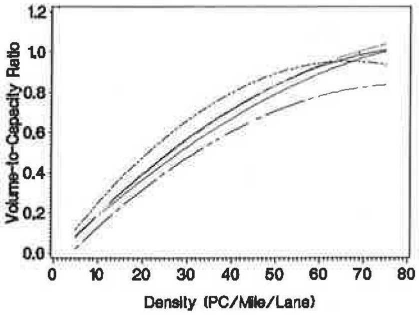


FIGURE 8 Plots of *v/c*-density models (*top to bottom*): freeway segments, 1-min data set, 5-min data set, 3-min data set, multilane highways.

Concerning the major parameters of stream flow used to define the level of service, Roess et al. (*14*) pointed out that the driver experiences speed and density but that the designer or analyst is most interested in the volumes that can be accommodated. Roess et al. recommended that the level of service be defined in terms of the parameters directly experienced by drivers: speed and density. These parameters should then be related to volume for the use of designers, analysts, and planners.

The density for a *v/c* ratio of 1.0 at toll plaza areas is at 67 passenger cars per mile per lane (pc/mi-ln), which is the critical density at which capacity most often occurs for stable flows. This value is the same as that used for freeway segments and multilane highways (see Table 9). This critical density is found to be valid for the flow-density models and *v/c*-density models used in this study. Therefore, current boundary values of density are adopted for the various levels of service for toll plaza areas. On the basis of the *v/c*-density relationships shown in Figures 5–8, the corresponding boundary values of the *v/c* ratio are then proposed (see Table 10). To be within a given level of service, the density criterion must be met. The *v/c* ratios indicated in the table are expected to exist for the given densities, although they may vary to some extent. Defining levels of service in this way takes into account parameters that are directly perceived by drivers and are of greatest interest to engineers.

Because most toll mechanisms exist within Interstate and multilane highways, it is a unique advantage to adopt density as the LOS criterion. This advantage warrants the continuance of the same criterion. The major difference lies in the decrease in average running speed.

Consider a stable platoon of traffic that enters a toll plaza area. The density will tend to decrease because of the area widening at the beginning of vehicle divergence; however, the evident reduction of travel speed in the toll plaza area may immediately offset this density increase.

When traffic volume and pattern are unchanged and the toll collection process is improved, as occurs with implementation of the AVI technique, the capacity will increase and the *v/c* value will decrease accordingly. Because this improvement also reduces the total travel time (from Equation 9), the corresponding density should also decrease. Adding a new lane produces a similar situation. It increases not only the

TABLE 8 REGRESSION MODELS FOR PRESENT *v/c*-DENSITY CRITERIA

	Model	R^2	PROB>F
Freeway Segments	$v/c = -0.000221K^2 + 0.0294K - 0.0228$	0.9746	0.0001
Multilane Highways	$v/c = -0.000143K^2 + 0.0231K - 0.0891$	0.9761	0.0001

TABLE 9 LOS CRITERIA IN 1985 HCM (*6*)

Design		Freeway Sections			Multilane Highways		
Speed		70 MPH	60 MPH	50 MPH	70 MPH	60 MPH	50 MPH
LOS	Density	v/c	v/c	v/c	v/c	v/c	v/c
A	≤ 12	0.35	—	—	0.36	0.33	—
B	≤ 20	0.54	0.49	—	0.54	0.50	0.45
C	≤ 30	0.77	0.69	0.67	0.71	0.65	0.60
D	≤ 42	0.93	0.84	0.83	0.87	0.80	0.76
E	≤ 67	1.00	1.00	1.00	1.00	1.00	1.00
F	> 67	—	—	—	—	—	—

TABLE 10 PROPOSED LOS CRITERIA FOR TOLL PLAZA AREAS

LOS	Density (pc/mi/ln)	v/c
A	< 12	0.24
B	< 20	0.40
C	< 30	0.57
D	< 42	0.74
E	< 67	1.00
F	> 67	—

capacity but also the area (A). Hence, v/c and density decrease at the same time. Therefore, the v/c-density relationships developed here still hold.

The following general operating conditions are proposed for each of the levels of service:

1. *Level of Service A.* There is very low density and delay. The operation of vehicles is virtually unaffected by the presence of other vehicles, although deceleration is necessary because of the toll plaza. The maximum average vehicle density is 12 pc/mi-ln. Most vehicles do not completely stop if change or a receipt is not required. Minor disruptions are easily absorbed. Standing queues will not form, and the general level of comfort and convenience is excellent.
2. *Level of Service B.* There is a maximum average density of 20 pc/mi-ln. Vehicles start to decelerate earlier upstream than for Level of Service A. The presence of other vehicles in the traffic streams begins to be noticeable. However, there is a good opportunity for lane change throughout the whole area. Minor disruptions can be easily absorbed at this level, although local deterioration in level of service at individual booths is more obvious.
3. *Level of Service C.* There is a maximum average density of 30 pc/mi-ln. The number of vehicles stopping is significant. The higher delays result from the early deceleration as well as the occasional complete stops for paying tolls. Minor disruptions can be expected to cause serious local deterioration in service. Queues establish at times.
4. *Level of Service D.* There is a maximum average density of 42 pc/mi-ln. Vehicles have little freedom of maneuver within the approaching section to choose a toll lane. Queue length becomes significant, and stop-and-go conditions are inevitable.
5. *Level of Service E.* There is a maximum average density of 67 pc/mi-ln. Every vehicle joins a queue before arriving at a toll booth. Stop-and-go traffic is a typical phenomenon. Maneuvering within the approaching section is almost impossible.
6. *Level of Service F.* There is an average density higher than 67 pc/mi-ln and a v/c ratio of more than 1. This condition often occurs when arrival flow rates exceed toll service rates. Queues continue to increase beyond the divergence point.

The level-of-service concept does not relate volume to density in a one-to-one relationship. For example, it is possible to have density at the range of LOS E while the v/c ratio is below that range in some instances. Similarly, it is possible to have the v/c ratio at the range of LOS C while the density is below that range. Nevertheless, the v/c ratio and density can be fairly reliable predictors of each other, as the regression models indicate in Table 7.

CONCLUSIONS

Service Times

For determining service time for a toll booth, an imaginary reference line was developed. This line is the point over which the rear of a vehicle passes when the following vehicle just stops to pay the toll. The reference line was located 55 ft from where the toll attendant stands for a general toll booth (50 ft for an exact-change booth). Service times at toll booths ranged from 12.87 to 14.88 sec for trucks and from 5.11 to 5.47 sec for automobiles.

Service times for trucks are significantly longer than those for automobiles. Automobile service times at exact-change booths are shorter or exhibit little or no difference than those at general booths.

Capacity

The capacity of a general booth is found to range from 650 to 705 passenger cars per hour. The capacity of an exact-change toll booth without a lifting barrier is between 665 and 745 passenger cars per hour. An exact-change toll booth with a lifting barrier has a capacity of 600 passenger cars per hour, which is lower than other arrangements because of the delay inherent in the automatic collection machines.

Density and Level of Service

Density is the most suitable variable for use in determining level of service for toll plaza areas. Density is a good indicator of the degree of freedom available to drivers and the operating speeds that can be achieved. Because vehicles are not evenly distributed throughout a toll plaza area, an average value is most appropriate.

Densities of toll plazas were found to be highly correlated to the corresponding v/c values with a quadratic model. Therefore, if density is used as the parameter for defining the level of service of a toll plaza area, volume can be used as a measurement. On the basis of these findings and current standards, LOS criteria were proposed for toll plaza areas in terms of density (see Table 7).

This proposal warrants two levels of analysis for traffic engineers who want to determine the level of service of an existing toll plaza or design a new one. The analysis procedures are similar to those presented in the 1985 HCM for other types of highway facilities.

Limitations

Geometric Condition

Because of geometric restrictions, on- and offramps may be placed next to toll plazas, which is not the case for any of the study sites. These ramps produce more traffic weaving and conflicts; thus, they affect capacity and increase travel times. Other adverse geometric conditions, such as sharp horizontal or vertical curves, would also affect travel times.

Vehicle Category

For simplicity, vehicles were divided into only two categories: (a) automobiles and (b) trucks and buses. If traffic consists of vehicles that can be clearly separated into more than two categories, the analysis should take this factor into account. For example, three categories could be used: (a) automobiles, (b) single-unit trucks, and (c) truck trailers.

Toll Collection Method

The Richmond-Petersburg Turnpike does not have advanced toll collection equipment. The results of this study should not be applied without modifying toll facilities with advanced electronic or optical collection mechanisms.

GLOSSARY

Toll Booth: a booth at which vehicles stop to pay tolls.
Toll Plaza: one or more toll booths.
Toll Plaza Area: a specific section that begins with vehicle convergence into the toll plaza and ends with vehicles reconverging into the highway traffic stream.
Service Time: the elapsed time for a vehicle to stop, pay a toll, and clear the area so that the next vehicle in the queue is positioned to pay the toll.
Density: the number of vehicles per unit area. Because vehicles are directionally guided rather than randomly spread in moving, the unit of area used is the mile-lane.

ACKNOWLEDGMENTS

This study was conducted with funds provided by FHWA through the Virginia Transportation Research Council. The authors gratefully acknowledge the assistance provided by both agencies.

REFERENCES

1. IRF Reviews Topics of World Interest. *World Highway*, Vol. XLI, No. 1, 1990.
2. S. C. Fehr. Toll Roads Gain in Popularity From Coast to Coast. *Washington Post*, Sept. 20, 1989.
3. New Toll Road Models. *AASHTO International Transportation Observer*, Oct. 1988.
4. Editorial. *The Urban Transportation Monitor*, Vol. 2, No. 24, 1988.
5. *Beyond Gridlock*. Highway Users Federation, Washington, D.C., 1988.
6. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
7. *Transportation Research Circular 319: Research Problem Statements: Highway Capacity*. TRB, National Research Council, Washington, D.C., June 1987.
8. H. C. Wood and C. S. Hamilton. Design of Toll Plazas from the Operator's Viewpoint. *HRB Proc.*, Vol. 34, 1955, pp. 127–139.
9. G. G. Henk, E. C. Pahlke, and M. C. Schaefer. Toll Road Facility Design Issues—Planning for Flexibility. Presented at Conference on National Perspectives on Toll Road Development, Irvine, Calif., Nov. 1988.
10. W. G. Bald and J. R. Underwood. An Investigation into the Effect of Toll-Booths on Morning Peak Traffic Flow. *Chartered Municipal Engineer*, Vol. 104, No. 11, 1977, pp. 205–208.
11. J. D. Griffiths and J. E. Williams. Traffic Studies on the Severn Bridge. *Traffic Engineering and Control*, May 1984, pp. 268–274.
12. L. D. Edie. Traffic Delays at Toll Booths. *Journal of the Operations Research Society of America*, Vol. 2, 1954, pp. 107–138.
13. J. F. Schwar. Quality of Traffic Service. *Traffic Quarterly*, Vol. XX, No. 1, 1966, pp. 136–146.
14. R. P. Roess, W. R. McShane, and L. J. Pignataro. Freeway Level of Service: A Revised Approach. In *Transportation Research Record 699*, TRB, National Research Council, Washington, D.C., 1979, pp. 7–16.

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

Flow and Capacity Characteristics on Two-Lane Rural Highways

Abishai Polus, Joseph Craus, and Moshe Livneh

Traffic operation on two-lane rural highways is unique; no lane change is permitted, and overtaking and passing are possible only in the opposing lane of the oncoming traffic. Traffic flow and capacity characteristics on two-lane highways were investigated. Several models were developed for the relationships between flow parameters. The relationships varied from one road to another and were dependent on the characteristics of each site. Thus, the hourly capacity volume is not a clear, single value. It primarily depends on the geometric and traffic characteristics of the highway section and on the time interval for which the flow data are collected. The two-way hourly capacity value was found to be about 2,650 passenger cars per hour in both directions. This value was close to the general (extended section) capacity given in the *Highway Capacity Manual* (HCM) for conditions similar to those in this study. The speeds at which capacity occurs, on the other hand, were lower than the HCM values: about 40 km/hr versus about 70 km/hr. Further research on the proper time interval for an analysis of highway capacity and on flow regimes and capacity values for various geometric, terrain, and traffic characteristics is suggested.

Traffic operation on two-lane rural highways is unique: no lane change is permitted, and overtaking and passing are possible only in the opposing lane of oncoming traffic. This situation results in interaction and influence between the two traffic directions. For example, as the ability to pass declines, drivers are forced to adjust their individual speeds.

The quality of service on two-lane highways is measured by three main traffic characteristics: average travel speed, percentage time delay, and capacity utilization. The average travel speed reflects the mobility function of the road; speeds over 90 km/hr are often observed on major highways. The percentage time delay reflects both the mobility and the access function. The 1985 *Highway Capacity Manual* (HCM) (*1*) defines this delay as the average percentage of time that all vehicles are delayed while traveling in platoons because of the inability to pass. Because it is difficult to measure this delay in the field, an evaluation should be made of the percentage of vehicles that travel in headways of less than 5 sec and this figure should be used as a surrogate variable. The third level-of-service (LOS) measure is a ratio of the demand flow rate to the capacity of the facility. This measure reflects the ability of drivers to drive in comfort without excess interaction with other vehicles.

Traffic flow and capacity characteristics on two-lane highways were investigated. To that end, several models that develop relationships between flow parameters, such as volume and density, were constructed. The basis for this work was a recent study sponsored by the Israel Public Works Department (*2*).

BACKGROUND

Speed-flow and delay-flow relationships for two-lane rural highways, under ideal conditions, have been presented by the HCM (*1*) and are shown in Figure 1. Ideal conditions on two-lane highways almost never exist, mainly because of the presence of trucks and buses and often because of the restriction imposed by the passing-sight distance. Naturally, both of these parameters have a direct and significant effect on the capacity of two-lane highways.

The methodology to determine capacity on two-lane highways is quite different for general terrain segments and for specific sections. Designers are often interested in the capacity of a specific section because it may constitute the critical one, such as the bottleneck of a long stretch of highway. The method for determining the capacity of a specific section is more complex than that for a long section because the speed at which capacity is reached is difficult to determine. The HCM (*1*) provides a model relating the flow rate and speeds at various levels of service. The HCM also gives a model that relates flow rate and speeds, at capacity, as follows:

$$S_c = 25 + 3.75(V_c/1,000)^2 \qquad (1)$$

where

S_c = speed at which capacity occurs (mph), and
V_c = flow rate at capacity (veh/hr).

The plots of the two models yield two curves that intersect. This intersection defines the flow rate at capacity and the speed at capacity for a specific two-lane rural section. An example is shown in Figure 2.

The capacity suggested by the new HCM (*1*) for two-lane rural highways is 2,800 passenger vehicles per hour, both directions, under ideal conditions. The old HCM (*3*) gave an ideal value of 2,000 passenger vehicles per hour. Most other studies in recent years have reported flows of more than 2,000 veh/hr. Yager's (*4*) survey of various studies showed values for two-directional flows ranging between 3,550 and 2,070 veh/hr.

The measurement of capacity is affected by traffic composition, particularly the presence of large trucks in the traffic stream. Capacity estimates are also sensitive to the assumed general form of the data and to inherent problems in the curve

A. Polus, K&D Facilities Resource Corp., 1021 West Adams, Chicago, Ill. 60607. J. Craus and M. Livneh, Department of Civil Engineering, Israel Institute of Technology, Technion City, Haifa, Israel.

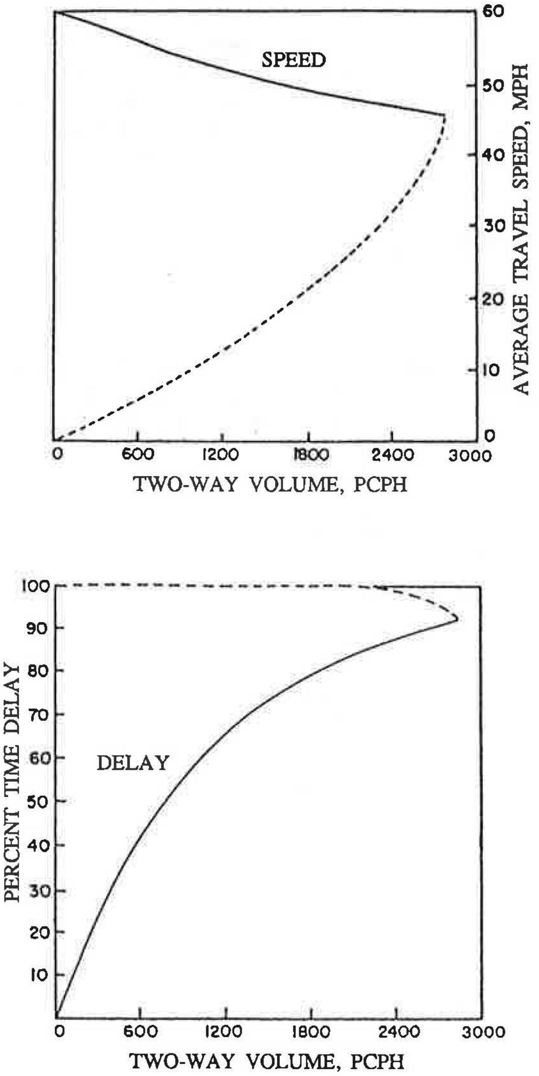


FIGURE 1 Speed-flow and delay-flow relationships at ideal conditions on two-lane rural highways (*1*).

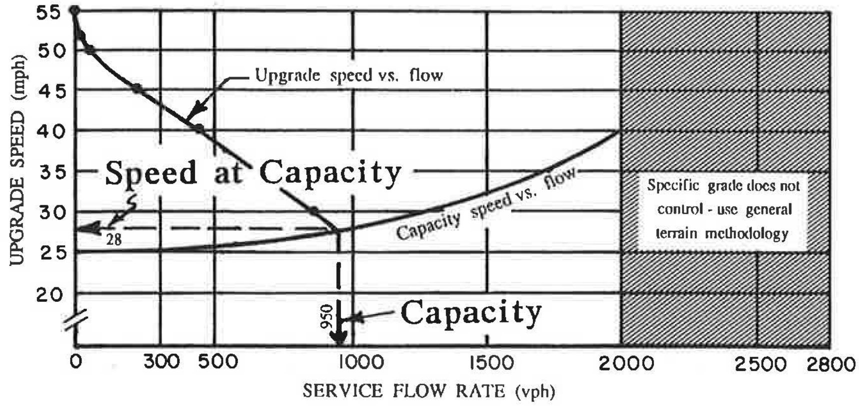
fitting of speed-volume data. Duncan (*5*) discusses the dangers of converting speed-concentration (density) data to speed-flow data. The danger exists particularly if the original data are composed of two different groups: one of free-flowing vehicles and the other of queued vehicles, the behavior of which is not rélevant to capacity.

The measurement of capacity is also dependent on the time interval for which data are collected. The intervals should not be too short (e.g., 1 min) because capacity may be overestimated. If the interval is too long (e.g., 1 hr), capacity will probably be underestimated because the demand may not be sufficient to provide saturation flows for the entire hour. This problem has been addressed by an Organization for Economic Cooperation and Development (OECD) Scientific Expert Group (*6*), which suggests that the time interval should be the longest possible period consistent with upstream demand.

It is clear, then, that the exact value of capacity is not easily determined. In fact, this measurement is so sensitive that it may vary on the basis of traffic demand fluctuations, short variations in speeds and traffic compositions, and other random phenomena (such as weather). In light of this variability, it can be argued that the exact value of capacity is of lesser importance in the design process of two-lane roads. Rather, it is of interest in determining the flow rates for the different levels of service, particularly LOS C and D, which often serve as design levels of service.

DATA COLLECTION AND PRELIMINARY OBSERVATIONS

Data for this study came from two busy two-lane highways in Israel. The two sites were initially selected because the average daily traffic (ADT) was in excess of 10,000 veh/day and, hence, the probability of obtaining dense flow conditions was high. Data were collected only during the midweek peak periods of each section. A video camera positioned at a high point at each site enabled the taping of a continuous section of about ½ km in length. The camera was equipped with a digital clock having an accuracy of 1/100 sec.



Intersection of Capacity speed vs. flow curve with Service Flow Rate vs. Speed curve defines Capacity, SF_E, and Speed at Capacity, S_C

FIGURE 2 Use of HCM method to determine capacity of a specific grade.

The method of data collection employed allowed the data to be used for analyses both of volumes and of density. To obtain speeds, pairs of tapeswitches were taped to the road surface at a known given distance. The tapeswitches were connected to a central processor that yielded the spot speeds of vehicles on the basis of the time intervals between the passing of the front wheels over the first and second tapes. Table 1 presents the pertinent traffic and geometric characteristics of the two sites. The ADT and the peak-hour volume are given for the two directions. The two sites were almost level (grades less than 1 percent), and there were no intersections at access points within 2 km (1.25 mi) of both ends of both sections. Figure 3 shows the cumulative probability density function of speeds for the two sites. These speeds were estimated for a range of hourly volumes.

The speed at Site 1 was considerably higher than that at Site 2, probably because of the better geometry and the lower average daily volume at Site 1. A comparison of passenger car and truck speeds at each site is shown in Figure 4. Relatively small differences occurred, most likely because of the low level of service and also the level alignment of each site.

DATA ANALYSIS

Accidents

Following the data collection and preliminary observations, an investigation was conducted of the annual number of accidents with casualties on the two road sections. Between 1981

TABLE 1 TRAFFIC AND GEOMETRIC DATA FOR TWO SITES

	Site 1	Site 2
ADT[1] (veh)	14850	18270
Hourly Volume[2] (vph)	1880	1490
Traffic Composition (%)		
Passenger cars:	53.0	62.8
vans:	23.7	19.8
trucks:	17.5	11.6
buses:	5.5	5.4
Speed[3] (km/h)	66.2	56.6
85th Percentile Speed (km/h)	74.0	65.0
Pavement Width (m.)	7.20	6.00
Shoulder Width (m.)	3.20	1.80
Terrain	Level	Level

1 - Average Daily Traffic, both ways
2 - Peak Hour Volume, both ways
3 - Average Spot Speed

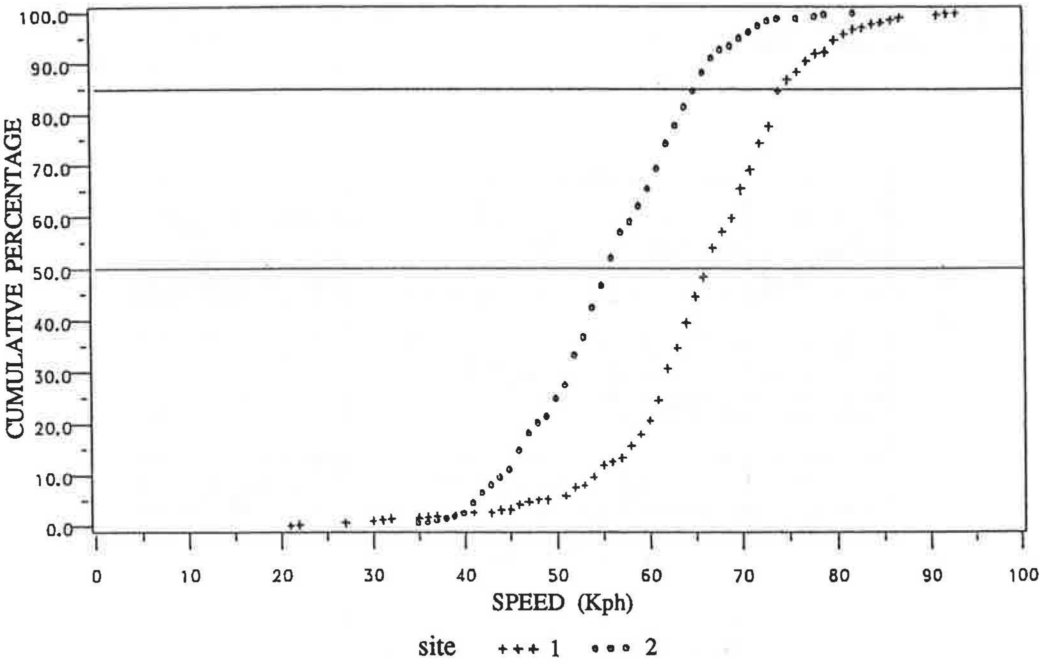


FIGURE 3 Cumulative density functions of speeds for two sites.

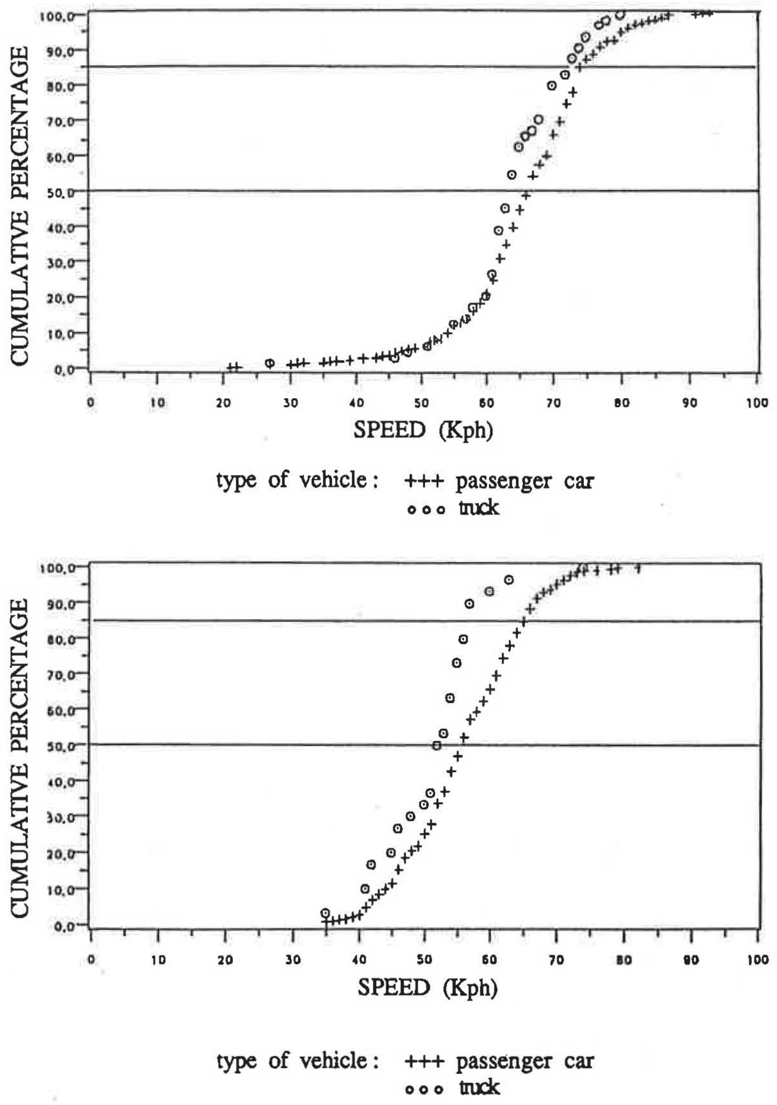


FIGURE 4 Cumulative density functions of speeds of cars and trucks for sites 1 and 2.

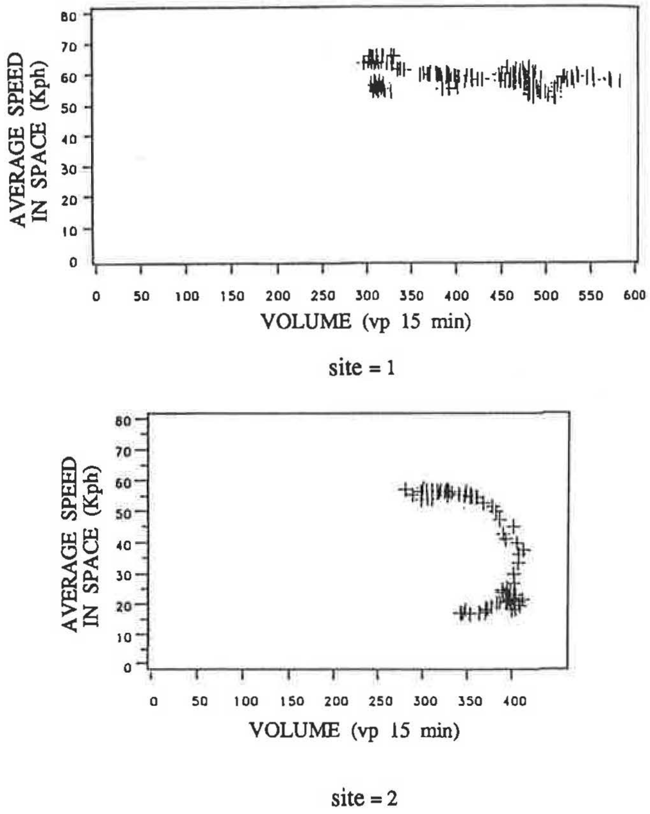
and 1985, the average number of accidents was 2.6 at Site 1 and 8.0 at Site 2. This difference was attributed to the improved geometry of Site 1, which included full, 7.20-m pavement with 3.0-m clear shoulders on each side.

Speed-Volume Relationship

The speed-volume relationship is often shown as a parabolic curve. The results obtained for this study are shown in Figure 5. The data at Site 2 result in a clear parabolic shape with a distinct peak, which indicates capacity. At Site 1, no peak was observed, and data were obtained only on the upper part of the curve.

These observations show two important findings. First, the flow regime at Site 2 is in the vicinity of capacity, above and below the maximum volume of about 420 vehicles per 15 min (both ways). At Site 1, the flow is below capacity, although some high volumes are obtained (more than 600 vehicles in 15 min). Second, the capacity zone at Site 2 occurs at a wide range of speeds, about 25 to 45 km/hr. This finding demonstrates that capacity on two-lane roads is a value that is not obtained at only one point on the speed-volume curve. It further shows that the flow data do not always behave according to the theoretical parabolic model. The relationship varies from one road to another and is dependent on the characteristics of each site.

Therefore, the hourly capacity volume is not a clear, single value. Furthermore, it is dependent on the time interval for which the data are collected. Naturally, a capacity based on 1-min observations will be higher than one based on 10- or 15-min time intervals. High capacity values may be obtained for short periods of time (i.e., short time intervals) because of flow characteristics: traffic on two-lane rural roads may flow with short headways (and therefore high volumes) only for short periods. As the time interval of the observation increases, the flow becomes inconsistent because average headways grow, as does the variation between headways. The



site = 1

site = 2

FIGURE 5 Speed-volume relationships for sites 1 and 2.

hourly volume, which is calculated from a longer time interval, therefore decreases.

Hourly capacity volumes are presented in Tables 2 and 3. The maximum one-way hourly volume shown in Table 2 for Site 1 is 1,980 veh/hr calculated from 1-min time intervals and 1,448 veh/hr calculated from 15-min intervals, or about a 37 percent difference. Table 3 presents the two-way maximum hourly volumes for the two directions together; the observed speeds obtained at capacity vary only slightly.

It is interesting to compare the observed capacity values with the HCM (*1*) values. To do so, the 15-min time intervals were used because the HCM analysis is based on observations at these intervals. The traffic directional split in this study was 60-40, and the percentage of trucks during the observations was approximately 14 percent. To convert the calculated, mixed two-way maximum hourly volume (2,320 veh/hr) to an equivalent hourly passenger car volume, a passenger car equivalent of 2 was used (level terrain). This calculation gave a maximum hourly volume of 2,645 veh/hr. The HCM value (directional split of 60-40, as shown on HCM page 8–6) is 2,650 passenger cars per hour, in both directions. The proximity of the two-way capacity values from this study and the HCM is evident.

TABLE 3 MAXIMUM FLOW VALUES FOR BOTH DIRECTIONS AND VARIOUS TIME INTERVALS

Time Interval (min.)	V_{max}	VPH	S (km/h)
1	48	2880	55
5	212	2544	55
10	392	2352	57
15	580	2320	58

V_{max} - Max. measured volume for the specified time interval

VPH - Calculated hourly volume, both directions

S - Observed space mean speed at capacity

The speeds, on the other hand, are not similar. The HCM gives a speed of about 72 km/hr (45 mph) as the speed at capacity (see HCM Figure 8-1). The average speed in this study was about 58 km/hr, and the speed at capacity was even lower (about 40 km/hr). It is concluded that, although the capacity values obtained by the HCM seem adequate, the speeds at which they occur are more sensitive to the geometric conditions of the roadway and may be lower than HCM-suggested values.

Speed-Density Relationship

Further analysis was conducted of the density data collected at Site 1. The speed-density curve shown in Figure 6 was based on average 15-min densities for a 1-km (0.62-mi) section. The densities varied widely between 10 vehicles per kilometer

TABLE 2 VOLUMES FOR VARIOUS INTERVALS AND CALCULATED HOURLY VOLUMES

Time Interval (min.)	SITE 1				SITE 2			
	1st Direction		2nd Direction		1st Direction		2nd Direction	
	V_{max}	VPH	V_{max}	VPH	V_{max}	VPH	V_{max}	VPH
1	30	1800	33	1980	23	1386	19	1140
5	132	1584	103	1236	96	1152	60	720
10	239	1434	195	1170	185	1110	105	630
15	362	1448	276	1104	276	1104	196	584

V_{max} - The maximum volume, as measured for a specified time interval

VPH - Calculated hourly volume

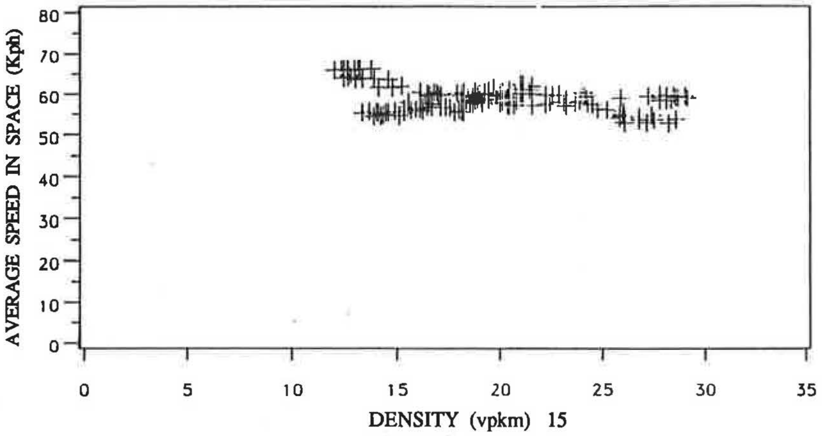


FIGURE 6 Speed-density relationship for Site 1.

(vpkm) and about 30 vpkm, which further indicates that traffic flow may be dense at certain times and light at other periods.

Volume-Density Relationship

The volume-density relationship for Site 1 is shown in Figure 7. There is a volume increase with 15-min interval observations as the density increases. The data spread is such that the resultant slope at high densities (about 25 vpkm) is smaller than that at low densities (about 15 vpkm). This finding indicates that a maximum volume was not reached at Site 1, as shown in Figure 5. Although the common volume-density relationship is parabolic, the peak volume could not be detected at high densities. The relationship between volume V in veh/hr and density D in vpkm can be expressed as a parabolic relationship as follows:

$$V = aD^2 + bD + c \tag{2}$$

in which a, b, and c are constants.

Calibration of this equation was performed on the volume-density data for various time intervals, and the results are presented in Table 4. The correlation coefficient increases with the increase in the time interval. This result is probably due to the reduction in variability of the data with the increase of the time period for which the data were collected.

CONCLUSIONS

Several conclusions can be drawn from this study. First, as expected, the capacity value is sensitive to the geometric char-

TABLE 4 CORRELATION COEFFICIENTS AND PARAMETERS OF VOLUME-DENSITY RELATIONSHIP

Time Interval (min.)	a	b	c	R^2
1	0.00	0.11	21.20	0.22
5	-0.06	7.16	24.04	0.62
10	-0.26	20.90	-27.16	0.76
15	-0.47	35.46	-87.86	0.83

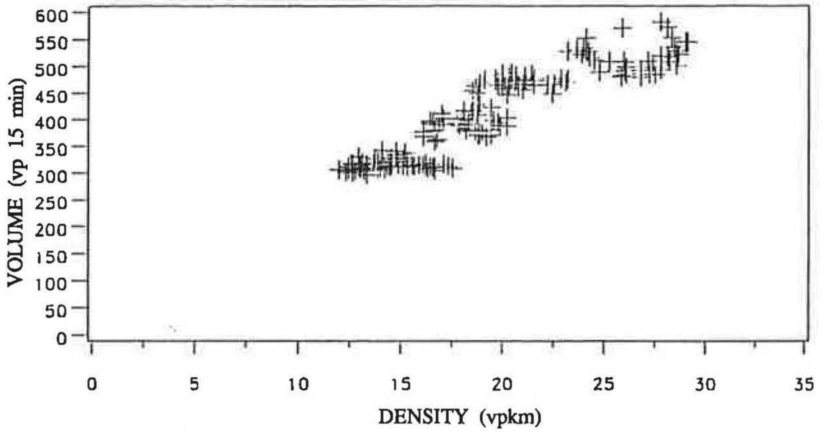


FIGURE 7 Volume-density relationship for Site 1.

acteristics of each site. Second, continuously changing traffic characteristics, such as the percentage of trucks, the directional split, the presence of a vehicle on the shoulder, or even the occurrence of a potential incident during dense flow conditions, have a great impact on road capacity. The flow regime, whether below or above HCM capacity, is very sensitive to the specific traffic conditions at any given time interval. Third, the value of capacity is highly dependent on the time interval for which data are collected and from which the capacity is calculated. A practical finding was that the capacity value, as calculated for 15-min time intervals, was similar to the HCM value. The speed at which this capacity occurs, however, was lower than the general speed recommended by the HCM.

Further research is suggested on two subjects. At the theoretical level, a calculation should be made of the proper time interval for an analysis of capacity. At the empirical level, investigation should be made of flow regimes and capacity values for various geometric conditions, terrains, and traffic characteristics.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. A. Polus, J. Craus, M. Livneh, and M. Gross. *Analysis of Capacity and Dense Flow Characteristics on Two-Lane Highways.* Research Report 88-125. Transportation Research Institute, Technion-Israel Institute of Technology, Haifa, July 1988.
3. *Special Report 87: Highway Capacity Manual.* HRB, National Research Council, Washington, D.C., 1965.
4. S. Yager. *Capacities for Two-Lane Highways.* Australian Road Research Board, Vermont South, Victoria, March 1983.
5. N. C. Duncan. A Further Look at Speed/Flow/Concentration Traffic Engineering and Control. *Traffic Engineering and Control*, Vol. 20, No. 10, Oct. 1979, pp. 482–483.
6. OECD Scientific Expert Group. *Traffic Capacity of Major Routes.* Road Transport Research, Organization for Economic Cooperation and Development, Paris, July 1983.

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

Characteristics and Level-of-Service Estimation of Traffic Flow on Two-Lane Rural Roads in Finland

MATTI PURSULA AND ÅSA ENBERG

An analysis of speed-flow relationships, time headways, and platooning of traffic on two-lane rural roads in Finland was conducted. In addition, the 1985 *Highway Capacity Manual* (HCM) level-of-service (LOS) calculation procedure was evaluated under Finnish road and traffic conditions. The speed-flow relationship on Finnish roads was found to be linear and not as steep as that given in the HCM. Simple loglinear regression models were estimated for the percentage of vehicles in platoons and mean platoon length as a function of a 15-min flow rate. A hyperbolic relationship was estimated between the platoon percentage and mean platoon length. The percentage of vehicles in platoons (headway less than 5 sec) was used as an approximation of the percent time delay (PTD) of the 1985 HCM. The analysis of the HCM LOS calculation procedure suggests that the service flow rates corresponding to the PTD values given in the HCM are clearly higher in Finland than in the United States. One reason for this finding may be that the speed-flow relationship and the adjustment factors of the HCM are both calibrated for U.S. road and traffic conditions.

A series of point measurements was carried out in the summer of 1984 to obtain information about the basic characteristics of traffic flow on two-lane rural roads in Finland. The aim of the measurements was to analyze the relationship between speed and flow and to determine the distribution of speeds and headways. After publication of the 1985 *Highway Capacity Manual* (HCM) (*1*), a further analysis was carried out to study the platoon characteristics and to evaluate the HCM level-of-service calculation procedure using these measurements.

The methods and results of these studies (*2–4*) are described in the following paragraphs.

STUDY MATERIAL

Study Locations

The results are based on data gathered at 10 points along the southern part of the Finnish trunk road network. The study locations were chosen to provide information from different kinds of roads. Roads with speed limits of 80 and 100 km/hr, roads with good and poor visibility, and roads of different widths and hilliness were included.

On the road sections where the measurement points were located (section length 2 to 3 km), the road width varied between 8.0 and 10.8 m, the pavement width between 7.2 and 9.6 m, and the lane width between 3.35 and 3.75 m. The hilliness index of the sections was 7 to 23 m/km and the curvature index 2 to 40 rad/km. The percentage of 460-m sight distance varied between 4 and 75 percent.

Measuring Techniques

The measurements were carried out using a traffic analyzer developed at the Laboratory of Transportation Engineering of Helsinki University of Technology. The equipment is able to measure and store the speed, time headway, and length of consecutive vehicles on the road. A pair of induction loops on both lanes were used as sensors.

At least two measurements were carried out in each location: one in the peak traffic period (mainly on Friday evenings) and one in normal weekday traffic. A total of about 5 hr of measurements were done at each point.

Basic Data Processing

The data were transferred to a mainframe computer and analyzed with special programs to give the mean speeds, flows, densities, and time headways at time intervals chosen for the analysis. The results presented are mainly based on 15-min intervals.

Most analyses were carried out both in the two directions of traffic separately (one-way traffic) and in the two directions together (two-way traffic). The highest 15-min flow rates measured were 1,500 to 1,600 vehicles per hour (vph) in the main direction and 1,900 to 2,000 vph in both directions together. The directional distribution of the traffic was about 50-50 on weekdays. On Friday evenings, the main direction carried about 70 to 80 percent of the traffic.

The percentage of heavy vehicles (measured vehicle length over 6 m) varied between 4 and 23 percent. In peak traffic, the share of heavy vehicles was about 5 to 8 percent.

SPEED-FLOW CURVES AND SPEED VARIATION

The relationship between speed and flow was found to be linear (see Figure 1). The average speed in one-way traffic decreased by 3 to 9 km/hr when the flow increased by 1,000 vph. The free speed was 89 to 92 km/hr on roads with 100-km/hr speed limit and 71 to 86 km/hr on 80-km/hr roads.

Laboratory of Transportation Engineering, Helsinki University of Technology, Rakentajanaukio 4A, SF-02150 Espoo, Finland.

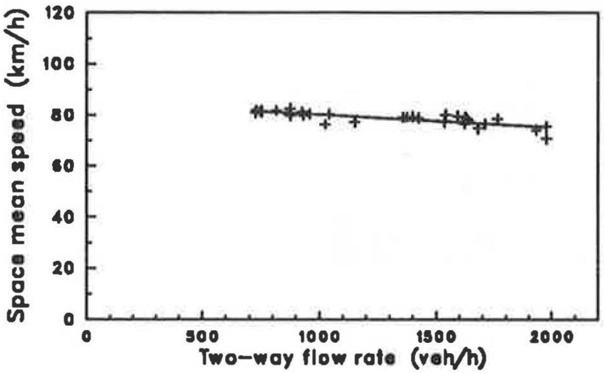


FIGURE 1 Example of measured speed-flow relationship for two-way traffic: Trunk Road 3, Riihimäki.

The speed-flow relationship was also studied in both directions together. The results were much the same. On roads with a speed limit of 100 km/hr, the mean free speed was about 87 to 92 km/hr, and on 80-km/hr roads about 75 to 85 km/hr. For every 1,000 vph flow increase, the speed decreased by 4 to 6 km/hr on the 100-km/hr roads and by 2 to 8 km/hr on the 80-km/hr roads.

The variation in mean speeds between different sites was considerable, as can be seen from Figure 2. However, no clear connection between the road width or other indicators of road standard and the speed values could be found.

In the 1985 HCM, under ideal conditions, the speed decreases by about 8 km/hr when flow increases by 1,000 vph. This decrease is larger than that found in most of the measuring points of this study (see Figure 2). According to the HCM, the decrease in speed is even faster when traffic and road conditions differ from the ideal ones. Hence, the speed-flow relationship in the HCM seems to be steeper than the one measured on Finnish two-lane roads.

The speed-flow relationship (one-way traffic) on long highway sections in Finland has been measured in other studies (*5*). The decrease of speed for a flow increase of 1,000 vph was found to be 6 to 10 km/hr, and the mean free speed was about 93 to 94 km/hr in 100-km/hr speed limit areas and about 87 to 88 km/hr in 80-km/hr speed limit areas.

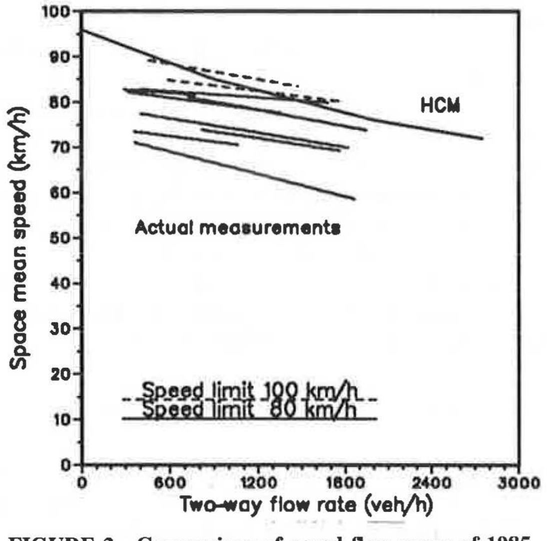


FIGURE 2 Comparison of speed-flow curve of 1985 HCM (ideal conditions) with actual measurements at 10 different locations on two-lane Finnish roads.

In general, the measured free speeds are tied to the traffic behavior of the time of the measurements. In Finland the free speed of all vehicles on 80- and 100-km/hr speed limit areas has risen about 1.2 km/hr from 1984 to 1988 (*6*).

The standard deviation of speed distributions varied between 5 and 12 km/hr. In the 100-km/hr speed limit areas, about 85 to 90 percent of the vehicles were driven with a speed below the limit. In 80-km/hr speed limit areas, there were some locations where only 50 percent of the drivers maintained a speed below the limit and others where practically all vehicles were driven with speeds under the limit. An example of the speed variation is shown in Figure 3.

The data gathered did not provide a direct opportunity to estimate the capacity values of the road at the measuring points because no breakdown conditions existed. The capacities estimated with flow-density models varied substantially. The general conclusion of the material was that the capacity in good road conditions is about 2,800 to 3,000 vph in both directions together. In other studies on long highway sections, an actual capacity of 1,400 to 1,500 vph in the main direction has been measured on two-lane trunk roads in Finland (*5*).

HEADWAY DISTRIBUTIONS

The time headway distributions of one-way traffic were analyzed graphically. That is, a number of distribution curves were drawn to determine the characteristics of the distributions, but no curve fitting was done. Figure 4 shows an example of these distributions. The mode of the curve is always between 1 and 2 sec. Also, the higher the volume, the higher the curve maximum.

The characteristics of the headway curves are quite similar to those found in international studies. The curves indicate that, at flow levels analyzed in this study, the headways do not follow the negative exponential distribution. Thus, traffic on a two-lane road is not free of interactions between vehicles, which is natural because of the passing problems on two-lane roads.

The usual variation area of the standard deviations of the headway distributions is shown in Figure 5. The variation of

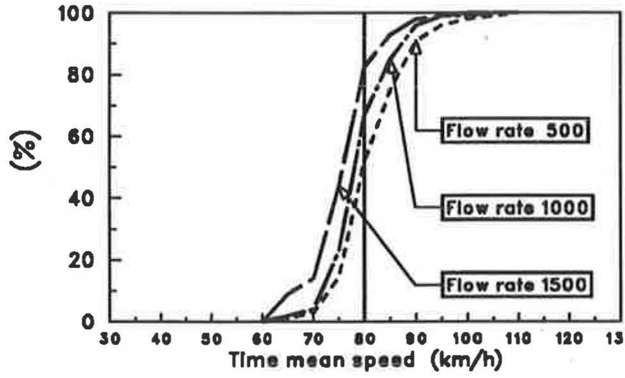


FIGURE 3 Example of cumulative speed distributions on a two-lane rural road: Trunk Road 3, Riihimäki.

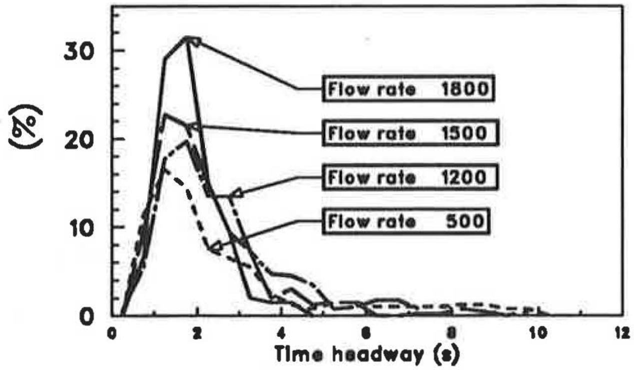


FIGURE 4 Example of measured time headway distribution curves for one-way traffic: Trunk Road 3, Riihimäki.

the measured headways is higher than that of a negative exponential distribution for one-way flows of up to 1,200 vph. This finding is an indication of the nature of the traffic on two-lane roads. A headway distribution with a standard deviation higher than that of the negative exponential distribution indicates that the traffic moves in long platoons with long headways between the platoons. When the standard deviation is smaller than that of the negative exponential distribution, the headways are more evenly distributed (*7*).

PLATOONING OF TRAFFIC

Principles of Analysis

The platoon criterion used was a 5-sec time headway between successive vehicles. The length of a platoon defined in this way is one vehicle shorter than the length of the corresponding vehicle cluster observed on the road. The first vehicle of the cluster is not in the platoon because its headway to the vehicle in front of it is longer than 5 sec.

The basic characteristics analyzed were the percentage of vehicles in platoons and the mean length of the platoons. The mean speeds of vehicles in platoons and those not in platoons were also analyzed, as well as the differences in vehicle composition of platoons compared with that of all traffic.

For the statistical analysis of the platoon percentage and platoon length, the base hypothesis was that the moving queues in one-way traffic follow the geometric distribution (*8*). From that assumption, a simple connection between the platoon percentage and the mean platoon length can be calculated:

$$E(Q) = 100/(100 - p) \quad (1)$$

where $E(Q)$ equals the mean platoon length (veh) and p equals percentage of vehicles in the platoons.

According to the theory, the mean length of the platoons is the same as the mean length of all vehicle clusters (including clusters of only one vehicle) on the road.

If the headway distribution is assumed to be negative exponential, the percentage of vehicles in the platoons and the mean platoon length are

$$p = 100 * [1 - \exp(-q * t)] \quad (2)$$

$$E(Q) = \exp(q * t) \quad (3)$$

where q equals traffic volume (veh/sec) and t equals time headway used as platoon criterion (5 sec in this study).

In logarithmic scale,

$$\ln[(100 - p)/100] = -q * t \quad (4)$$

$$\ln[E(Q)] = q * t \quad (5)$$

Thus, in one-way traffic with negative exponential headway distribution and geometric platoon length distribution, the

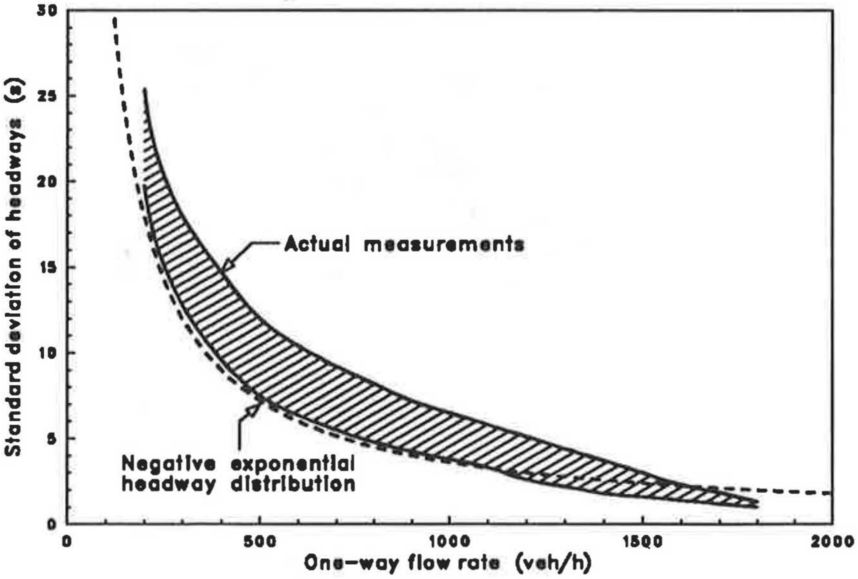


FIGURE 5 Variation area of standard deviations of headway distributions as a function of 15-min flow rate for one-way traffic.

natural logarithm of platoon length and the share of vehicles outside the platoons have the same numerical value but opposite sign. In two-way traffic, the formulas are somewhat more complex because the directional split of traffic must be taken into account.

The analysis of platoon percentages and platoon lengths was carried out in 15-min time intervals. In these time periods, the mean speeds and vchiclc composition of traffic in platoons and outside platoons were also registered. Platoons were not cut off at the borders of the time intervals, which means that the length of the interval was not always precisely 15 min.

The data were grouped on the basis of the measuring locations and analyzed in these groups. For some calculations, all of the data were used. The main mathematical tool was multiple linear regression, using logarithmic values of platoon percentage and platoon length. A basic assumption of loglinear relationships was used even though the analysis of headways showed that they do not follow the negative exponential distribution.

Results

The percentage of vehicles in the platoons and the mean platoon length follow a loglinear relationship quite well, as shown in Figure 6. In Figure 7, the actual relationships between traffic volume and platoon percentage are drawn. The figure shows clear differences between the measuring locations. These differences are partly because of road conditions and partly because of traffic composition.

In Figure 8, the corresponding curves of mean platoon length versus traffic volume are shown. Comparison of the measured curves and the curves of traffic with negative exponential headways in Figures 7 and 8 clearly shows that there are more

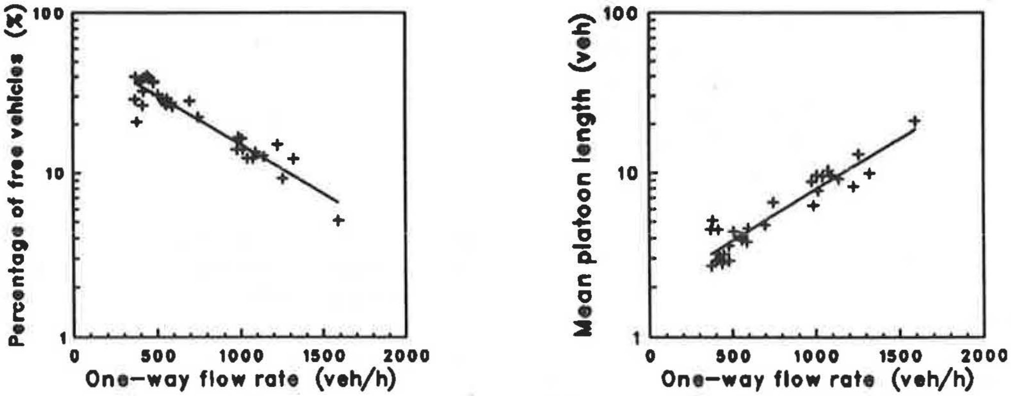


FIGURE 6 Example of loglinear relationship of platoon percentage and mean platoon length with flow rate: Trunk Road 3, Riihimäki.

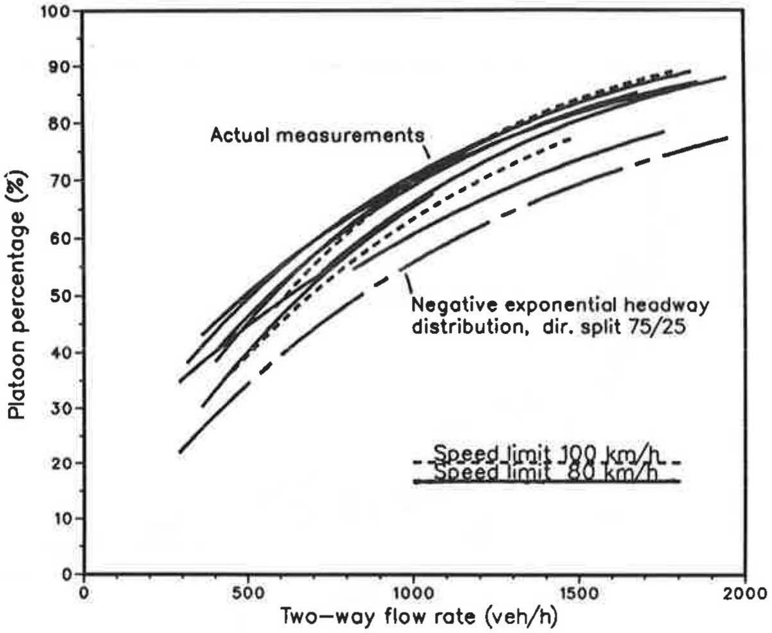


FIGURE 7 Relationship of platoon percentage and flow rate on Finnish two-lane roads.

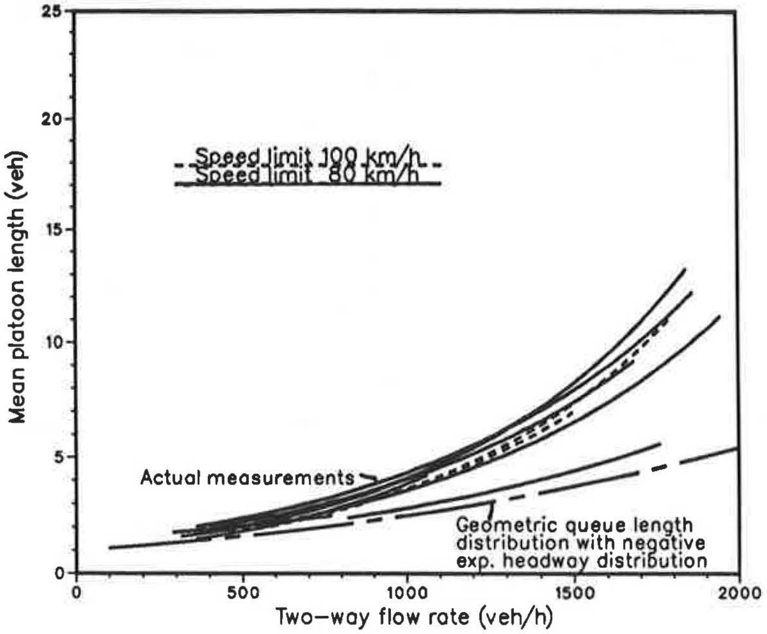


FIGURE 8 Relationship between mean platoon length and flow rate on Finnish two-lane roads.

and longer platoons on Finnish two-lane roads than the simple theory would suggest. According to the theory, a close relationship between platoon percentage and mean platoon length exists. Figure 9 shows the theoretical curve and the basic data of this relationship.

Figure 10 shows a summary of the findings. Using this figure, the average platoon percentage and the mean platoon length on Finnish two-lane roads can be approximated as a function of traffic volume. Two different platoon criteria are shown in the figure, namely, 5- and 3.5-sec time headways.

Figure 10 shows that the relationship between platoon percentage and platoon length is not dependent on the platoon criterion. This finding is in accordance with the theory presented. The simple regression equations of the basic relationships in Figure 10 (two-way traffic, 5-sec platoon criterion) are given in Equations 6–8.

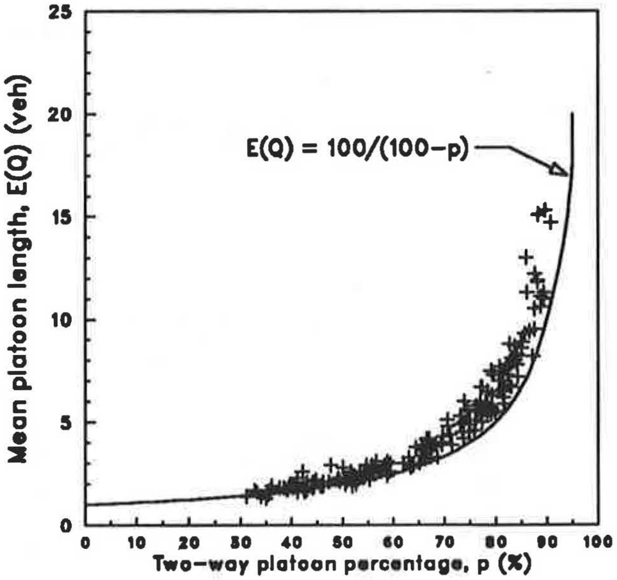


FIGURE 9 Basic data of the relationship of mean platoon length and platoon percentage.

Mean platoon length $E(Q)$ as a function of flow rate q:

$$\ln [E(Q)] = 0.056 + 0.00126 * q \qquad (6)$$
$$R^2 = 0.894$$

Platoon percentage as a function of flow rate q:

$$\ln (100 - p) = 4.57 - 0.0011 * q \qquad (7)$$
$$R^2 = 0.926$$

Mean platoon length $E(Q)$ as a function of platoon percentage p:

$$E(Q) = -0.57 + 141.8/(100 - p) \qquad (8)$$
$$R^2 = 0.952$$

Figure 11 shows an example of the difference between the mean speeds of free and platoon vehicles. In general the difference varied between 1 and 5 km/hr and was found to increase slightly when the traffic flow increased.

Figure 12 shows the result of an analysis of the effect of the opposite flow on the platoon lengths and platoon percentage when the flow in the analyzed direction is constant. The material was so limited that a more general analysis of this phenomenon could not be done. Nevertheless, the figure shows that the level of the opposite flow has a substantial effect on platooning.

The analysis indicated that heavy vehicles (vehicle length more than 6 m) are more often platoon leaders than their

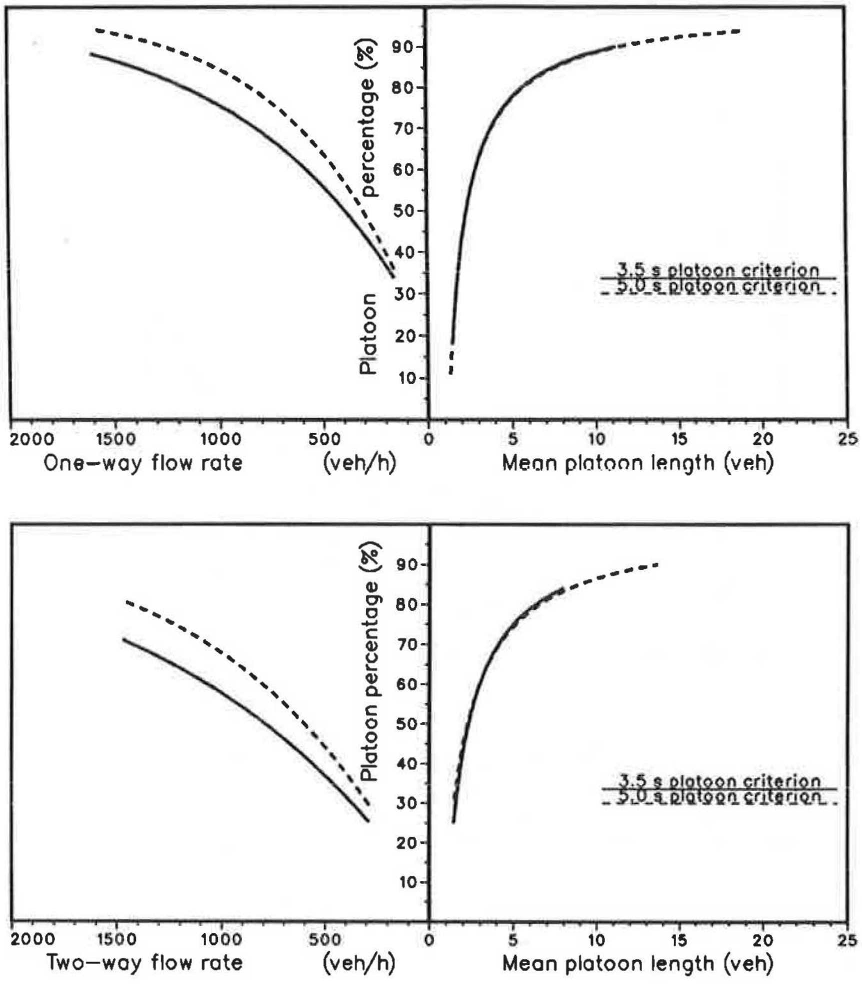


FIGURE 10 Relationship among flow rate, platoon percentage, and mean platoon length on two-lane Finnish roads.

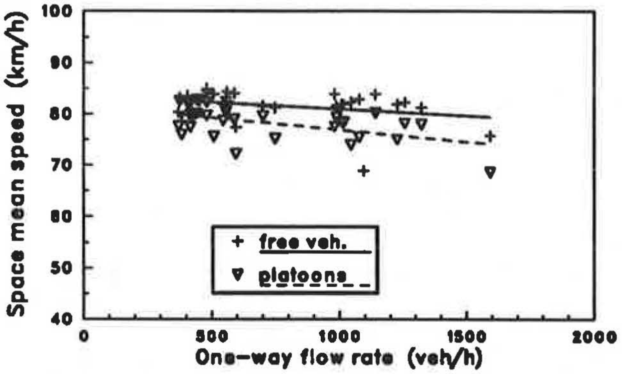


FIGURE 11 Example of speed-flow relationship of platoons and free vehicles: Trunk Road 3, Riihimäki.

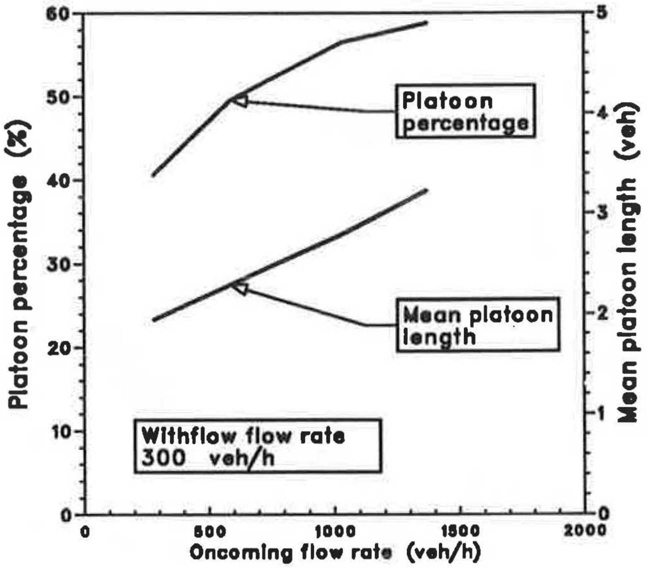


FIGURE 12 Effect of opposite flow rate on platoon percentage and mean platoon length.

share of the traffic would suggest. An example of this finding is shown in Figure 13.

LOS

In the 1985 HCM, the LOS of two-lane rural roads is calculated using the percent time delay (PTD), which, according to the manual, can be approximated with the percentage of vehicles with headways less than 5 sec (*1*). Thus, the platoon percentage of this study can be used as a PTD estimate.

In order to make a comparison between the PTD values and actual traffic volumes on Finnish roads and the 1985 HCM, the following procedure was used. In every measuring location, the maximum service flow rates were calculated using the equation given in the 1985 HCM:

$$SF_i = 2{,}800 * (v/c)_i * f_d * f_w * f_{HV} \quad (9)$$

where

SF_i = total service flow rate in both directions for prevailing roadway and traffic conditions, for LOS *i* (vph);
$(v/c)_i$ = ratio of flow rate to ideal capacity for LOS *i*;
f_d = adjustment factor for directional distribution of traffic;
f_w = adjustment factor for narrow lanes and restricted shoulder width; and
f_{HV} = adjustment factor for the presence of heavy vehicles in the traffic stream.

The maximum service flow rates were then plotted against the PTD values given in the HCM as LOS criteria. The results are shown with crosses in Figure 14. The variation of service flow rates corresponding to a given PTD value is substantial because of the variations in road and traffic conditions of the measuring locations.

The curves in Figure 14 indicate the mean relationships between flow rate and PTD value at those measuring points where the speed criterion of the LOS was met. The figure shows that the actual measured average service flow rates on Finnish two-lane rural roads are generally higher than those calculated according to the 1985 HCM or, vice versa, the PTD values corresponding to a certain flow level are higher in the HCM. However, in good road conditions, the differences are small.

A different PTD criterion (namely, 3.5-sec headways) with slightly changed *v*/*c* ratios has also been suggested in the literature for LOS analysis (*9*). The calculations described previously were also carried out with these criteria. The results (see Figure 14) are about the same even though the differences between the Finnish and the American values are smaller.

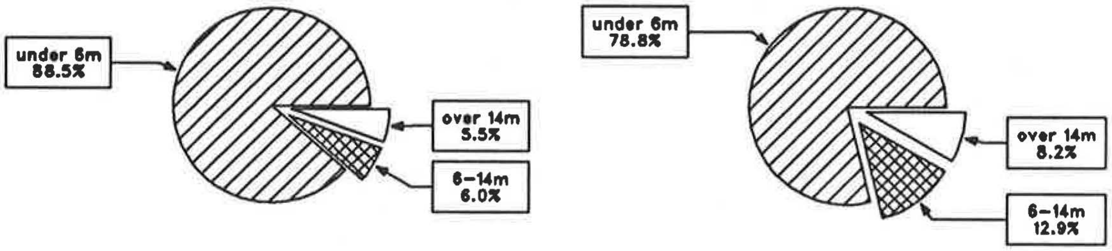


FIGURE 13 Example of vehicle composition of whole traffic (*left*) and that of platoons (*right*).

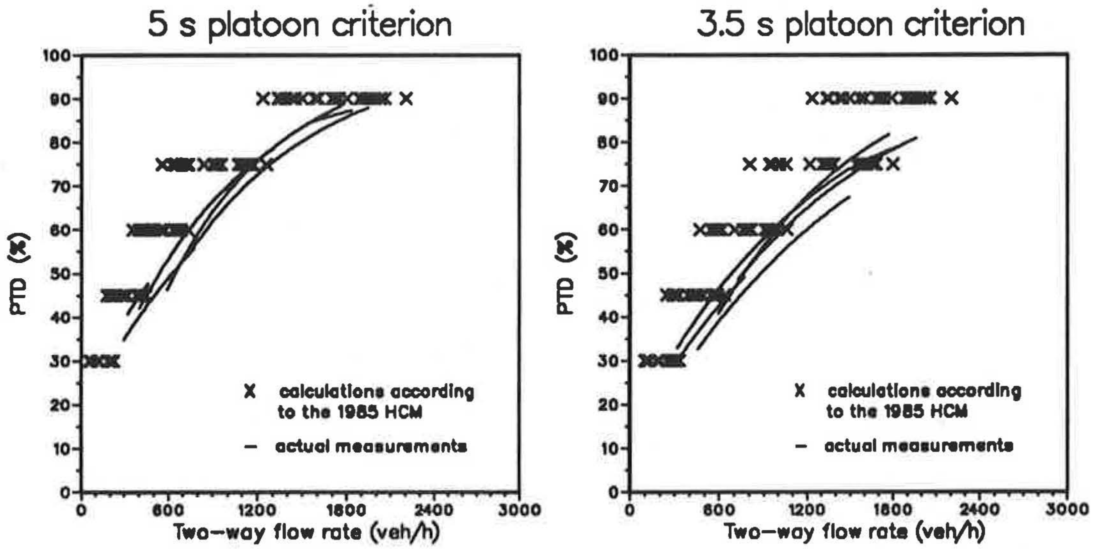


FIGURE 14 Calculated service flow rates (HCM) and measured relationships of flow rate and PTD on two-lane rural roads in Finland.

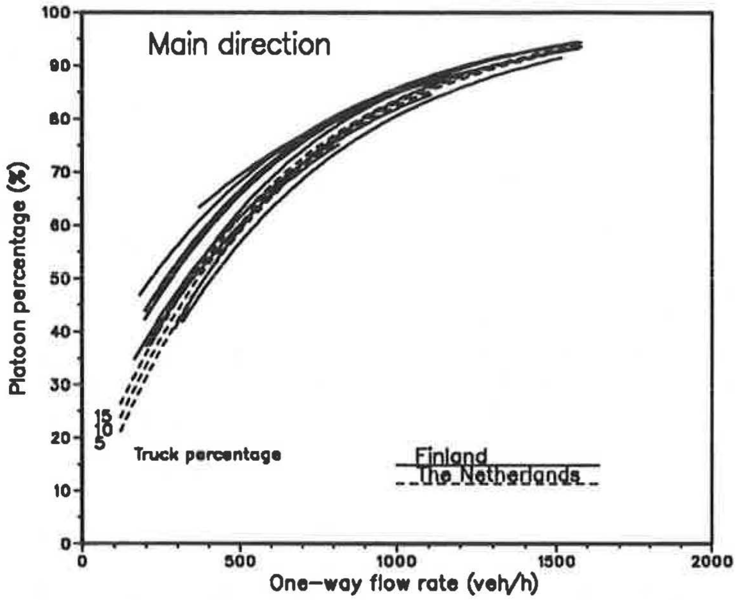


FIGURE 15 Comparison of measured Finnish and Dutch platoon percentages as a function of flow rate.

DISCUSSION OF RESULTS

The studies described are in accordance with U.S. and other international findings on the flatness and linearity of speed-flow curves and on the magnitude of the capacity of two-lane rural roads. The decrease in speed when flow increases seems to be faster in the HCM than in actual traffic on Finnish roads.

In the study, it was assumed that the platoon length and headway distributions were simple and could be represented by geometric and negative exponential distributions. A closer look at the headway distributions and statistical studies on platoon length distributions clearly indicated that the assumptions are not generally valid. Nevertheless, the simple theory provided simple relationships that turned out to be effective in analyzing the basic relationships among the platoon percentage, mean platoon length, and flow rate.

Platooning in Finland seems to be similar to that in the Netherlands. In Figure 15, the Finnish results of platoon percentage are compared with a Dutch model (*10*).

In the analysis of the level of service, the percentage of vehicles with headways less than 5 sec was used as an approximation for the PTD. This procedure is in accordance with the HCM, but it is possible that the differences found between the Finnish traffic and the HCM are an indication of the differences in real PTD values and the approximation used. An analysis of this difference was not possible in this study, and the conclusions drawn are based on the assumption that the approximation is not seriously biased.

Under Finnish road and traffic conditions, the flow levels corresponding to a given PTD value are often considerably higher than those in the HCM. The difference is small on good roads. The Dutch study, as opposed to the Finnish one, concludes that there is a good similarity between the Dutch results and the 1985 HCM (*10*). The reason for the different conclusion may be the different road conditions. The Dutch measurements were carried out in good road conditions, whereas the Finnish results were also near the values given in the HCM.

The similarity of results in good road conditions implies that, in other conditions, the adjustment factors of the HCM calculations may be a reason for the differences. The proper values of the adjustment factors in Finland probably differ from those given in the HCM. So, changing the adjustment factors is one way to adjust the measured service flow rates with the PTD values. Another possible way is to change the critical *v/c* ratios that are closely related to the steepness of the speed-flow curve. So far, no changes to any HCM values have been made. More material is needed to correct the procedure properly for Finnish conditions.

ACKNOWLEDGMENTS

The research described was done at The Helsinki University of Technology, Laboratory of Transportation Engineering, with the sponsorship of the Finnish National Roads Administration. The valuable comments and suggestions of Sulevi Lyly during the preparation of this paper are greatly appreciated.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. M. Pursula and H. Sainio. *Basic Characteristics of Traffic Flow on Two-Lane Rural Roads in Finland* (in Finnish). TVH 741 824, Finnish National Roads Administration, Helsinki, 1985.
3. A. Enberg and M. Pursula. *Platooning on Two-Lane Rural Roads in Finland* (in Finnish). Finnish National Roads Administration, Helsinki, 1987.
4. A. Enberg and M. Pursula. Platooning and Level of Service on

Two-Lane Roads (in Finnish). In *Tie ja Liikenne 88*. Suomen Tieyhdistys, Helsinki, 1988.

5. H. Summala, J. Hietamäki, A. Lehikoinen, K. Töttölä, and J. Vierimaa. *Highway Capacity and Travel Speeds on Long Distances* (in Finnish). Report 7:1986, Traffic Research Unit, University of Helsinki, Helsinki, 1986.
6. *The Speeds of Motor Vehicles on Trunk Roads in 1988* (in Finnish). TVH 741 836-88, Finnish National Roads Administration, Helsinki, 1989.
7. Kapacitetsutredning. *Litteraturstudie och Analys*. TV 118, Statens Vägverk, Stockholm, 1983.
8. D. R. Drew. *Traffic Flow Theory and Control*. McGraw-Hill, New York, 1968.
9. D. L. Guell and M. R. Virkler. Capacity Analysis of Two-Lane Highways. Presented at 67th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1988.
10. H. Botma. Traffic Operation on Busy Two-Lane Rural Roads in The Netherlands. Presented at 65th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1986.

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

Saturation Flow: Do We Speak the Same Language?

S. Teply and A. M. Jones*

Saturation flow is the basis for the determination of traffic signal timings and for the evaluation of intersection performance. Most current analysis and design methods are based on some form of the critical lane technique, in which the critical lane, or group of movements, is determined from the relationship between the volume it carries and its saturation flow. The general descriptions of saturation flow presented in the *Highway Capacity Manual*, the *Canadian Capacity Guide for Signalized Intersections*, and an Australian Road Research Board special report on traffic signal capacity and timing analysis are more or less identical for similar, close-to-ideal intersection conditions. Nevertheless, the definitions implied by the recommended survey techniques are significantly different. Moreover, saturation flow values reported by some researchers in the past have been based on measurement methods that differ from all three documents. The differences in the meaning of saturation flow, as understood in different regions and by different authors, are discussed. The implications are illustrated using the three documents and a set of 1989 Edmonton surveys as examples. It is found that values reported in the literature may not be directly comparable and that saturation flows measured with the assistance of one document may not be compatible with calculation techniques taken from another.

Saturation flow is the basis for the determination of traffic signal timings and for the evaluation of intersection performance. Most current analysis and design methods are based on some form of the critical lane technique, in which the critical lane, or group of movements, is determined from the relationship between the volume it carries and its saturation flow.

A small change in the saturation flow value may result in a relatively large change in the calculated cycle time and the duration of the necessary green intervals. The impact becomes especially critical in situations in which volumes are just below or exceed capacity.

Because saturation flow represents a major factor in the determination of measures of effectiveness used in traffic engineering practice, its value also influences the performance evaluation of signalized intersections. Almost all computer programs for design and analysis of individual intersections, signal progression, or network operations use saturation flow as one of the principal input parameters. The accuracy of these programs is usually sensitive to saturation flow estimates.

Most documents for the design and analysis of signalized intersections, such as the *Highway Capacity Manual* (HCM) (*1*), *Canadian Capacity Guide for Signalized Intersections* (CCG) (*2*), or Australian Road Research Board (ARRB) Report 123 (*3*), recommend the use of measured saturation flows, rather than the default values provided in these documents or in many computer programs.

Nevertheless, although the general descriptions of satura tion flow are more or less identical for similar, close-to-ideal intersection conditions, the definitions implied by the recommended survey techniques are not. To make the issue more complicated, saturation flow values reported by some researchers in the past have been based on measurement methods that differ from all three documents.

The differences in the meaning of saturation flow, as understood in different regions and by different authors, are discussed. The implications are illustrated using the three documents and a set of 1989 Edmonton surveys as examples. The objective is to make researchers and practitioners aware that values reported in the literature may not be directly comparable and that saturation flows measured with the assistance of one document may not be compatible with calculation techniques taken from another.

CONCEPT OF SATURATION FLOW

The HCM (*1*) describes the saturation flow rate as the flow, in vehicles per hour per lane, that can be accommodated by the lane assuming that the green phase is always available to the approach.

The CCG (*2*) defines saturation flow as the rate of queue discharge from the stop line of an approach lane, expressed in passenger-car units per hour of green (pcu/hr green).

ARRB Report 123 (*3*) defines saturation flow as the maximum constant departure rate from the queue during the green period, expressed in through-car units per hour (tcu/hr).

These definitions do not mean that there is a continuous hour of green, but imply the usual stopping and moving operation for the normally used range of cycle times and green intervals; thus, saturation flow reflects the uniform service rate used in most applications of queueing theory for the problem of intersection capacity.

All three general definitions are based on the conventional graphical representation of saturation flow shown in Figure 1. The solid line in the figure shows the traditional concept, which assumes that the discharge rate from a stop line remains constant after some initial period. Canadian experience (*2*) has indicated that the discharge begins to decline after 30 to 50 sec of green time. Both interpretations are based on demand at or above capacity.

This traditional concept assumes that, after an initial hesitation immediately following the beginning of the green interval, traffic discharges at a constant rate (the saturation flow

*Winners of the 1991 D. Grant Mickle Award for outstanding paper in the field of operation, safety, and maintenance of transportation facilities.

Department of Civil Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2G7.

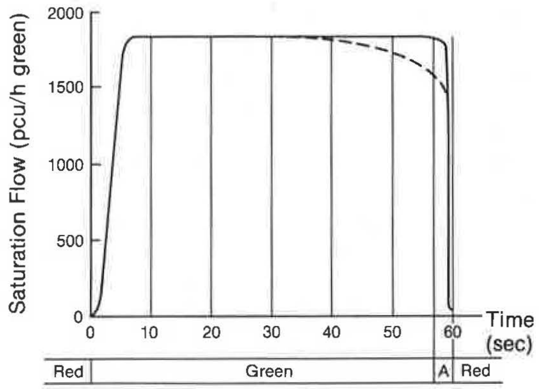


FIGURE 1 Saturation flow concept.

rate) until the queue is exhausted or until shortly after the beginning of amber, when a sharp drop in the flow occurs. The departure rate is lower during the first few seconds, while vehicles accelerate to normal running speed, and after the end of the green interval, as the flow of vehicles declines.

The HCM, CCG, and ARRB report agree on the great variability in saturation flows caused by different urban conditions and various geometric and traffic configurations. Factors used to modify the saturation flow of a given lane or approach generally include the urban environment; local driver behavior; lane width; turning radius; gradient; pedestrian interference; parking or transit interference; interaction with priority, opposing or adjacent flows; and limited queueing or discharge space. The CCG also allows the user to consider weather conditions and the duration of the green interval.

There is a major difference among the three documents in their treatment of traffic composition (i.e., cars, trucks, and buses). The CCG notes that it is convenient to represent all flows as homogeneous entities, therefore suggesting conversion of all volumes of individual vehicle categories to passenger-car units. The values of pcu equivalents have been determined from surveys in several Canadian cities, both in summer and in winter conditions, using a least squares optimization technique (*4*).

Both the HCM and the ARRB report consider the basic saturation flow in vehicles/hr green, applying a traffic composition adjustment along with the adjustment factors based on geometric and other traffic conditions. All three documents include calculation procedures involving the special effect of turning vehicles on the operation of single lanes that carry different movements (shared lanes).

Although there are many similarities in the saturation flow adjustments recommended in these three documents, actual procedures differ. The Canadian concept of the so-called "basic" (or ideal) saturation flow has been adopted for use here. Basic saturation flow is defined as the number of pcu that can discharge across the stop line of an intersection approach lane 3.0 to 3.5 m wide, with no slope or curves. Traffic is moving straight ahead without any additional traffic friction (i.e., no bus stops, pedestrians, parking, or other limiting factors are present), under ideal weather conditions and during an optimum duration of the green interval.

For consistency in this comparison, traffic composition adjustments have been applied to all measured results according to the following rules:

- *HCM.* The percentage of heavy vehicles (defined as vehicles with more than four tires on the road) was derived from the surveyed data, and the corresponding heavy vehicle adjustment factor (f_{HV}) was taken from Table 9–6 of the HCM. The basic saturation flow was then divided by this factor to increase the flow, accounting for the traffic composition in a manner similar to the pcu conversion in the CCG.
- *CCG.* All surveyed vehicles were converted to pcu before processing of the data, using the pcu equivalents suggested by the CCG.
- *ARRB Report 123.* The traffic composition factor (f_c) is a weighted average determined by the proportions of various vehicle types in combination with turning movements. Because only straight-through flows were surveyed, the through-car-unit equivalent values were limited to 1 for cars and 2 for all heavy vehicles. Heavy vehicles are defined as vehicles having more than two axles or having dual tires on the rear axle. The average tcu/vehicle was multiplied by the basic saturation flow to convert the flow to tcu/hr green. Naturally, for the basic (i.e., straight-through) saturation flows, pcu and tcu are directly comparable.

METHODS OF MEASURING SATURATION FLOW

Much of the background work on saturation flow took place at the end of the 1950s and in the 1960s, although the pioneering work by Greenshields (*5*) started much earlier.

One of the major problems with the body of literature on saturation flow is that many authors do not report the details of their survey techniques. As a result, the transferability of the values is questionable. Even when potential users are aware of the range of possible fluctuations in saturation flows, these flows are difficult to compare. The following section attempts to illustrate this problem.

It is assumed that there are no gross mistakes, which are sometimes found in practical applications. A typical example of this category when applying the HCM method is a wrong determination of the average headway by starting the time count at the passage of the fifth vehicle (the HCM clearly recommends use of the time of the fourth vehicle). The error can be compounded by dividing the duration of the surveyed period by the number of vehicles and not by the number of gaps. Such large mistakes, however, can easily be spotted because they usually result in saturation flow values well over 2,000 cars or vehicles/hr green/lane. Although such values are possible, they are unusual.

In essence, there are two major categories of saturation flow surveys, as follows.

The first group of techniques is based on the successive times (not necessarily true headways) of vehicle discharge at a specified reference line. Stop lines, nearside crosswalk boundaries, nearside intersection boundaries, or other nearside points (such as the cross section at which a nearside signal is located) have been used to that end, although farside intersection boundaries or farside crosswalk lines have also been used. Vehicles have been considered discharged when their front bumpers, front wheels, rear wheels, or rear bumpers have passed the reference line. As can be seen, a large number of possible combinations of reference lines and vehicle discharge criteria exists.

The second group of techniques, which count the number of vehicles passing the reference line during short portions of the green interval, is best represented by one of the first reports on the subject, published by the (then) Road Research Laboratory (RRL) and titled *A Method for Measuring Saturation Flow at Signalized Intersections* (*6*). In principle, the CCG also applies this basic technique.

Nevertheless, although both the RRL and CCG methods designate the stop line as the reference line, there is a difference in the identification of the point in time at which a vehicle is considered discharged. In the RRL report, discharge is given by the moment when the vehicle fully clears the stop line; the report recommends the use of the time at which the rear wheels cross the stop line. The CCG uses the passage of the front bumper over the stop line as the time of discharge, primarily because it is consistent with the usual definition of a headway (time differential between the passage of front parts of consecutive vehicles over a fixed cross section).

The HCM and ARRB report are, in fact, combinations of both basic groups of techniques because they are based on the determination of the average headways during a specifically defined portion of the green interval. This portion starts with the passage of the fourth vehicle in the HCM method and after 10 sec of green in the ARRB technique. The CCG survey method includes the entire initial period of flow.

The HCM, CCG, and ARRB report also demonstrate the different preferences in the identification of the reference cross-section line and the part of the vehicle that is considered representative of its discharge.

As an additional consideration, Branston and Van Zuylen (*7*) reported on a comparison of asynchronous and synchronous counting periods as defined by Haight (*8*), applied to saturation flow measurement. Although the synchronous technique starts and terminates observations at the instant of a vehicle departure (however defined), as in the HCM, the asynchronous method starts and terminates at arbitrary points in time, as in the CCG or ARRB report. Branston and Van Zuylen found some advantages of the synchronous method, although both techniques seem to be consistent if the data are manipulated correctly.

Greenshields (*5*) pioneered the headway method. He measured headways of passenger cars as their front parts passed the line connecting the nearside intersection curbs and found that, after the passage of the fifth vehicle, the headways stabilized at 2.1 sec. This value corresponds to about 1,715 vehicles per hour of green per lane.

Headway-based methods have been popular in Continental Europe, although the definitions of the reference line, and the part of the vehicle that defines discharge, vary (*7,9–16*). The use of the intersection exit as a reference line is less popular but has recently been used in Canada in a study of the effects of winter maintenance on traffic flow (*17*).

Many current saturation flow surveys still use the 1963 RRL method (*4,6,18–21*). Axhausen et al. (*16*) apparently used it to supplement headway measurements.

Although most of the British work consistently applies the 1963 method (which uses the passage of the rear axle for time reference), recent work by the British Transport and Road Research Laboratory (*22*) on the values of pcu equivalents defined the headways as "the time elapsed between the front of a vehicle crossing the stopline and the front of the following vehicle crossing the stopline."

Some of the differences in the headway values and their sequence that are caused by the various reference points in space and time can be explained with the assistance of Figures 2 and 3. Figure 2 shows, in a time-space diagram, a perfectly

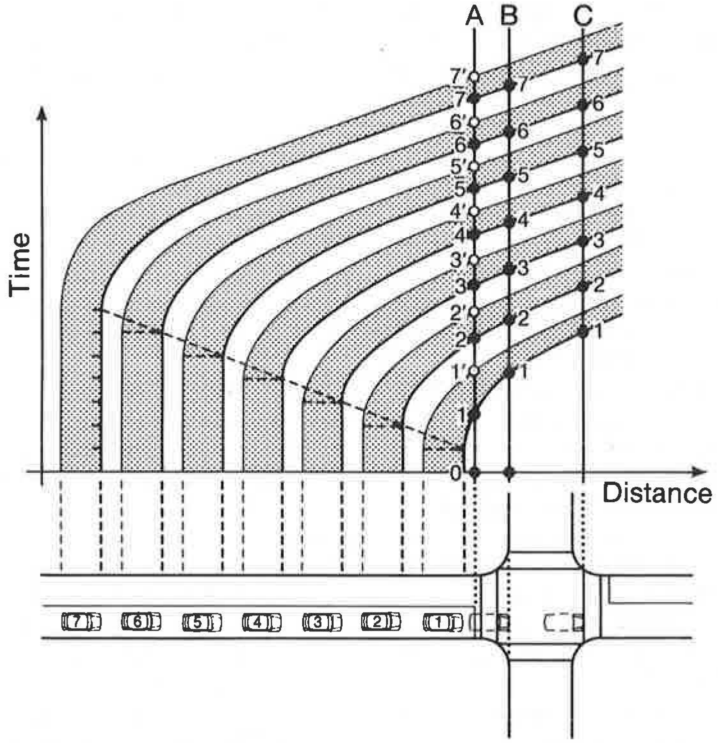


FIGURE 2 Perfectly regular discharge from standing queue, with uniform acceleration.

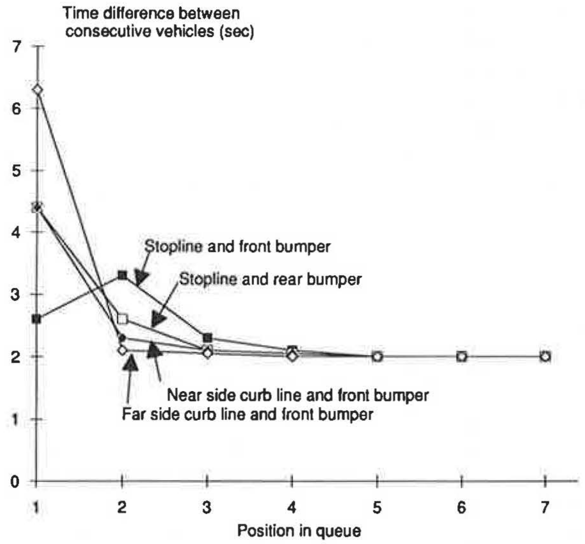


FIGURE 3 Variation in time difference between consecutive vehicles from Figure 2 for combinations of reference point and vehicle discharge criteria.

regular discharge of passenger cars from a standing queue with all cars accelerating uniformly at the same rate from the stopped position to a constant speed. Three reference lines are identified in the figure, which allows for the following combinations:

A. Stop line + front bumper,
A′. Stop line + rear bumper,
B. Nearside curb line + front bumper, and
C. Farside curb line + front bumper.

The trajectories of the vehicles intersect the vertical lines, which identify the reference lines in space. Figure 3 shows that the headways (or their rear-end equivalents) for the leading four or five cars vary not only in absolute values but also in the shape of the graph. The two basic shapes are easily recognizable, as follows:

1. The greatest time difference represents the time needed by the first vehicle, and
2. The greatest time difference occurs for the second vehicle.

The first case is typical for the Greenshields intervals; the second for many Continental European studies. The diagram attempts to illustrate that what is sometimes interpreted as a regional difference may merely be a result of a different measurement method. Naturally, the practical problem is more complex because of such factors as different vehicle lengths, rates of acceleration, acceleration noise of individual vehicles, and, consequently, the increased randomness that occurs at the far end of the intersection rather than at the stop line.

Some of the true regional factors that influence the derivation of saturation flows depend on traffic engineering practices and legal requirements. For example, the period immediately preceding the beginning of the green interval is identified in some countries by a short display of the amber signal in addition to the red signal. This practice may have a significant impact on the discharge during the first part of green (especially when motorcycles are present). Another factor that increases the international variability of flow, and influences the practice of defining the reference line and discharge, is the standard application of nearside or farside signals, or a combination of both.

An important factor related to the use of the surveyed data on headways or saturation flows is the well-known concept of the effective green interval and how this concept is introduced in design and analytical practice. Theoretically, it can either mitigate or increase the extent of inconsistencies caused by the application of a survey method that has not been meant for the calculation procedure at hand. Figures 2 and 3 illustrate the problem: the initial lag (start-up lost time) is significantly longer for methods that use reference points downstream of the stop line and rear parts of the vehicle. Figure 1 shows that such an increase may not be fully compensated for by the time gained after the end of the green interval.

COMPARATIVE SURVEYS

The main objective of the comparative research described here was to examine the relationship among the three saturation flow survey methods as described in the CCG, HCM, and ARRB report. By comparing the three methods, some insight might be gained as to the reliability of the surveyed results for further use in design or analysis computations. In the course of this study, variations in survey techniques used by several other researchers were also identified.

A total of 10 saturation flow surveys were conducted during the summer of 1989 in Edmonton, Alberta, Canada. More than 4,800 vehicles were observed during 325 signal cycles. The survey locations were as follows (see Figure 4):

1. St. Albert Trail/137 Avenue,
2. St. Albert Trail/Yellowhead Trail,
3. 127 Street/Yellowhead Trail,
4. 97 Street/137 Avenue,

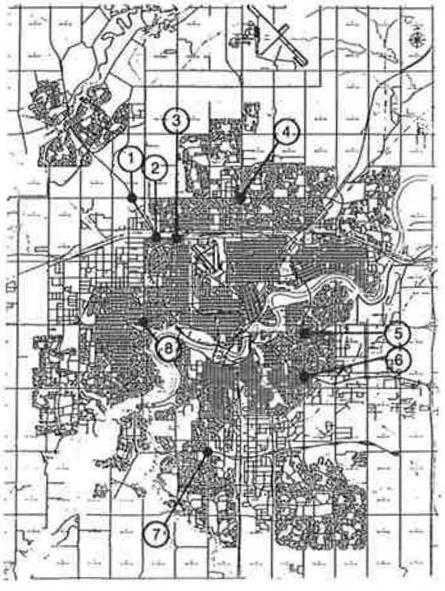


FIGURE 4 Survey locations within city of Edmonton.

5. 75 Street/98 Avenue,
6. 75 Street/Whyte Avenue,
7. Whitemud Drive/111 Street, and
8. 142 Street/Stony Plain Road.

These locations were carefully selected to conform with the general definition of basic saturation flow as designated by the CCG. The High Geometric Standard/Low Activity intersection approach category was chosen, meaning primarily divided arterial roadways with low activity levels for pedestrian and other adjacent land uses. This designation corresponds to ARRB Class A: ideal or nearly ideal conditions for free movement of vehicles (both approach and exit sides), good visibility, few pedestrians, and almost no interference because of loading and unloading of goods vehicles or parking turnover. The selected sites were also chosen to give the observer an unobstructed view of the surveyed lane and its stop line, and, hence, to maximize the accuracy of the results. The location shown in Figure 5 represents a typical intersection representative of basic saturation flow in Edmonton. In the figure, Site 1 was the observed lane for Survey 1 and Site 2 was the observed lane for Surveys 2 and 10.

Each survey consisted of 30 valid cycles with adequately saturated portions of the green interval of longer than 20 sec. For this reason, only the busier morning and afternoon peak periods were used. Several surveyed cycles had to be discarded due to unusual circumstances, such as the presence of emergency vehicles or stalled vehicles.

The surveys were conducted using a cassette recorder. Events in the observed lane were noted as they occurred (i.e., the beginning of the green interval, the length of the queue, the passage of the front bumper and rear axle of each passing vehicle over the reference point, the end of saturated flow, and the end of the green interval). Although the recording process required a great deal of concentration, no major problems were encountered with the ability to correctly identify the series of events.

The major benefit of using the same sample for all three techniques—simultaneously recording all data necessary for each of them—was that each of the methods could be applied to the same set of events. As a result, a direct and reliable comparison of the results was possible. In addition, if there were any inconsistencies due to the observer's perception of an event, they were equally reflected in all three methods.

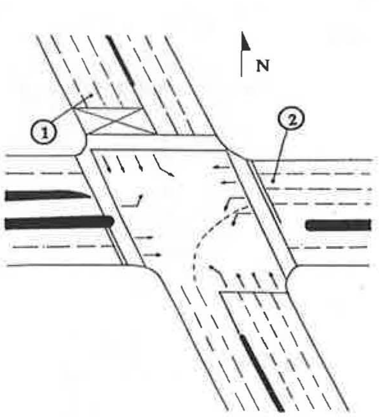


FIGURE 5 Typical survey site (Location 1): St. Albert Trail and 137 Avenue.

The survey methods used are described in detail in the source documents (CCG, HCM, and ARRB Report 123). Raw data were transcribed to field data sheets for each survey in the formats suggested by the individual manuals. The following section briefly outlines the three procedures.

CCG

The CCG survey method is designed for application to a single observed lane (similar to the HCM and ARRB report). A 15-sec time check just before and after each survey was recorded to calibrate the speed of the cassette recorder. The following vehicle categories were used: passenger cars, vans and pickup trucks, single-unit trucks, multiunit trucks, buses, and motorcycles.

Vehicles were considered to be discharged when their front bumpers had passed the reference point (stop line). The end of the platoon also included vehicles that joined the queue and proceeded as part of the lineup after the start of green. Because of the short duration of some green intervals, a 5-sec time slice was used in the data transcription. A comparison with the use of a 10-sec time slice showed that the choice of duration made little difference to the final basic saturation flow.

Saturation flow was computed as the average number of pcu per valid (fully saturated) time slice, multiplied by 3,600 and divided by the duration of the slice.

The results are illustrated in two formats (see Figure 6):

1. As a conventional saturation flow graph, with the flow for each 5-sec time slice plotted directly and the average flow calculated from all the values; and
2. As a cumulative average graph, with the average flows calculated from the beginning of green to the end of the given time slice (i.e., 0 to 5, 0 to 10, 0 to 15 sec, etc.). The curve climbs asymptotically to the eventual constant saturation flow rate as the effect of the initial loss becomes less pronounced for longer green intervals. Thus, the saturation flow is the asymptotic value to which the graph converges or, in the event that the rate starts declining, the peak of the graph, usually occurring at 30 to 40 sec of elapsed green time. Such a decline has been identified in Canada (*2*) and elsewhere (*18*).

The cumulative representation is favored by the CCG because it reflects the true flow from the onset of the green interval to any desired period (i.e., the saturation flow for any selected green interval includes the initial time period, during which the flow is lower). Moreover, for statistically significant samples, the cumulative graph exhibits a great degree of consistency. Consequently, it can be used for an initial assurance that the sample is valid and for a quick estimation by extrapolation of the basic saturation flow value in cases where the saturated flow does not persist long enough to be measured beyond, say, 10 sec.

The average saturation flow calculated using both formats is 1,840 pcu/hour green. The difference between the basic saturation flow determined from the two formats is generally negligible.

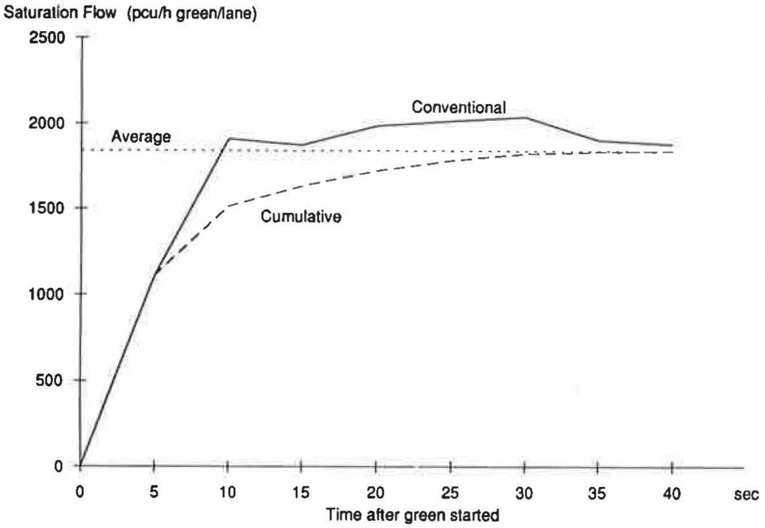


FIGURE 6 Basic saturation flow in conventional and cumulative formats for Survey 10.

HCM

This method differs from the CCG method in that a vehicle is considered discharged when its rear axle crosses the stop-line reference point. The period of saturation flow begins when the fourth vehicle in the queue crosses the reference point and ends when the last queued vehicle crosses the same point. Thus, unlike the CCG, the initial period of hesitant flow is not included. It is therefore to be expected that the HCM will yield a higher saturation flow than the CCG.

Flows were calculated by dividing 3,600 sec by the average headway for the saturated portion of the green interval (after the fourth queued vehicle). The flows calculated for each of the 30 surveyed cycles were averaged, and the heavy vehicle adjustment factor was then applied to obtain the basic saturation flow, consistent with the CCG for comparison.

ARRB Report 123

The time surveyed during each cycle was divided into three separate periods as follows:

1. The first 10 sec of green,
2. The remainder of the green period while saturated, and
3. The period after the end of green, including both amber and red.

The saturated time is considered to be the time to clear the vehicles that are stopped during the red interval, as well as those that arrive and join the back of the queue during the green interval (i.e., Intervals 1 and 2). Discharging vehicles were recorded as their front bumpers crossed the stop line.

This method stipulates that the saturation flow is calculated on the basis of the average headway of vehicles discharging only during Period 2 (i.e., not including the first 10 sec of green). The initial period of lower saturation flow is disregarded. As a result, it can be expected that the ARRB method would yield a higher saturation flow than the CCG method. Further, this saturation flow will likely be higher than that measured by the HCM method as well because, under normal circumstances, the passage of four passenger cars uses a period shorter than 10 sec (i.e., HCM discounts a smaller portion of the initial flow).

Saturation flow was then multiplied by the traffic composition factor to obtain a flow in through-car units, which, because only straight-through lanes were surveyed, is comparable to the CCG's passenger car units.

PCU Program

To have a generic basic saturation flow value to compare with the values found using the previously described methods, the computer program known as "PCU" was used. This program was developed by the University of Alberta (*4*) and was based on an older procedure suggested by Fischer (*9*). The main assumption is that, under equal conditions and for homogeneous traffic flow, the number of passenger-car units crossing the stop line of a single lane of an intersection approach should be constant during green intervals of equal duration. Variations are therefore attributed to differences in the behavior of individual vehicle categories. The PCU program finds the values of the pcu equivalents using a least squares optimization technique to minimize the variations in the number of pcu discharging for green intervals of equal duration. This method of data manipulation is similar to the application of linear regression described elsewhere (*7*). Because the PCU program calculates its own pcu equivalents, it is not biased with respect to the three documents (CCG, HCM, and ARRB Report 123). Nevertheless, its application includes the start-up lag as in the CCG.

The data input to the program are in time slices, before pcu conversion. The output may be in either the conventional or cumulative formats; again, the cumulative format was assumed to be the most reliable and was therefore chosen to serve as the basis for the comparisons.

RESULTS

Table 1 presents the results of the 10 surveys conducted in Edmonton, with the saturation flows calculated using the PCU cumulative format, the CCG cumulative format, and the HCM and ARRB procedures. All flows are in units adjusted for the presence of heavy vehicles in the traffic stream. Figure 7 shows these results graphically for easicr comparison.

The difference between the results of the PCU program and the CCG is negligible; on average, the CCG method yields flows only 0.13 percent higher than those derived from the PCU method. The correlation between the two methods was expected because of the flow at the start of green. Moreover, the pcu equivalents used by the CCG were derived in previous research with the assistance of the PCU program.

The ARRB method yields saturation flows up to 23 percent higher than the PCU flows—the range is 5 to 23 percent. This result corresponds to the expectation of higher values because of the exclusion of the first 10 sec of the green interval, which leaves only the most stable flow in the evaluation (see Figure 2).

The HCM values are in most cases higher than the PCU results. Surveys 5, 9, and 10 for the PCU program and the HCM are, however, quite comparable. Overall, the magnitude of the difference between the PCU and HCM values ranges from 0.3 to 11 percent. The variability, and the low range of the difference, is somewhat surprising because the first part of the flow is omitted, more or less similarly to the ARRB method, and the HCM and ARRB values were expected to be similar. This finding seems to be true only for Surveys 1, 3, and 6.

There seems to be no immediately identifiable consistent relationship between any factor and the magnitudes of the differences among the three basic methods. Initially, it was thought that the differences might be attributed to different pcu conversions in combination with the share of heavy vehicles. Nevertheless, Figure 7, which presents the surveys in the order of increasing percentage of heavy vehicles, does not confirm that hypothesis. As a result, it appears that it is not possible to determine the saturation flows for one method from the measurements based on another.

Because it has been shown that saturation flow exhibits a great degree of consistency over time (*4*), it was judged feasible to check the accuracy and consistency of the surveys by repeating one of the surveys under identical conditions after a short period of time. The survey at Location 1 (see Figure

TABLE 1 SATURATION FLOW SURVEY RESULTS AND PERCENTAGE OF HEAVY VEHICLES FOR 10 EDMONTON SURVEYS

Survey # / Location	Direction / Peak	Saturation Flow				% Heavy Vehicles
		PCU	CCG	HCM	ARRB	
1. St. Albert Trail / 137 Avenue	Southbound / A.M.	1890	1900	2100	2140	2.9
2. St. Albert Trail / 137 Avenue	Westbound / A.M.	1900	1910	2000	2190	3.3
3. St. Albert Trail / Yellowhead Trail	Northbound / P.M.	1710	1730	1900	1930	3.5
4. 137 Avenue / 97 Street	Westbound / P.M.	1600	1640	1710	1970	8.1
5. 142 Street / Stony Plain Road	Eastbound / A.M.	1800	1730	1810	1900	1.9
6. 98 Avenue / 75 Street	Northbound / P.M.	1690	1710	1850	1900	3.5
7. 127 Street / Yellowhead Trail	Westbound / P.M.	1790	1810	1880	2060	5.1
8. Whyte Avenue / 75 Street	Southbound / P.M.	1670	1670	1780	1910	4.3
9. Whitemud Drive / 111 Street	Eastbound / A.M.	1630	1650	1640	1720	12.2
10. St. Albert Trail / 137 Avenue	Westbound / A.M.	1890	1840	1960	2020	4.2

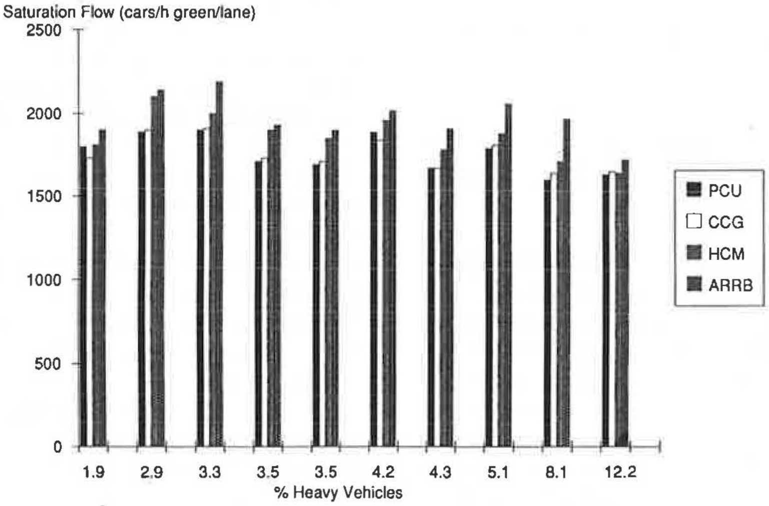


FIGURE 7 Comparison of surveyed saturation flows according to survey method and percentage of heavy vehicles.

4) was selected and repeated exactly 2 weeks after the initial survey (i.e., Survey 10 was designed to duplicate Survey 2). The PCU results for the two surveys indicate only a 0.58 percent difference (i.e., 1,900 versus 1,890 pcu/hour green), which can be considered insignificant. Consequently, the accuracy of the surveys appears to be adequate.

OTHER RESULTS

The CCG suggests that the passenger-car-unit equivalents for individual vehicle categories may vary regionally. Because the PCU program also calculates appropriate pcu equivalents, it was possible to compare these equivalents with those in the CCG. The results are presented in Table 2 and indicate a good fit, considering that the surveys described represent a small sample compared with the magnitudes of the survey efforts for the CCG. The CCG recommended values are a composite of summer and winter measured passenger-car-unit equivalents.

Vehicle headways were calculated from a number of randomly selected surveyed cycles, and the results are shown in Figure 8. For the reasons explained in connection with Figure 2, the traditional Greenshields curve (*5*) is not supported by the initial portion of the graph (the first two to three vehicles), but the average headway of 2.12 sec calculated here for the fifth and successive vehicles closely corresponds to Greenshields' reported value of 2.1 sec. Although not subject to this study, previous headway studies for Edmonton winter conditions exhibited patterns identical to that of Greenshields, even though headways were measured as front bumpers crossed the stop line (see Figure 9). Similar patterns for stop lines and front bumpers are frequently observed in regions with high degrees of red signal violations. A cautious entry by the first driver of the queue into the intersection is a likely explanation of this phenomenon and, perhaps, an added reason for survey consistency and careful interpretation of results.

Finally, the CCG reports a value of 1,650 pcu/hour green as the basic saturation flow in Edmonton for intersections of the type described in this discussion (for 1978 to 1984). The survey results presented are often significantly higher, ranging from 1,600 to 1,900 pcu/hour green. The average morning peak flow is 1,820 pcu/hour green (as calculated by PCU), whereas the average afternoon peak flow is 1,690 pcu/hour green, as shown in Figures 10 and 11. Higher saturation flows for morning periods with a more homogeneous and more aggressive driver population are mentioned in the CCG.

TABLE 2 COMPARISON OF PASSENGER-CAR-UNIT EQUIVALENTS

Vehicle Category	Passenger Car Unit Equivalent			
	Surveyed (PCU)	CCG	HCM	ARRB
Passenger Cars, Vans, Pick-Ups	1	1	1	1
Single-Unit Trucks	1.32	1.5	2	2
Multi-Unit Trucks	2.24	2.5	2	2
Buses	1.93	1.75	2	2
Motorcycles	0.75	0.5	1	1

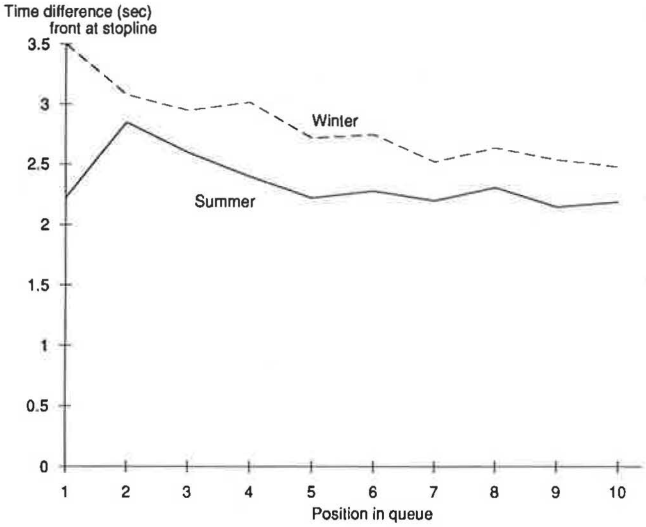


FIGURE 9 Comparison of vehicle headways for typical summer and winter conditions in western Canada.

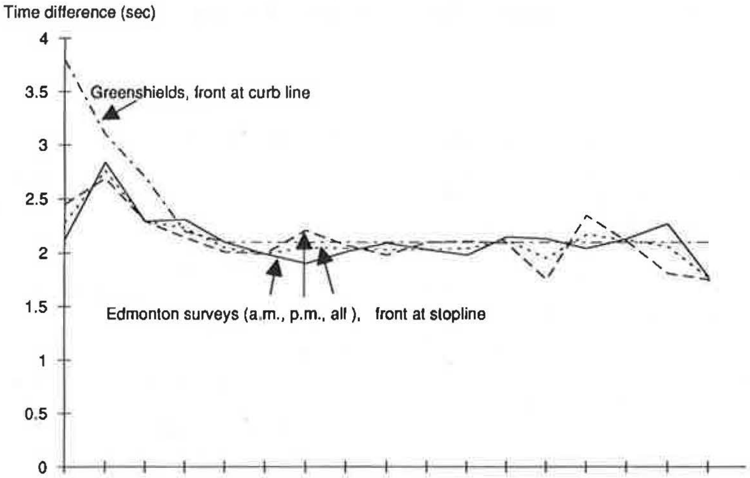


FIGURE 8 Vehicle headways calculated from 10 Edmonton surveys.

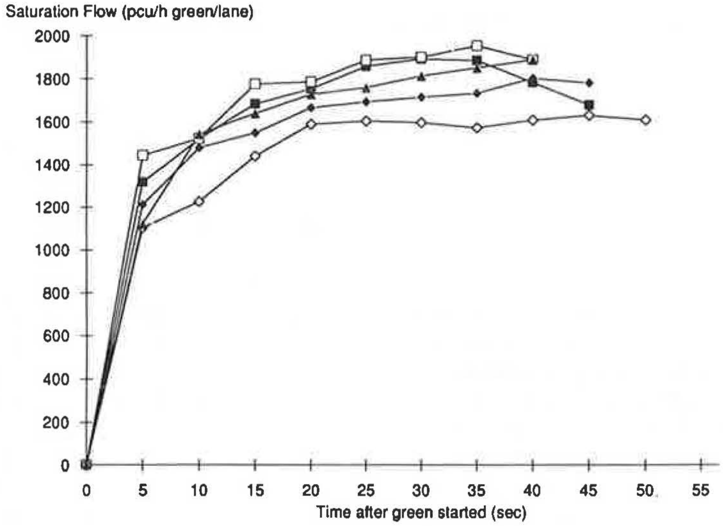


FIGURE 10 Morning peak saturation flows calculated by PCU program for five Edmonton surveys in cumulative format.

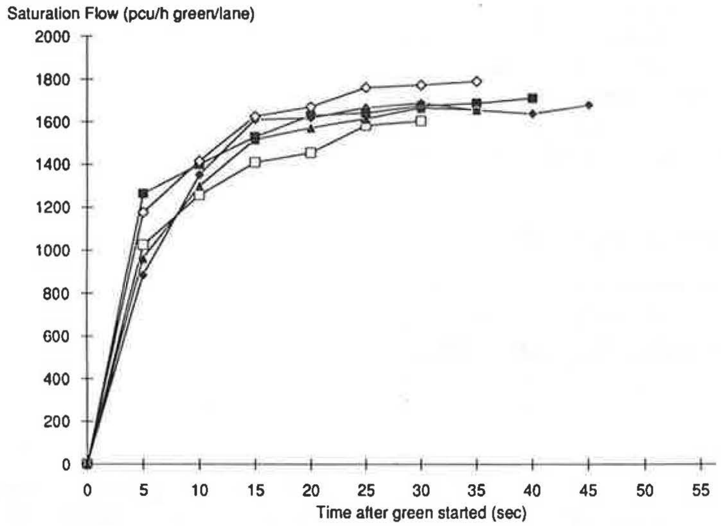


FIGURE 11 Afternoon peak saturation flows calculated by PCU program for five Edmonton surveys in cumulative format.

An increase in saturation flows over the past several years, and a corresponding increase in intersection capacity, has been observed by the city of Edmonton traffic engineering staff and in several smaller communities in the commuter shed. A logical explanation can be found in improved economic conditions in recent years. On the other hand, saturation flows in Calgary, which shares much of Alberta's economic prospects, remained unchanged. This finding emphasizes the value of periodic field saturation flow surveys in providing the most accurate and up-to-date basis for the analysis and design of signalized intersections.

As an additional observation, a tendency for calculated saturation flows to decrease over time for longer green intervals has been identified during the surveys (see Figure 8).

CONCLUSIONS

All sources suggest that accurate saturation flow measurements are necessary for reliable analysis and design of signalized intersections.

The deliberations during this study, and the results of the surveys presented, illustrate that the value of such surveys greatly depends on the method applied. The three methods commonly used in North America differ significantly with respect to use of the following attributes:

1. Roadway reference point,
2. Reference point on the vehicle that denotes discharge,
3. Start time for measurement and analysis, and
4. Heavy vehicle conversions.

The problem is compounded because the relationship among the values determined by various techniques is not consistent. Consequently, a method taken from one document cannot be reliably converted to the values needed for the calculation procedures used by another document.

The implications bear not only on the values of saturation flow itself but also on the applications of lost time, effective green interval, and effective red interval. As a result, care should be taken, both in practice and research, to correctly interpret the methodology used in the measurement of saturation flows.

This research study also confirmed some of the previous Canadian findings, such as the values of pcu equivalents for the basic vehicle categories and the consistency of saturation flows over shorter time spans. Although the consistency makes saturation flow an excellent basis for analysis and design of intersection signalization, saturation flow also exhibits longer trends that can only be detected by periodic surveys, which, naturally, must use methods commensurate with design and analysis practices in the region.

ACKNOWLEDGMENT

An operational research grant by the Natural Sciences and Engineering Research Council of Canada, which made this study possible, is gratefully acknowledged.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. D. Richardson, B. Stephenson, J. Schnablegger, and S. Teply. *Canadian Capacity Guide for Signalized Intersections*, 1st ed. (S. Teply, ed.). District 7, Institute of Transportation Engineers, and the University of Alberta, Edmonton, Canada, 1984.
3. R. Akcelik. *Traffic Signals: Capacity and Timing Analysis.* Research Report 123. Australian Road Research Board, Victoria, Australia, 1981.
4. S. Teply. Saturation Flow Through a Magnifying Glass. In *Proc., 10th International Symposium on Transportation and Traffic Flow Theory*, Toronto University Press, Toronto, Canada, 1981.
5. B. D. Greenshields, D. Schapiro, and E. L. Ericksen. *Traffic Performance at Urban Street Intersections.* Technical Report 1, ENO Foundation for Highway Traffic Control, Westport, Conn., 1947.
6. *A Method for Measuring Saturation Flow at Signalized Intersections.* Road Note 34/196, Road Research Laboratory, Crowthorne, England, 1963.
7. D. Branston and H. Van Zuylen. The Estimation of Saturation Flow, Effective Green Time and Passenger Car Equivalents at Traffic Signals by Multiple Linear Regression. *Transportation Research*, Vol. 12, 1978.
8. F. A. Haight. *Mathematical Theories of Traffic.* Academic Press, New York, 1963.
9. R. Fischer. Leistung von Kreuzungen im Gemischten Stadtverkehr. *Strasse und Autobahn*, Vol. 9, 1958.
10. W. Glueck and R. Schelzke. Startzeit und Zeitluecken an Signalgesteuerten Knoten. *Strasse und Autobahn*, Vol. 8, 1960.
11. C. Heideman. Messung der Leistungsfaehigkeit Signalgesteuerter Knoten. *Strasse und Autobahn*, Vol. 8, 1964.
12. R. Zuberbuehler. Elemente fur Dimensionierung von Lichtsignalanlagen bei Strassenknoten. *Strasse und Verkehr*, Vol. 8, 1966.
13. R. Krell. *Die Bemessung Lichtsignalgesteuerter Strassenknoten ohne Iteration.* Forschungsarbeiten aus dem Strassenwesen, Neue Folge, Heft 53, 1962.
14. W. Doerfler. Fahrdynamik an Signalgesteuerten Kreuzungen. *Strassenbau und Strassenverkehrstechnik*, Vol. 38, 1965.
15. G. Schmidt and F. Gothan. Ein Beitrag zur Anwendung der Zeitbeiwerte nach Greenshields bei der Berechnung der Signalanlagen. *Strasse und Autobahn*, Vol. 12, 1959.
16. K. W. Axhausen, M. Fellendorf, and D. Hook. Zur Abhangigkeit der Zeitbedarfswerte von der Knotenpunktsgeometrie. *Strassenverkehrstechnik*, Vol. 5, 1989.
17. J. P. Braaksma, R. C. Ridley, and P. H. Jones. The Effect of Roadway Salting on Safety and Mobility. *Canadian Journal of Civil Engineering*, Aug. 1987.
18. OECD Scientific Expert Group. *Traffic Capacity of Major Routes.* Road Transport Research, Organisation for Economic Cooperation and Development, Paris, France, July 1983.
19. D. Branston. Some Factors Affecting the Capacity of Signalized Intersections. *Traffic Engineering and Control*, Aug./Sept. 1979.
20. R. M. Kimber and M. C. Semmens. An Experiment To Investigate Saturation Flow at Traffic Signal Junctions. *Traffic Engineering and Control*, March 1982.
21. R. M. Kimber, M. McDonald, and N. B. Hounsell. The Prediction of Saturation Flow for Road Junctions Controlled by Traffic Signals. *Digest of Research Report RR67*, Transport and Road Research Laboratory, Crowthorne, England, 1986.
22. *Passenger Car Equivalents of Goods Vehicles of Different Lengths at a Signal-Controlled Junction.* Leaflet LF686, Transport and Road Research Laboratory, Crowthorne, England, 1978.
23. S. Teply. A Canadian Problem: Winter Intersection Capacity. In *Proc., 3rd Annual Meeting of the Institute of Transportation Engineers, District 7*, Calgary, Alberta, Canada, 1978.

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

Delay Effects on Driver Gap Acceptance Characteristics at Two-Way Stop-Controlled Intersections

WAYNE K. KITTELSON AND MARK A. VANDEHEY

This paper examines the 1985 *Highway Capacity Manual* (HCM) definition of critical gap for two-way stop-controlled (TWSC) intersections. Minor street left-turning movements at unsignalized three-leg intersections are examined. On the basis of this examination, a revision is recommended for the HCM definition, which better reflects actual driver behavior and provides a better estimate of the capacity of TWSC intersections. The effects of front-of-queue delay on the length of the critical gap were investigated. On the basis of a limited amount of field data, the critical gap was found to be significantly affected by the amount of front-of-queue delay incurred by individual drivers. It is therefore recommended that any delay-based level of service (LOS) criterion for TWSC intersections should incorporate lower delay thresholds than are used for signalized intersections at least in the LOS *D*, *E*, and *F* regions. Finally, the collected field data demonstrate that the critical gap for minor street left-turning vehicles is affected by the type of major street conflict (same direction versus opposite direction) experienced. From this finding, it is concluded that the distribution of major street traffic can have a substantial effect on the capacity of the minor street left-turn movement.

Two-way stop controlled (hereafter referred to as TWSC) intersections are one of the most common intersection types within the United States and abroad. They are also among the most complex to analyze with respect to their capacity and level of service (LOS) characteristics. This is because these intersections provide a high degree of discretion to individual drivers in how they react to conflicting traffic streams. Thus, driver behavior characteristics play an important role in defining how a TWSC intersection operates, resulting in greater variability than is typically found in more controlled settings such as those at signalized intersections.

Driver gap acceptance characteristics have long been identified as a bellwether parameter for establishing individual behavior patterns and their impact on the operation of a TWSC intersection. The 1985 *Highway Capacity Manual* (HCM) (*1*) procedure for analysis of TWSC intersections (Chapter 10) relies on the critical gap as the basis for estimating the potential capacity of a minor movement. The HCM procedure also recognizes that the critical gap is not a constant value even for individual intersections, but is affected by the average running speed on the major road, the basic number of through lanes on the major road, and the type of minor movement being made.

Although critical gap adjustments reflecting the effects of these independent variables are probably appropriate to include within the HCM procedure, there remain several deficiencies with the method:

- The HCM procedure defines the critical gap to be the median gap size that is accepted by drivers in a given situation. This definition is not consistent with the definition currently being used within the profession. Additionally, it is a definition that leads to inaccurate estimates of actual driver behavior characteristics. Thus, there is need for a revised definition.
- The HCM procedure assumes that the critical gap remains constant over time. In fact, it seems reasonable to expect that drivers will accept shorter gaps as their delay time (including both total delay time and also delay time in the front position of a minor movement queue) increases.
- The HCM procedure assumes that there is only a single critical gap for each minor movement, and that this same critical gap applies equally to all conflicting traffic flows. In fact, it seems reasonable to expect that drivers evaluate the acceptability of a gap at least partly on the amount of time they expect to be exposed to the major street conflict. Thus, a left-turning driver from the minor street will require a longer gap between conflicting vehicles traveling in the direction in which the driver intends to travel than between conflicting vehicles traveling in the opposite direction.

Each of these issues is addressed through the collection and analysis of real-world data. To simplify the analysis and minimize the effects of other variables not specifically under consideration, this discussion is limited to the minor street left-turn movement at TWSC intersections that are outside the influence area of signalized intersections and possessing a two-lane cross-section on the major street (one through lane in each travel direction). Additionally, the analysis is limited to intersections where the minor street left-turn movement occurs in a marked or de facto exclusive turn lane. Finally, the analysis is also limited to data collected at *T*-intersections to minimize the potential for the findings to be affected by influences from other opposing minor street movements. Even so, the results of this analysis are considered to be equally applicable to four-leg TWSC intersections.

DEVELOPING A PRACTICAL DEFINITION FOR CRITICAL GAP

The critical gap is currently defined within the HCM procedure as the median gap size that is accepted by drivers in a

Kittelson & Associates, Inc., 610 S.W. Alder, Suite 700, Portland, Oreg. 97205.

given situation. The flaw in this definition as it relates to the HCM procedure can best be illustrated through a simple and purely hypothetical example. Consider the minor street left-turn movement onto a major street in which the traffic flow is unidirectional and the traffic flow rate is uniform at one vehicle every 30 sec (see Figure 1). If minor street vehicles arrive at this intersection randomly, then they are likely to accept lag gaps ranging in length from 3 to 30 sec. (A lag gap is the gap that is defined by the arrival of the minor street vehicle at its beginning and the arrival of a conflicting vehicle at its end.) For those minor street vehicles arriving such that the lag gap is less than 3 to 4 sec., it is likely that they will reject the lag gap in favor of the following 30-sec gap. Using the HCM definition, the critical gap will likely be computed to be something over 15 sec, even though it is clear that almost all drivers are willing to accept gaps that are considerably shorter.

At the crux of this problem is the implicit assumption in the HCM definition that gap acceptance data are the only meaningful information necessary to determine driver gap acceptance behavior. Such an assumption is not valid: it is intuitively obvious that most drivers will accept 15-sec gaps, and the fact that they do sheds no light on how these same drivers might react to a 5- or 6-sec gap. Worse, reliance on such an assumption will usually lead the analyst to an incorrect conclusion regarding an appropriate critical gap length. Thus, it is important to consider both gap acceptances and gap rejections in estimating the actual critical gap.

This observation has been made by many other transportation professionals (*2–4*), some of whom have also developed procedures for obtaining a more reliable estimate of the critical gap length. As early as 1960, Bissell suggested a better definition for critical gap as the median probability of accepting a gap of a given size. This definition takes account of both acceptance and rejection characteristics, and is clearly superior to the current HCM definition.

In order to illustrate the applicability of this revised definition in a real-world situation, consider the actual gap acceptance frequency distribution shown in Figure 2. This distribution reflects observations of minor street left-turning movements at a T-intersection on a major street with one lane in each direction of travel. The design speed of the major street is between 40 and 50 mph. Inspection of this figure suggests that meaningful information regarding driver gap acceptance behavior ceases beyond gaps longer than about 11 or 12 sec, because nearly all of the presented gaps are accepted beyond this point. Yet the large number of accepted gaps longer than 12 sec has the effect of forcing the HCM-defined critical gap to be unrealistically high. Using the entire gap frequency distribution, the median accepted gap length can be calculated to be between 8.5 and 9.0 sec. If, however, only the accepted gaps equal to or shorter than 12.0 sec are considered, then the median gap length decreases to between 5.0 and 6.0 sec. Thus, the longer gaps cause a systematic bias to the defined critical gap that is unwarranted in its effect, unpredictable in its magnitude, and misleading to all subsequent intersection capacity analyses.

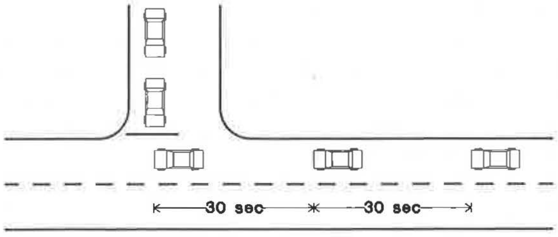


FIGURE 1 Hypothetical vehicle arrival profile.

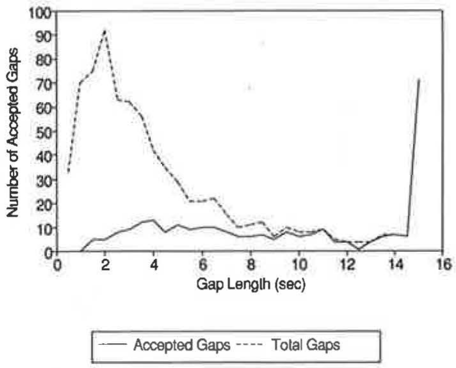


FIGURE 2 Frequency distribution of accepted gaps.

Figure 3 identifies, for the same data set shown in Figure 2, the estimated critical gap length using the suggested definition revision. In this case, drivers were observed to accept a 6-sec gap with a 50 percent probability. This observation is consistent with the results of the intuitive approach described earlier, and confirms the reasonableness of the suggested revision to the HCM critical gap definition. In the following sections, this revised definition is used in estimating the critical gap for left-turning movements at an unsignalized intersection.

The findings of this paper also have important implications to current procedures described in AASHTO's 1984 *Policy on Geometric Design of Highways and Streets* (*5*) for computing intersection sight distance requirements. Specifically, Figure 4 is taken from this document and illustrates the recommended intersection sight distance requirements for minor-street left turns at at-grade intersections under three conditions:

- Case B–1 illustrates the minimum sight distance required for a passenger vehicle to turn left into a two-lane highway

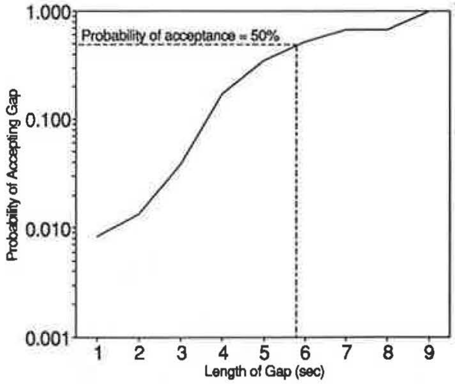


FIGURE 3 Estimated critical gap using revised definition.

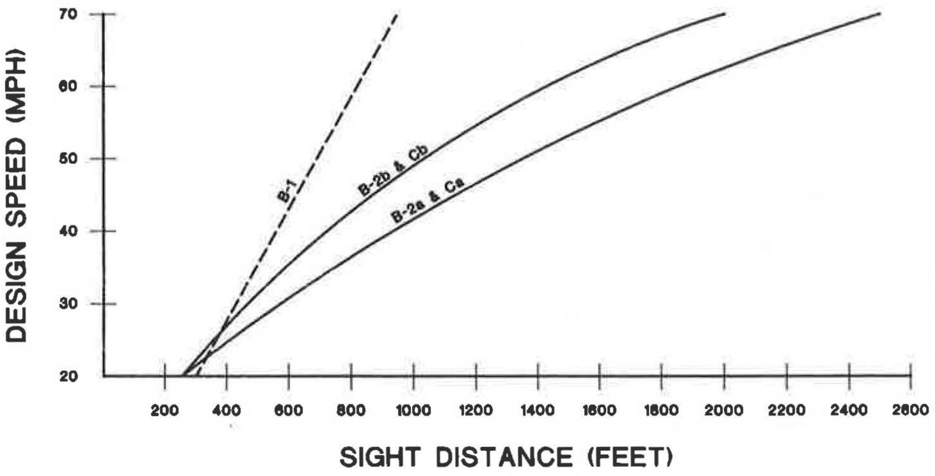


FIGURE 4 Intersection sight distance recommendations for at-grade intersections.

and across a passenger vehicle approaching at design speed from the left.

- Case B–2a illustrates the minimum sight distance required for a passenger vehicle to turn left into a two-lane highway and attain the design speed without being overtaken by a vehicle approaching from the right and maintaining the design speed.
- Case B–2b illustrates the minimum sight distance required for a passenger vehicle to turn left into a two-lane highway and to attain the average running speed without being overtaken by a vehicle approaching from the right and reducing its speed from the design speed to the average running speed.

Figure 5 shows the translation of each of these distance requirements into time gaps on the basis of the major street design speed. From the perspective of an intersection sight distance, Curve B–2a is the most restrictive, and it is difficult to imagine that side street drivers would have any reasonable inhibitions about accepting gaps longer than those defined by this curve. The same can also be said of Curves B–2b and B–1 with respect to the field data shown in Figures 2 and 3: at a design speed of 40 to 45 mph, the least restrictive curve (B–1) suggests a minimum gap length of approximately 9 sec, whereas Figure 3 shows side street vehicles regularly accepting much shorter gaps than this. In fact, Figure 3 indicates that there is nearly a 100 percent probability that side street drivers will accept any gap equal to or longer than 9 sec. Thus, the AASHTO guidelines for intersection sight distance do not match well with actual driver behavior patterns, and result in length recommendations that are much more generous than drivers either require or will use.

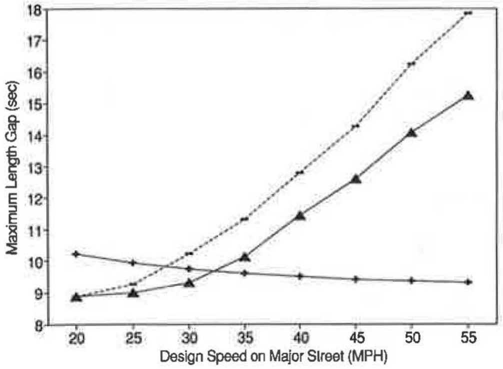


FIGURE 5 Intersection sight distance requirements expressed as gap lengths plus sign, Curve B-1; bar, Curve B-2a; solid triangle, Curve B-2b.

ESTIMATING EFFECTS OF DELAY ON CRITICAL GAP LENGTH

It was originally hypothesized that drivers will accept shorter gaps as their delay time in the front position of a minor movement queue increases. This is a common-sense hypothesis that is generally supported in the published literature (*6*), but almost always without the benefit of specific quantitative findings. Also not well documented is the extent to which critical gaps are affected by front position delay. Therefore, this investigation sought to corroborate the hypothesis that delay causes a reduction in critical gap length, and, if true, to quantify the magnitude of the delay effect on critical gap.

It is important to recognize that the front-of-queue delay is only one component of the total delay incurred by minor-street left-turning drivers. The delay time spent in the queue behind other vehicles is also considered by many to be an important factor in determining overall driver frustration levels. Although this may be true, it is also likely that the two delay parameters are highly correlated: long delays in the front-of-queue position usually go hand-in-hand with long total delays. Therefore, it is not expected that any significant bias was introduced into this study by concentrating on only the front-of-queue delay time.

The investigation focused on the minor street left-turn movement at unsignalized intersections where the major street consisted of only one lane in each direction of travel. Special attention was given to the site selection process to ensure that the following conditions prevailed:

1. The intersection was sufficiently removed from upstream and downstream traffic signals so that major street traffic

flows were not affected by the platooning these signals cause. In all cases, the intersections were separated from signalized intersections by at least 1 mi.

2. All intersection approaches were relatively flat, and intersection sight distance was unrestricted.

3. Traffic volumes on the major and minor streets were sufficient to cause a range of front position delays to left-turning vehicles, ranging from 0 to 10 sec at the low end to beyond 40 sec at the high end.

4. The cross-section on the major street was such that no center area refuge was provided for the minor street left-turning movements. Thus, all left turns were made as a single movement.

Three sites were investigated that met all of these site selection criteria. Although this is not a large number of sites, it was considered sufficient for the purposes of this preliminary investigation.

Data were collected at each site through use of a laptop computer programmed to accept keystrokes identifying the arrival and departure of vehicles on each major and minor movement, and to record the real time (to the nearest 10th of a second) that each such arrival and departure occurred. The data were recorded on diskettes and later analyzed in an office environment using additional customized computer programs. A total of 5 hr of data were collected, including 300 left-turn observations and 1,500 gaps.

To determine the effects of delay on driver gap acceptance characteristics, the following data were collected for each minor street left-turning vehicle:

1. The arrival time of the vehicle at the front of the queue;
2. The vehicle's departure time;
3. The length of the accepted gap; and
4. The number of rejected gaps that were longer than the gap that was ultimately selected.

The data were grouped into 10-sec delay increments and summarized as shown in Figure 6. The dashed line shown in this figure represents the best-fit regression line through the data points shown. Each data point represents an average of observations taken at all sites. Clearly, Figure 6 shows a strong relationship between the amount of delay incurred by a driver while in the front of the queue and the number of longer gaps that were previously rejected. On this basis, it is concluded that there is quantitative support for the hypothesis that drivers reduce the length of the gap they are willing to accept as a function of the amount of delay they incur.

In order to determine the magnitude by which critical gaps are affected by delay, critical gap lengths were computed for the following vehicle groupings:

1. Only those minor street drivers whose accepted gap was longer than all gaps they previously rejected. Presumably, these drivers were either unaffected by delay in the front queue position or accepted a gap sufficiently long that it masked the effects of the delay.

2. Only those minor street drivers whose accepted gap was shorter than one or more gaps they had previously rejected. Presumably, at least some of the delay effects on these drivers can be seen through examination of the gaps that they accepted.

The results of this analysis are shown in Figure 7, and clearly indicate a strong relationship between the previous rejection of longer gaps and the length of the observed critical gap. This relationship, combined with the relationship shown in Figure 6, suggests that delay has an important effect on the critical gap length.

These findings are significant because they indicate that minor-street drivers attempt to reduce their exposure to long delays at unsignalized intersections by accepting shorter gaps. Thus, the total amount of delay incurred by minor-street vehicles at an unsignalized intersection does not necessarily reflect the true level of service in the same way that it does at signalized intersections. This difference occurs because, at unsignalized intersections, drivers are able to reduce their delay by accepting shorter gaps, even though doing so results in greater risk and suggests a higher degree of driver frustration than would be indicated if only the total amount of delay were considered. Thus, if delay is to be used as the primary measure of LOS at an unsignalized intersection, then the delay thresholds should be lower than those currently in use for signalized intersections.

It might also be hypothesized that minor-street drivers who accept shorter gaps do so at the expense of additional delay to main-street through traffic. Thus, it may be that in accepting a shorter gap, minor-street drivers cause some slowing of the main street through traffic and, in effect, force a longer

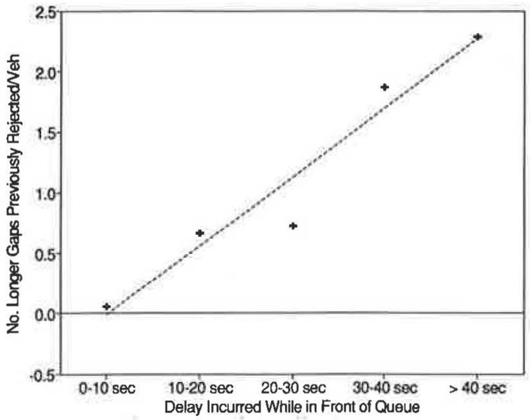
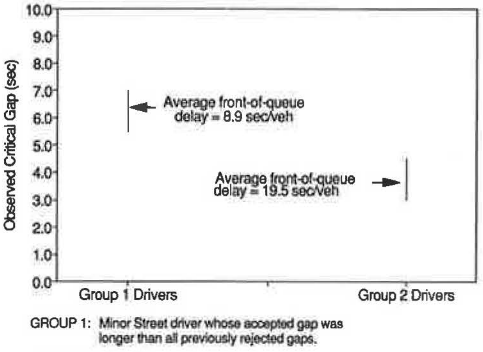


FIGURE 6 Effects of delay on gap acceptance.



GROUP 1: Minor Street driver whose accepted gap was longer than all previously rejected gaps.

GROUP 2: Minor Street driver whose accepted gap was shorter than one or more previously rejected gaps.

FIGURE 7 Effects of front-of-queue delay on critical gap.

gap than would have otherwise existed. No data were collected during this study to either confirm or reject this hypothesis.

ASSESSING EFFECTS OF CONFLICT TYPE ON CRITICAL GAP LENGTH

Further examination of Figure 7 reveals a very short critical gap (between 3.0 and 4.0 sec) for those minor-street drivers whose accepted gap was shorter than gaps they had previously rejected. This accepted gap is so short that, on first glance, it raises questions regarding the safety and judgment of drivers who would accept such gaps. Yet the fact that this is the median accepted gap indicates that many drivers find this condition to be acceptable. Further, the intersections where these data were collected were not noteworthy with respect to their accident history.

In order to explain driver behavior in such situations, it was hypothesized that drivers evaluate the acceptability of a gap at least partly on the amount of time they expect to be exposed to the current major-street conflict. Thus, a left-turning driver from the minor street will require a longer gap between conflicting vehicles approaching from the right than from the left. This is because major-street drivers approaching from the right are traveling in the same direction as is intended by the turning vehicle, and therefore have the potential to overtake the turning vehicle some time after the turn has been completed. Further, this potential for an overtaking conflict remains until the turning vehicle has accelerated to the average operating speed. In contrast, a major-street vehicle approaching from the left poses only a crossing conflict to the turning vehicle, and this conflict disappears as soon as the turn is completed.

The hypothesis was tested by disaggregating the collected data according to the direction of the major-street vehicle defining the end of each accepted gap. Thus, gap acceptance characteristics for minor-street left turns were compared between drivers whose primary conflict was a vehicle traveling in the same direction and drivers whose primary conflict was a vehicle traveling in the opposite direction. This comparison was first made for the entire data set, including drivers with short as well as long delays. The results of this initial comparison are shown in Figure 8, and reveal no particular difference in the overall gap acceptance characteristics between the two disaggregated data sets. The distribution pattern for accepted gaps is similar in both data sets, and even though there were clearly more shorter gaps accepted against opposing traffic than against same-direction traffic, no clear pattern can be easily discerned.

The situation changes dramatically when only those drivers who were clearly being challenged in their gap acceptance behavior are considered. Figure 9 shows that in this case, there is a distinct difference between the gap acceptance characteristics of the two data sets, with much shorter gaps being accepted against opposite-direction traffic. For the data set used to compile Figure 9, 89 percent of all observed accepted gaps less than or equal to 3.0 sec in length were accepted against opposite-direction traffic.

On the basis of the information shown in Figures 8 and 9, it is concluded that, for minor-street left turns, gap acceptance characteristics are significantly affected by the directional distribution of the conflicting major-street tr ffic flows. This finding is important because it indicates that the directional distribution of major-street traffic can significantly affect the capacity and LOS experienced by minor-street left-turning traffic.

CONCLUSIONS

The analysis results should be considered to be preliminary because they are based on a limited amount of data collected at only three different sites in a single geographic region. Nevertheless, there appears to be a strong correlation between delay and gap acceptance characteristics. Additionally, several unexpected findings have resulted from this preliminary effort. The following summarizes the key analysis findings:

- A revised definition of critical gap would better reflect actual driver behavior, and consequently would provide a better estimate of the capacity of TWSC intersections.
- The critical gap is affected by the amount of front-of-queue delay that is incurred by individual drivers. Specifically, drivers accept shorter gaps as front-of-queue delay increases. This finding implies that the capacity of the minor movements increases as the LOS degrades.
- Given that drivers accept shorter gaps as front-of-queue delay increases, it seems reasonable to conclude that any

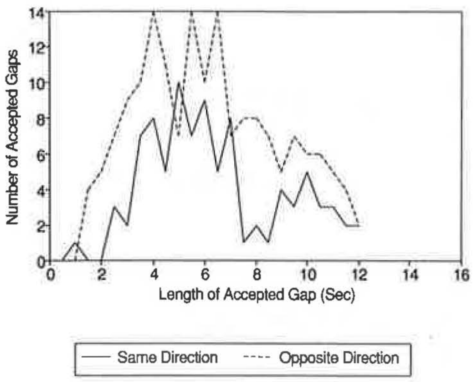


FIGURE 8 Effect of conflict type on critical gap for all vehicles.

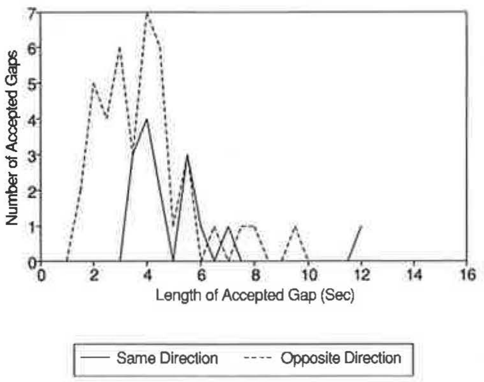


FIGURE 9 Effect of conflict type on critical gap for delayed vehicles.

delay-based LOS criterion for TWSC intersections should incorporate lower delay thresholds than are used for signalized intersections at least in the LOS *D*, *E*, and *F* regions. This is because the acceptance of shorter gaps implies more risk taking by minor-street drivers. Overall driver safety and comfort are also elements to be considered in the definition of LOS. Therefore, it follows that the higher degree of risk taken by minor-street drivers in the LOS *D*, *E*, and *F* regions should reasonably translate, for any given delay level, into a lower LOS for an unsignalized intersection than for a signalized intersection.

- The critical gap for minor-street left-turning vehicles is also affected by the type of major-street conflict (same direction versus opposite direction) that is experienced. Thus, the directional distribution of major-street traffic can have a substantial effect on the capacity of this movement.

FUTURE RESEARCH NEEDS

There is an acute need for additional research into driver behavior characteristics at TWSC intersections:

- The analysis findings described earlier should be confirmed through development and evaluation of a larger data base reflecting driver characteristics observed over a much wider geographic area.
- The findings and conclusions focus on only the minor-street left-turning movement. Subsequent analyses should be expanded to include a detailed examination of driver characteristics associated with all other minor-street movements at TWSC intersections.
- The data base used was specifically limited to *T*-intersections. Subsequent analyses should also address driver behavior characteristics at four-leg intersections.
- The data base used was specifically limited to level-grade TWSC intersections located on two-lane major streets with unlimited intersection sight distance and outside the influence area of nearby traffic signals. Subsequent analyses should consider the effects of varying some of these variables on driver gap acceptance characteristics.

Substantial benefits are likely to be obtained from additional research into the operating characteristics of TWSC intersections. A partial listing of some of the benefits likely from such research includes the following:

- Improved warrants for intersection signalization may be possible. By recognizing conditions under which TWSC intersections can continue to operate effectively, unnecessary signals (which represent a continuing maintenance cost for government agencies and additional delay to motorists) can be avoided.
- Better-access management strategies can also be developed. Understanding how each TWSC intersection operates within a system-wide context will result in a more efficient use of existing arterial streets. It will also assist in defining the appropriate number of access drives for private and public developments.
- Modifications to current minimum intersection sight distance requirements may be possible for urban areas on the basis of an improved understanding of driver gap acceptance characteristics. Current AASHTO procedures consider vehicle acceleration and deceleration dynamics, but do not specifically account for the gap acceptance characteristics of side street vehicles.
- Improved procedures for estimating the capacity and LOS of shared-lane and permitted left-turn movements at signalized intersections may be possible. In both of these situations, left-turning drivers are provided the freedom to select what they consider to be an appropriate gap in the opposing traffic flow, and therefore are in a similar position to that which occurs at TWSC intersections.
- A more realistic procedure for estimating the capacity and LOS of TWSC intersections is likely. Although Chapter 10 of the HCM is a significant advancement over previously available procedures, users continue to be concerned about some of its predictive abilities. A better understanding of underlying driver behavior characteristics at TWSC intersections is key to improving this predictive ability.

Taken together, these benefits have the potential to affect virtually every facet of transportation engineering. The importance of focusing additional formal research activities in the area of unsignalized intersections can therefore not be overstated.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. H. H. Bissell. *Traffic Gap Acceptance From a Stop Sign.* Graduate Research Report. University of California Institute of Transportation and Traffic Engineering, Berkeley, 1960.
3. R. Ashworth and C. G. Bottom. Some Observations of Drivers' Gap Acceptance Behaviour at a Priority Intersection. *Traffic Engineering and Control*, Vol. 18, pp. 569–571.
4. A. J. Miller. Nine Estimators of Gap Acceptance Parameters. In *Proc., 5th International Symposium on the Theory of Traffic Flow and Transportation* (G. F. Newell, ed.), 1971, pp. 215–235.
5. *A Policy on Geometric Design of Highways and Streets*, AASHTO, Washington, D.C., 1984.
6. W. R. McShane and R. P. Roess. *Traffic Engineering.* Prentice-Hall, Englewood Cliffs, N.J., 1990.

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

Capacity and Delay Characteristics of Two-Way Stop-Controlled Intersections

Michael Kyte, Chris Clemow, Naseer Mahfood, B. Kent Lall, and C. Jotin Khisty

Three objectives are sought: (a) to present a data base for two-way-stop-controlled (TWSC) intersections that can be used to investigate the factors that influence delay and capacity, (b) to identify some of the factors that affect delay and capacity at a TWSC intersection, and (c) to develop a set of preliminary models to estimate delay and capacity. Traffic flow, delay, and geometric data were collected at nine TWSC intersection sites in the Pacific Northwest encompassing a total of 13 hr of intersection operation. A total of 970 minor-street (subject approach) vehicles were observed and nearly 2,000 accepted and rejected gaps were identified and recorded. Each site has several common characteristics: four approach legs, single lanes on each approach, and a 25-mph speed limit on the major street. Several factors were identified that may influence delay and capacity at a TWSC intersection. The time waiting in queue (queue time) is affected by the traffic flow rate on the subject and opposing approaches, whereas the time spent waiting at the stop line (service time) is affected by the traffic flow rate on the conflicting approaches. The capacity of the subject approach is also affected by the flow rate on the conflicting approaches. The size of the accepted gap is affected by the length of time that a vehicle has been delayed, the flow rate on the conflicting approaches, and the directional movement of the subject vehicle.

Literature describing both empirical and theoretical studies of two-way stop-controlled (TWSC) intersections is plentiful. One major source is a compendium of papers presented at the 1988 International Workshop, *Intersections Without Traffic Signals* (*1*). These papers represent the latest research efforts in Europe, Asia, and the United States on this topic. Common to much of this work are two elements: (a) the determination of the critical gap as a function of the intersection geometry and major-street flow characteristics, and (b) the hierarchical classification of conflicting traffic streams as the basis for computing movement capacities.

These two elements are essential components of the standard analysis methodology used in the United States, defined by the 1985 *Highway Capacity Manual* (HCM) (*2*). Despite its widespread use, there is a general consensus that the HCM methodology requires substantial modification (*3*). Several empirical and theoretical limitations are most commonly cited:

1. The measure of effectiveness (MOE) used in the HCM procedures is reserve capacity. Although reserve capacity is directly related to delay in the German and Swedish procedures (*4,5*), the HCM provides only a broad description of the delay likely to be encountered for each range of reserve capacity. Thus, it is not now possible to forecast the performance of an intersection operating under various control conditions (e.g., signal control or two-way-stop control) using a single MOE such as delay.
2. The HCM procedure is based on a German method developed in 1972 and has been verified with only limited studies for U.S. conditions (*6*). There is an obvious need to develop a data base for U.S. conditions for TWSC intersections.
3. The Germans have since identified several theoretical problems with their original procedure and have now developed a new set of paper-and-pencil procedures and a computer simulation model (*7*). Also, the German procedures were developed only for single-lane approach sites. Multilane sites in Germany are always signalized.
4. There are several methodological problems with the current HCM procedure, primarily related to the use of the critical gap. First, the methodology used to estimate the critical gap usually results in an overestimation of this parameter. Kittelson and Vandehey in a companion paper in this Record propose a new method to obtain a more accurate estimate of the critical gap; this method deserves consideration. Second, it has been suggested that the critical gap is not constant and that drivers tend to accept smaller gaps the longer they wait in queue or at the stop line. This variability in the critical gap needs to be quantified. Finally, it is likely that the critical gap depends on the direction of the minor-street vehicle. The current HCM procedure assumes that the critical gap varies only with the travel speed and number of lanes on the major street. One solution to this problem may be to adopt the approach used in the United Kingdom (*8*), where empirically based models for delay and capacity have been developed that do not rely on estimation of the critical gap.

Clearly, there is a need to address several major issues with respect to the understanding of traffic flow at TWSC intersections. Three objectives begin to address some of these issues:

1. To present a data base for TWSC intersections that can be used to investigate the factors that influence delay and capacity,
2. To identify some of the factors that affect delay and capacity, and
3. To develop a preliminary set of models that can be used to estimate delay and capacity.

M. Kyte, C. Clemow, and N. Mahfood, Department of Civil Engineering, University of Idaho, Moscow, Idaho 83843. B. K. Lall, Department of Civil Engineering, Portland State University, Portland, Oreg. 97207. C. J. Khisty, Department of Civil Engineering, Washington State University, Pullman, Wash. 99163.

DEFINITIONS OF TERMS

Intersection Approaches

Queue activity is monitored on one of the minor-street approaches. This approach is called the "subject approach." The other (facing) minor-street approach is called the "opposing approach." The two major-street approaches are called the "conflicting approaches."

Microscopic or Instantaneous Major-Street Flow Rate

The microscopic analysis requires a definition of the major-street flow as seen by each subject approach vehicle. Consider the following sequence of events. A subject vehicle arrives at time t_1. Two conflicting approach vehicles pass by at times t_2 and t_3, respectively. The gaps $t_3 - t_2$ and $t_2 - t_1$ are rejected by the subject vehicle. The subject vehicle departs at t_4, whereas the next conflicting vehicle passes at t_5. The definition of the conflicting flow rate as seen by this particular subject vehicle is the number of observed conflicting vehicles divided by the observation time, or

$$\text{Flow rate} = \frac{3}{t_5 - t_1} \tag{1}$$

This definition differs from the standard (macroscopic) method of estimating flow rates, in which averages are reported for some fixed period, usually 15 min or 1 hr. This distinction will be noted whenever the microscopic or instantaneous flow rate is used.

Accepted Gap

The standard definition for the accepted gap is $t_5 - t_3$. However, it is common for the conflicting vehicle representing the end of the accepted gap to pass through the intersection long after the departure of the subject vehicle. Thus, the subject vehicle never sees this conflicting vehicle, and the calculated accepted gap is not meaningful. Kittelson and Vandehey in a companion paper in this Record propose that a maximum value be used for the accepted gap based on criteria of sight distance and major-street speed. They suggest in most cases that a maximum of 12 sec is reasonable.

Delay

Stopped delay was calculated for each minor-street subject vehicle by measuring two separate components. The time between the arrival at the end of the queue and the time when the vehicle arrives at the stop line is defined as the "queue time." The time between the arrival at the stop line and the departure from the stop line is defined as the "service time."

COLLECTION AND REDUCTION OF TRAFFIC OPERATIONS DATA

Data were collected during nine different days at four different intersections over a period of 13.1 hr. Each of these days is classified as a sample site. Each site has four approach legs, a single lane on each approach, and a speed limit of 25 mph on the major-street approaches.

Each sample site was videotaped with the camera in a position to observe vehicles on each approach as they passed through the intersection as well as the queue activity on one minor-street approach.

Data collection software was used to record flow rate and delay data while observing the videotapes from each sample site (*9*). The time that each vehicle entered the intersection was recorded and its directional movement (left-turn, through, or right-turn) was noted. Flow rates were calculated on the basis of these data. For one minor-street approach (referred to as the subject approach), the times that each vehicle entered the end of the queue, arrived at the stop line, and entered the intersection were also noted. Delays were calculated on the basis of these data.

OVERVIEW OF SITE CONDITIONS: THE DATA BASE

The first objective is to present a data base for TWSC intersections. Table 1 presents some of the important geometric, traffic flow, and delay data for the nine sample sites.

Sight Distance Data

The sight distance data were assessed for each of the intersection approaches at the nine sample sites. Four categories were used in this assessment, including excellent, good, fair, and poor. Although admittedly qualitative, the assessment may be useful in determining if sight distance affects driver behavior and thus delay and capacity characteristics.

Upstream Control and Platoon Characteristics

The arrival patterns of the traffic on the conflicting approaches have an effect on the behavior of the subject vehicle. Up-

TABLE 1 SITE DATA

Site	Site Distance	Observed Major Street Platoon	Subject Volume, vph	Opposing Volume, vph	Conflicting Volume, vph	Mean Service Time, sec	Mean Queue Time, sec
1	Fair	Cyclic	137	12	642	8.3	3.4
2	Fair	Random	177	27	183	5.6	11.7
3	Excellent	Cyclic	38	69	1100	9.2	2.1
8	Good	Uniform	245	62	279	5.7	8.3
9	Good	Cyclic	388	200	464	6.9	17.7
11	Good	Cyclic	392	187	472	7.0	18.1
12	Fair	Cyclic	299	58	283	5.7	6.3
14	Good	Random	288	277	506	6.8	19.7
15	Fair	Cyclic	69	38	1100	10.3	0.9

stream control devices often have the effect of regulating this arrival pattern in some way. The type of control and the distance upstream to this control device for traffic arriving on the conflicting approaches from both the left and the right of the subject approach and the resulting observed platoon patterns were noted.

Flow Rate Data

Traffic flow rate data were collected for each intersection approach at the nine sample sites. Intersection flow rates varied from about 400 veh/hr at Site 2 to 1,200 veh/hr at Sites 3 and 15.

Delay Data

Queue and service times were measured for each minor-street subject vehicle. High mean delays, about 25 sec/vehicle, were observed at Sites 9, 11, and 14. The lowest delays, about 11 and 12 sec, were measured at Sites 1, 3, 12, and 15. The distribution of the queue time and service time components varies widely among the sites.

Major-Street Gap Data

Table 2 presents several important gap data for the nine sample sites. The critical gap was calculated as the gap size at which the percentage of accepted gaps is equal to the percentage of rejected gaps. The critical gap ranges from 5.5 to 7.5 sec. Data from Table 10–2 of the HCM for major streets with 30-mph speed limits and single lanes on each approach exhibit a range for the critical gap of 5.0 to 6.5 sec. Thus the data collected in this study compare favorably with ranges suggested in the HCM.

The zero gap was calculated as the highest measured gap that was rejected by all drivers.

Accepted and Rejected Gap Data

All gaps seen by each subject vehicle were measured. Each gap was classified into one of eight categories. The first gap observed by a subject vehicle is a lag. All others are gaps. Accepted gaps or lags were used only by a subject vehicle. Shared gaps or lags were used jointly by the subject vehicle and one or more opposing vehicles. Used gaps or lags were used by an opposing vehicle and were thus unavailable to the subject vehicle. Rejected gaps or lags were not used by the subject vehicle or any opposing vehicle. Table 2 presents the mean values for each of these gap categories for each site.

Several points can be made about the data presented in Table 2. First, expected differences between gap categories are supported by the data. For example, gaps or lags shared by more than one vehicle are larger than gaps that are accepted or used by only one vehicle. Also, rejected gaps or lags are smaller than accepted gaps or lags. Second, the large values of accepted gaps or lags (all mean values are greater than 11 sec) support the need to develop adjusted values for accepted gaps, as described earlier. Gaps greater than some value (suggested as 12 sec) do not provide any useful information in the analysis of the gap-capacity relationship.

IDENTIFICATION OF BASIC RELATIONSHIPS—MACROSCOPIC ANALYSIS

The second objective is to identify some of the factors that influence delay and capacity. In this section, mean values for several variables from each site are compared to determine, on a macroscopic basis, those factors that influence delay and capacity on the minor-street approach of a TWSC intersection.

Vehicle Delay

Stopped vehicle delay data were collected by measuring two components, queue time and service time. Plots of service time versus both conflicting flow and subject flow are shown in Figures 1 and 2. These flow rates are the measured macroscopic flow rates. Plots of queue time versus conflicting flow and subject flow are shown in Figures 3 and 4, respectively.

Service Time

Service time, or the time spent waiting at the stop line, is affected most directly by the rate of flow on the major street. As major-street flow increases, all other factors being equal, the service time should also be expected to increase. The data shown in Figure 1 indicate a strong correlation between service time and major-street flow rate. If a linear relationship is assumed, the coefficient of determination (R^2) is 0.94.

Figure 2 shows a slight decline in service time with increasing subject approach flow. One supposition that may be made

TABLE 2 GAP DATA

Sample Site	Critical Gap	Zero Gap	Mean Gap	Accepted Gap	Shared Gap	Accepted Lag	Shared Lag	Used Lag	Rejected Lag	Avg Used Gap	Avg Rejected Gap
1	5.5	4.5	5.9	13.8	21.5	18.1	24.2	5.4	3.4	4.9	2.8
2	7.5	4.5	10.2	12.0	20.3	15.6	26.8	11.1	3.8	7.1	3.3
3	6.5	6.5	4.0	12.4	14.1	17.2	0.0	6.1	2.0	3.0	3.4
8	6.5	3.5	12.5	13.0	22.6	17.2	29.1	6.7	2.8	8.6	3.0
9	5.5	3.0	7.8	20.9	22.2	19.7	25.5	8.4	3.1	8.1	2.7
11	6.0	3.0	7.7	15.1	24.4	21.6	25.9	8.0	2.5	9.6	2.5
12	6.5	3.5	13.5	14.8	26.6	17.8	26.2	6.1	3.3	11.8	3.2
14	6.5	3.5	7.3	16.4	21.8	19.4	21.7	7.9	2.5	8.9	2.4
15	5.5	3.5	4.0	11.0	11.1	11.6	8.7	6.5	3.0	3.9	2.9

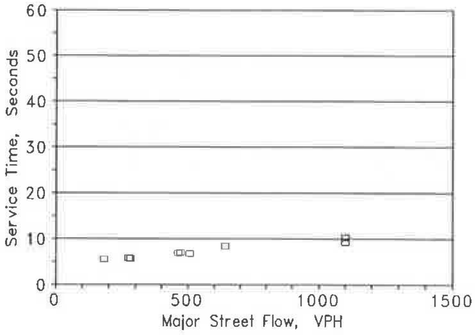


FIGURE 1 Service time versus major-street flow rate.

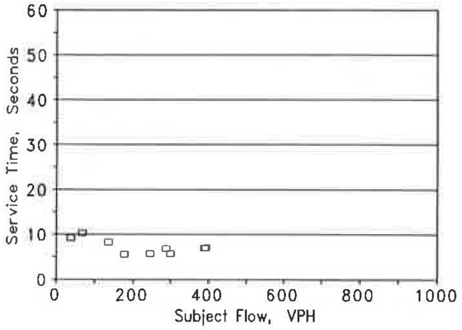


FIGURE 2 Service time versus subject approach flow rate.

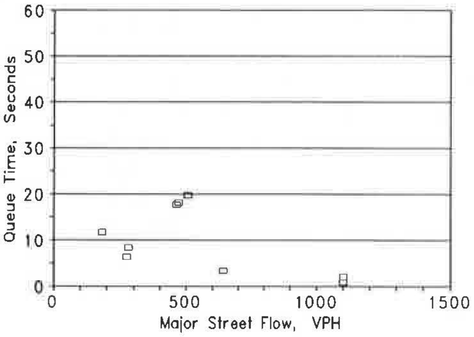


FIGURE 3 Queue time versus major-street flow rate.

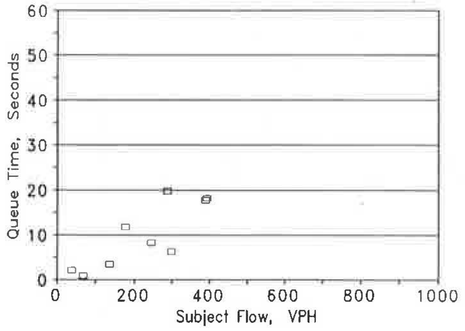


FIGURE 4 Queue time versus subject approach flow rate.

from the data in Figure 2 is that the pressure from increasing subject approach flow rates results in drivers accepting smaller major-street gaps and thus experiencing lower service times.

Queue Time

Queue time, or the time spent waiting in queue until arriving at the stop line, is a function of both the number of vehicles that are waiting in the queue and the rate at which the vehicle at the stop line is able to depart. The data shown in Figure 4 indicate a good correlation between queue time and the subject approach flow rate. Although other variables such as arrival patterns may also be important, the subject approach flow rate may be a good indicator for the number of vehicles likely to be in a queue on the minor street. If a linear function is assumed, the coefficient of determination (R^2) for this relationship is 0.70.

The data shown in Figure 3 do not, at first glance, support the contention that queue time is affected by major-street flow. However, a more detailed analysis may, in fact, bear out this supposition. If, for example, the three points in the lower right quadrant of Figure 3 are eliminated from the plot, the data indicate that queue time does increase with service time. The elimination of these data may in fact be justified because these three data points (sample Sites 1, 3, and 15) are from sites with the lowest subject approach flow rates (less than 150 veh/hr). Because these are cases in which subject approach queuing is low or nonexistent, it would not be expected that a variation in the major-street flow rate would have any effect on the queue time. However, the limited number of data points prevents a definitive conclusion at this time.

Minor-Street Capacity

Minor-street (subject approach) capacity was calculated using Equation 2 and the service time for each subject approach vehicle.

$$\text{Capacity} = \frac{3{,}600}{\text{Service Time}} \qquad (2)$$

Figure 5 shows a strong linear correlation between the minor-street capacity and the major-street (conflicting approaches)

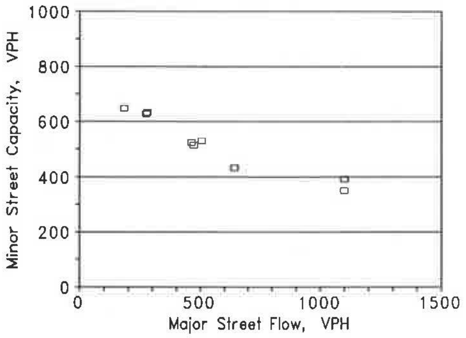


FIGURE 5 Minor-street capacity versus major-street flow rate.

flow rate. Although other variables are likely to affect capacity (e.g., major-street approach speed and turning proportions), the coefficient of determination (0.91) indicates the importance of major-street flow rate.

Figure 6 shows a plot of accepted gap versus major-street flow rate. The accepted gap data shown in this figure have not been modified as described earlier. One possible hypothesis that can be suggested on the basis of Figure 6 is that the size of the accepted gap first increases as major-street flow increases, and then decreases. The change point may represent the point at which a driver's frustration level caused by the delays experienced at higher major-street flow rates begins to affect the decision-making process.

IDENTIFICATION OF BASIC RELATIONSHIPS—MICROSCOPIC ANALYSIS

In order to further meet the second objective, a microscopic analysis of the data was undertaken. In this microscopic analysis, the conditions faced by each subject approach vehicle were taken into account; that is, consideration was given to the major-street flow rate as seen by the driver, the gaps that were presented, and so on. Several other points should be made about this analysis.

1. The gap acceptance and delay data for each minor-street subject vehicle were calculated and the gaps were classified according to the definitions described earlier.
2. Only those cases in which a subject vehicle did not face any opposing approach vehicles were considered (in effect simulating the operation of a one-way minor street crossing a two-way major street).

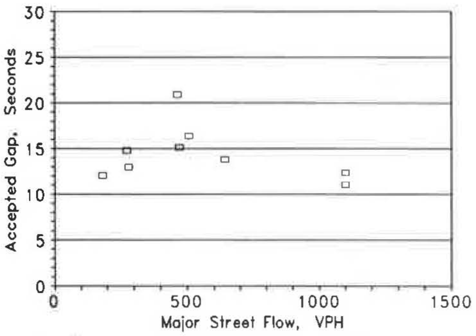


FIGURE 6 Accepted gap versus major-street flow rate.

3. The microscopic major-street flow rate and the accepted gap were calculated as described earlier.
4. Not all of the sample sites are shown in every assessment: a sample site is only shown if all cells for a particular table had at least 10 observations.

Accepted Gap

Four assessments of accepted gap are presented here. These assessments are intended to identify whether the following variables influence the size of the gap accepted by a driver: whether a vehicle waited in queue or not, the major-street flow rate, the queue time, the service time, and movement direction of the subject approach vehicle.

Several conclusions may be stated from the following assessments. The most important is that the accepted gap is not a constant, but varies as a result of several factors: queue time, service time, number of gaps that have been rejected, major street flow, and directional movement of the subject vehicle.

Accepted Gap as a Function of Number of Rejected Gaps and Presence of Queue

The assessment presented in Table 3 indicates that most vehicles that wait in a queue accept shorter gaps than vehicles that do not wait in a queue. Table 3 also indicates that most vehicles that reject one or more gaps eventually accept a shorter gap than vehicles that accept the first gap. One possible inference is that the longer a vehicle waits, the greater the driver's frustration, and a gap that previously seemed too short now becomes acceptable. Another possible inference is that a learning process takes place as the driver waits at the intersection. The longer the wait, the better a driver can estimate the size of an acceptable gap. Both of these inferences are tentative conclusions, and they should receive further study.

Accepted Gap as a Function of Major Street Flow

Table 4 indicates that the size of the accepted gap decreases as the microscopic major-street flow rate increases. With increasing major-street flow rates, the size of the available gaps

TABLE 3 ACCEPTED GAPS, NUMBER OF REJECTED GAPS, AND QUEUE TIMES

	Accepted Gap, Seconds			
	Zero Gaps Rejected		One or More Gaps Rejected	
Sample Site	No Queue	Queue	No Queue	Queue
1	11.3	10.8	10.2	8.9
2	10.1	10.0	9.0	8.4
3	-	-	-	-
8	10.0	9.6	9.5	10.5
9	10.8	10.3	9.7	9.4
11	10.7	9.4	10.5	9.2
12	10.9	10.2	9.5	9.4
14	10.8	9.7	9.9	9.2
15	-	-	-	-

TABLE 4 ACCEPTED GAPS AND MICROSCOPIC (INSTANTANEOUS) MAJOR-STREET FLOW RATES

Sample Site	Accepted Gap, Seconds			
	Major St Flow 0-300 vph	Major St Flow 300-600 vph	Major St Flow 600-900 vph	Major St Flow >900 vph
1	12.0	10.2	10.5	6.8
2	12.0	9.6	8.3	6.1
3	12.0	10.5	9.0	8.1
8	12.0	11.6	9.0	6.9
9	12.0	10.6	7.8	6.9
11	12.0	11.0	8.8	6.8
12	12.0	10.9	7.5	6.1
14	12.0	11.4	7.7	5.6
15	-	11.7	8.9	7.7

becomes smaller. But these results may also mean that when faced with higher flow rates, drivers may anticipate the need to accept what is available, even if the available gap size is smaller than desired.

Accepted Gap as a Function of Service Time

Table 5 indicates a slight decrease in the size of the accepted gap as the service time increases. More data are needed to determine if these results are statistically significant. Again, the trend may point to two factors discussed earlier: the accepted gap may be influenced by driver frustration level or by learning, or both. Both factors directly relate to the length of time spent at the intersection stop line (service time).

Accepted Lag as a Function of Directional Movement

Table 6 indicates the importance of the directional movement of the minor-street vehicle and the size of the lag that is accepted by the subject vehicle. Because a through-vehicle must only clear the intersection, the size of the required lag is smaller than for a turning vehicle, which must also merge into the conflicting-vehicle stream. The data presented in Table 6 seem to confirm this assertion.

Delay

Two assessments of vehicle delay are presented here. Both assessments tend to confirm the conclusions reached in the macroscopic analysis.

Service Time as Function of Major-Street Flow

Table 7 indicates that, for most sites, service time and total delay increase as major-street flow rates increase. This relationship was also observed in the macroscopic analysis.

Total Delay as a Function of Major-Street Flow

Table 8 indicates that, for most sites, total delay increases as major-street flow rates increase.

DEVELOPMENT OF PRELIMINARY MODELS FOR CAPACITY AND DELAY

On the basis of the analysis presented earlier, a set of equations has been developed that indicates some important fac-

TABLE 5 ACCEPTED GAPS AND SERVICE TIMES

Sample Sites	Accepted Gap, Seconds		
	Service Time, 0-5 sec	Service Time, 5-10 sec	Service Time, > 10 sec
3rd/Hayes (Sites 2, 8, 12)	10.1	9.9	9.7
29th/Freya (Sites 9, 11, 14)	9.9	9.7	9.6

TABLE 6 ACCEPTED LAGS AND DIRECTIONAL MOVEMENT

Sample Sites	Accepted Lag, Seconds			
	Left Turns	Throughs	Right Turns	All
2	10.0	8.7	10.9	10.0
8	10.0	9.2	10.6	9.8
12	10.7	10.0	11.1	10.6
9	-	9.6	11.2	10.4
11	-	9.1	10.9	9.9
14	-	8.8	11.1	10.0

TABLE 7 SERVICE TIMES AND MAJOR-STREET FLOW RATE

Sample Sites	Service Time, Seconds		
	Major St Flow < 300 vph	Major St Flow 300-600 vph	Major St Flow > 600 vph
2	3.8	4.1	8.0
8	3.9	4.6	4.8
12	4.2	5.0	6.7
9	3.3	4.5	6.5
11	3.8	3.8	5.2
14	4.1	3.8	5.7

TABLE 8 TOTAL DELAYS AND MAJOR-STREET FLOW RATE

Sample Sites	Total Delay, Seconds		
	Major St Flow 0-300 vph	Major St Flow 300-600 vph	Major St Flow > 600 vph
1	4.7	9.5	13.5
2	13.7	14.9	23.0
3	9.5	9.1	11.1
8	10.0	13.0	14.5
9	16.4	16.2	21.4
11	17.1	19.1	23.9
12	8.4	9.6	10.2
14	15.4	24.6	25.9
15	7.1	6.2	13.1

tors that affect minor-street capacity, total minor-street delay, queue time, and service time.

Minor-Street Capacity

A correlation analysis was undertaken to identify the factors that affect minor-street capacity. The strongest linear correlations are between minor-street capacity and the volume on the conflicting approaches ($R^2 = 0.91$), the critical gap ($R^2 = 0.42$), the mean gap ($R^2 = 0.88$), and the volume on the subject approach ($R^2 = 0.31$).

Equation 3 shows the best model resulting from a regression analysis with minor-street capacity as the dependent variable.

$$C = 687 + 0.307q_c \qquad (3)$$

The model of Equation 3 has been compared with capacity models estimated in two other countries. Equation 4 was developed from simulation data in Poland by Marion Tracz (*10*). Equation 5 was developed by the Transportation Road Research Laboratory in the United Kingdom (*11*). There is a strong similarity between the three equations in both intercept and slope, indicating that the capacity model developed with data from this current study compares favorably with models developed in Poland and the United Kingdom.

$$C = 887 - 0.547q_c \qquad (4)$$

$$C = 675 - 0.400q_c \qquad (5)$$

Total Minor-Street Delay

A correlation analysis was undertaken to identify the factors that affect total minor-street delay. The strongest linear correlations are between total minor-street delay and the volume on the subject approach ($R^2 = 0.60$), the volume on the opposing approach ($R^2 = 0.83$), and the minor-street approach volume-to-capacity ratio ($R^2 = 0.68$).

Equation 6 indicates the best model resulting from a regression analysis with total minor-street delay as the dependent variable and the volumes on the subject and opposing approaches as the independent variables.

$$d_t = 8.0 + 0.0153q_s + 0.0505q_0 \qquad (6)$$

Minor-Street Queue Time

A correlation analysis was undertaken to identify the factors that affect minor-street queue time. The strongest linear correlations are between minor-street queue time and the volume on the subject approach ($R^2 = 0.70$), the volume on the conflicting approaches ($R^2 = 0.30$), the volume on the opposing approach ($R^2 = 0.74$), and the volume-to-capacity ratio ($R^2 = 0.70$).

Equation 7 indicates the best model resulting from a regression analysis with minor-street queue time as the dependent variable and the volumes on the subject and opposing approaches as the independent variables.

$$d_q = 0.0428q_0 + 0.0245q_s \qquad (7)$$

Minor-Street Service Time

A correlation analysis was undertaken to identify the factors that affect minor-street service time. The strongest linear correlations are between minor-street service time and volume on the conflicting approaches ($R^2 = 0.94$), and the volume on the subject approach ($R^2 = 0.43$).

Equation 8 indicates the best model resulting from a regression analysis with minor-street service time as the dependent variable and volume on the conflicting approaches as the independent variable.

$$d_s = 0.0048q_c \tag{8}$$

SUMMARY AND CONCLUSIONS

The first objective was to develop a data base for traffic operations at TWSC intersections. To meet this objective, nine TWSC sample sites were studied covering 13 hr of traffic operations. A total of 970 minor-street subject vehicles and nearly 2,000 gaps that were accepted or rejected were studied.

The second objective was to identify some of the factors that affect delay and capacity. Several conclusions can now be stated on the basis of the macroscopic and microscopic analyses.

When the average delay and flow rate data for each of the nine sites are compared, several general delay and capacity relationships can be identified.

1. The average time spent waiting in queue (queue time) increases as the subject approach flow rate increases.
2. The average time spent waiting at the stop line (service time) increases as the flow rates on the conflicting approaches increase.
3. The minor-street capacity decreases as the major-street flow rate increases.

A microscopic analysis of the behavior of individual minor-street vehicles, effectively operating as a one-way minor street crossing a two-way major street, confirms some of these same conclusions and suggests several additional ones:

1. The mean accepted gap tends to be lower for drivers who have waited in a queue than for drivers who have not waited in a queue.
2. The mean accepted gap tends to decrease as the major-street flow rate increases.
3. The mean accepted gap tends to decrease as the queue time or service time increases.
4. The mean accepted gap is lower for through vehicles than for turning vehicles.
5. The time in queue is not correlated to the major-street flow rates.
6. Service time and total stopped delay increase as the major-street flow rates increase.

The third objective was to develop a set of equations to forecast delay and capacity. The equations for forecasting minor-street capacity depend on traffic volumes on the conflicting and opposing approaches. Total delay on the minor street (subject approach) depends on traffic volumes or the volume-to-capacity ratios on the subject and opposing approach. Queue time depends on traffic volumes or volume/capacity ratios on the subject and opposing approaches. Service time depends on the traffic volume on the conflicting (major-street) approaches.

ACKNOWLEDGMENTS

Financial support for this study was provided by the University of Idaho, the Idaho Transportation Department, and TRANSNOW. The assistance of Robert Smith, research engineer for the Idaho Transportation Department, and Dan McComb, Elmer Kassens, and Scott Guernsey of the Idaho Transportation Department Traffic Section was greatly appreciated.

Acknowledgment is also gratefully given to Wayne Kittelson, for his new and challenging ideas in trying to solve a difficult problem; to Martha Ford, for her substantial contributions in the development of this paper; and to Joseph Schacher, for his assistance in the data collection and reduction.

REFERENCES

1. W. Brilon (ed.). *Intersections Without Traffic Signals*. Springer-Verlag, Berlin, 1988.
2. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
3. *Transportation Research Circular 319: Research Problem Statements*. TRB, National Research Council, Washington, D.C., June 1987.
4. W. Brilon. Recent Developments in Calculation Methods for Unsignalized Intersections in West Germany. In *Intersections Without Traffic Signals*, Springer-Verlag, Berlin, 1988.
5. J. Anveden. Swedish Research on Unsignalized Intersection. In *Intersections Without Traffic Signals*, Springer-Verlag, Berlin, 1988.
6. J. D. Zegeer. Status of Unsignalized Intersection Capacity Research in the United States. In *Intersections Without Traffic Signals*, Springer-Verlag, Berlin, 1988.
7. W. Brilon and M. Grossman. Recommendation for a New Computational Procedure in Unsignalized Intersection Design. Lehrstuhl fur Verkehrswesen, Ruhr Universitat Bochum, July 1990.
8. R. M. Kimber and R. D. Coombe. *The Traffic Capacity of Major/Minor Priority Junctions*. Supplementary Report 582, Transport and Road Research Laboratory, Crowthorne, England, 1980.
9. M. Kyte and A. Boesen. Traffic Data Input Program, Program Documentation and User's Manual, Version 3.0, Department of Civil Engineering, University of Idaho, Moscow, 1991.
10. M. Tracz. Research on Traffic Performance of Major/Minor Priority Intersections. *Intersections Without Traffic Signals*, Springer-Verlag, Berlin, 1988.
11. R. M. Kimber. The Design of Unsignalized Intersections in the UK. *Intersections Without Traffic Signals*, Springer-Verlag, Berlin, 1988.

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

Toward a New German Guideline for Capacity of Unsignalized Intersections

WERNER BRILON, MICHAEL GROSSMANN, AND BIRGIT STUWE

The upcoming German guideline for calculating the capacity of unsignalized intersections is described. Information is included on fundamental aspects of the theory as well as on the proposed procedure that results from several research projects carried out on behalf of the West German Federal Minister of Transport. The new guideline will be a revision of the German procedure introduced in 1972, which has been transferred as Chapter 10 of the 1985 *Highway Capacity Manual* (HCM). Furthermore, a selection of formulas for roundabout capacity is given. Capacity and delays of unsignalized intersections have been investigated in Germany in a long series of research projects during the 1960s and 1970s. These results led to the German guideline of 1972, which was mainly based on Harders' pioneering work. The problems with the 1972 procedure and Chapter 10 of the HCM can be summarized as follows: (a) differences to the simple Poisson case, (b) concept and size of critical gaps, (c) determination of impedance factors, (d) double introduction of Rank 2 and Rank 3 traffic streams, and (e) delays and level of service. After the publication of the 1985 HCM, the problem of capacity at unsignalized intersections received new attention in Germany. A series of research projects was started, the aim of which was to develop a new guideline for practical application. The results of these investigations will be introduced into the German committee's discussions within a short time.

The upcoming German guideline for calculating the capacity of unsignalized intersections is described. Information is included on fundamental aspects of the theory as well as on the proposed procedure that results from several research projects carried out on behalf of the West German Federal Minister of Transport. The new guideline will be a revision of the German procedure introduced in 1972 (*1*), which has been transferred as Chapter 10 of the 1985 *Highway Capacity Manual* (HCM) (*2*). Furthermore, a selection of formulas for roundabout capacity is given. Capacity and delays of unsignalized intersections have been investigated in Germany in a long series of research projects during the 1960s and 1970s, a survey of which was given by Brilon (*3,4*). These results led to the German guideline of 1972, which was mainly based on Harders' pioneering work (*5*). The problems with the 1972 procedure and Chapter 10 of the HCM can be summarized as follows: (a) differences to the simple Poisson case, (b) concept and size of critical gaps, (c) determination of impedance factors, (d) double introduction of Rank 2 and Rank 3 traffic streams, and (e) delays and level of service. After the publication of the 1985 HCM, the problem of capacity at unsignalized intersections received new attention in Germany. A series of research projects was started, the aim of which was to develop a new guideline for practical application. The results of these investigations will be introduced into the German committee's discussions within a short time.

HANDLING THE DIFFERENCES BETWEEN ACTUAL TRAFFIC STREAMS AND THE SIMPLE POISSON MODEL

The former versions of guidelines (*1,2*) are based on a simple Poisson model for each of the traffic streams at an intersection. Moreover, these procedures are only valid for streams with non-time-dependent traffic volumes. Siegloch (*6*) has studied the influence of non-Poisson properties of traffic streams on the main road. He found that more realistic headway distributions could change the capacity of the intersection considerably. On the one hand, slight disturbances of the headway distribution decrease the capacity. However, normal or heavy disturbances (i.e., platooned traffic) can increase capacity considerably, for example, up to 1.2 times the Poisson case [see Figure 4 in Brilon (*3*)]. These effects have been confirmed by several authors [see Zhang (*7*) and Chodur et al. (*8*)], but Siegloch's (*6*) figures and tables still seem to be the most extensive documentation of this phenomenon. Nevertheless, under average conditions and realistic distributions of headways in the priority stream, the capacity is similar to the one calculated with the help of a simple Poisson model.

Therefore, the simple Poisson case will still be the basis of the future German guideline. The intention is to use Siegloch's capacity formula to describe the maximum number of minor-stream vehicles that can pass through a priority traffic stream of a given volume:

$$G = \frac{3,600}{t_f} \cdot e^{-pt_0} \tag{1}$$

where

G = capacity of the minor stream [passenger car units per hour (pcu/hr)];
p = $q_p/3,600$;
q_p = volume of the priority stream [vehicles per hour (veh/hr)];
t_0 = $t_g - t_f/2$;
t_g = critical gap (minimum headway between vehicles in the main street traffic streams that enables at least one vehicle in a minor traffic stream to pass through the intersection); and
t_f = move-up time (minimum headway between vehicles in a minor traffic stream entering into the same main stream gap).

Lehrstuhl für Verkehrswesen, Ruhr Universität Bochum, Universitätsstrasse 150, 4630 Bochum 1, Germany.

The results of the formula are similar to Harders' (*5*) original formula

$$G = q_p \frac{e^{-p(t_g - t_f)}}{e^{p t_f} - 1} \qquad (2)$$

which was the basis of the Chapter 10 procedure. The differences between Equations 1 and 2 are within a margin of 10 pcu/hr. Results of Equation 1 are shown by Figure 1. In the final version of the new guideline, a series of four similar figures will be included for left-turning traffic from the priority road, right-turning traffic from the minor road, through movements, and left-turning traffic from the minor road.

The German guideline only applies to single lanes for each movement. According to other German guidelines, unsignalized intersections in multilane streets are to be avoided because of traffic safety.

CONCEPT AND SIZE OF CRITICAL GAPS AND MOVE-UP TIMES

In a recent publication, Kimber (*9*) reiterates his former argument that the concept of mathematical models with gap acceptance theory leads to unrealistic estimations of capacity. He has developed an empirical approach by using regression techniques. These ideas, of course, should be considered and evaluated carefully and comparisons should be begun as soon as possible.

However, further study over a number of years is required to develop this empirically supported concept as an alternative to the gap acceptance theory. Moreover, there is the problem of transferring the results to the situation in another country and the tremendous expenditure for comprehensive empirical evaluations. Therefore, the next German guideline for unsignalized intersections will still be based on gap acceptance theory.

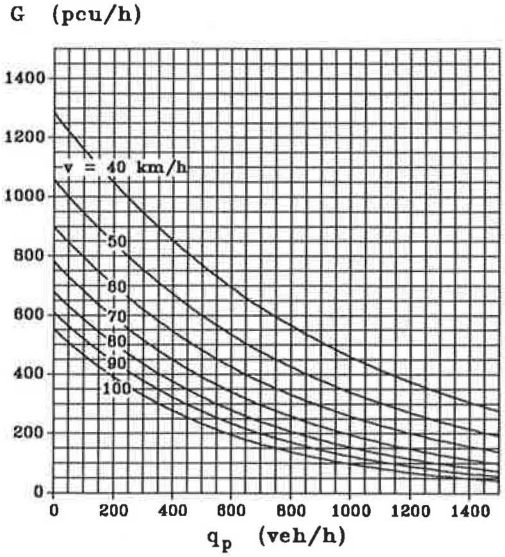


FIGURE 1 Capacity calculated from Equation 1 in relation to priority volume q_p and average main-road velocity v_m for through traffic on the minor road.

The only empirical results for representative critical gaps and move-up times under German conditions are those of Harders (*10*) (see Table 1), described in English by Brilon (*3*). This data base might have become obsolete and driver behavior could have changed in the meantime. Another objection is that the influence of the velocities on the major street might be overestimated in these values. Nevertheless, Harders' results are still the most recent and above all the most comprehensive values for critical gaps and move-up times. During the study, 33 intersections were observed for more than 109 hr, resulting in 30,000 measured values. Moreover, some sample comparisons conducted with priority street velocities of 50 km/hr confirmed Harders' results. These values characterize driver behavior in West Germany in 1976. Before applying the formulas in other countries, these parameters must certainly be evaluated again.

Another influence on capacity is caused by the variations in critical gaps and move-up times among drivers and even for one driver over time. This effect has been studied, for example, by Ashworth (*11*). Theoretically, he found a relationship between reduction in minor-road capacity and variance of critical gap distribution for different levels of major-road volume. The reduction in capacity can vary up to 40 percent depending on the other parameters. On the other hand, a recent research project (*12*) has found that this effect is compensated if at the same time a more realistic headway distribution is used.

DETERMINATION OF IMPEDANCE FACTORS

At an intersection, different ranks of priority can be assigned on the basis of highway code regulations. Rank 1 is defined for movements with priority over all other conflicting movements. Each traffic stream (movement) that has to give priority to another movement of Rank *r* is at least of Rank *r* + 1 (see Figure 2).

Impedance factors in the former guidelines included this hierarchy for the different movements at an intersection [for more details, see Chapter 10 in the HCM (*2*)]. Originally, in the basic queuing model, the impedance factor is the probability that no vehicle is queuing in the minor traffic stream. It is a fundamental property of queuing systems with one service channel that

$$p_0 = 1 - a \qquad (3)$$

TABLE 1 CRITICAL GAPS t_g AND MOVE-UP TIMES t_f FOR PASSENGER CARS IN WEST GERMANY (*10*)

Type of minor stream		Average velocity v_m on the major road						
		40	50	60	70	80	90	km/h
Left turns from the major road	t_g t_f	4.5 1.7	5.2 2.1	5.8 2.5	6.5 2.8	7.1 3.2	7.8 3.6	s s
Right turns from the minor road	t_g t_f	5.0 2.1	5.8 2.6	6.5 3.1	7.2 3.6	7.9 4.1	8.7 4.5	s s
Through traffic on the minor road	t_g t_f	5.1 2.8	5.8 3.4	6.5 4.0	7.3 4.6	8.0 5.3	8.7 5.9	s s
Left turns from the minor road	t_g t_f	5.6 2.7	6.4 3.3	7.2 3.9	8.0 4.5	8.8 5.1	9.6 5.7	s s

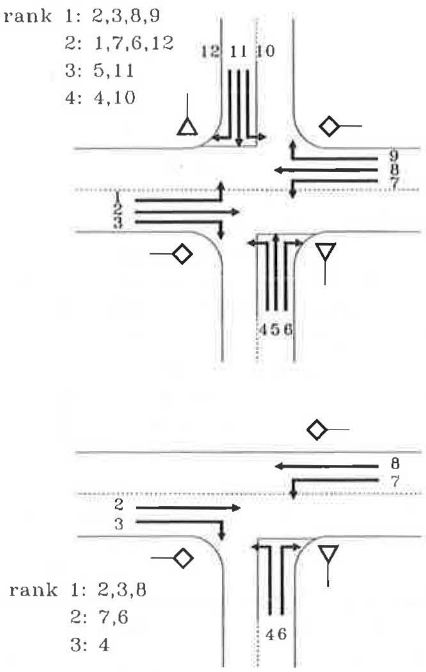


FIGURE 2 Traffic streams and their rank of priority.

where

p_0 = probability of the empty system,
a = utilization factor ($= q_n/G$),
G = capacity of the minor stream (pcu/hr), and
q_n = volume of the minor stream (pcu/hr).

This regularity has already been pointed out by Siegloch (*6*). The original curve for p_0 in the 1972 German procedure (*1*) and the HCM (*2*)—which is wrong—was based on approximations from Harders (*5*). Equation 3 can be applied for the probability of no queuing in movements of Rank 2. Impedance factors in different streams of Rank 2 can be multiplied as proposed by Figure 10–4 in Chapter 10 of the HCM (*2*) or Table 5 in Brilon (*3*), because queuing in these streams is independent of each other, which has also been proven by extensive computer simulations (*12*).

INTRODUCTION OF RANK 3 AND RANK 4 TRAFFIC STREAMS

In former publications, several authors objected that traffic streams of Ranks 2 and 3 are introduced twice into the calculations according to the German (*1*) and American (*2*) procedures. These streams are on the one hand added to the main-street volumes [see Figure 10–2 in the HCM (*2*)]. On the other hand, they are introduced into the impedance factors by which the basic capacities G of streams of Ranks 3 and 4 are diminished to calculate the actual capacities L by

$$L = p_0 G \tag{4}$$

Brilon (*3*) has discussed arguments that support this double introduction (see his Figures 7 and 8).

The reasons for this are as follows:

- During times of queuing in Rank 2 streams (e.g., left turners from the priority street) the Rank 3 vehicles (e.g., left turners from the minor street at a T-junction) cannot enter the intersection because of traffic regulations by highway code. Because the portion of time provided for Rank 3 vehicles is p_0, the basic capacity calculated from Equation 1 for Rank 3 streams has to be diminished by the factor p_0 for the corresponding Rank 2 streams (Equations 3 and 4).
- Even if no Rank 2 vehicle is queuing, these vehicles influence Rank 3 operations, because a Rank 2 vehicle approaching the intersection within a time of less than t_g prevents a Rank 3 vehicle from entering the intersection.

A recent research project (*12*) has proven that among the possibilities considered, the existing approach (Equation 4) for Rank 3 operations is the best one to account for these relations. For Rank 4, however, a minor change of the procedure proved to be necessary, because the probabilities $p_{0,r3}$ (= no vehicle is queuing in Rank 3 streams) depend on $p_{0,r2}$ for Rank 2 movements. This relationship is shown in Figure 3. The diagonal would represent the case of statistical independence between the impedance factors of Rank 2 and Rank 3 movements. This assumption of independence, as the basis of the procedures in the HCM (*2*), underestimated the actual capacity L. To enable an easy consideration of statistical interrelations between $p_{0,r2}$ and $p_{0,r3}$, the solid line in Figure 3 has been developed. It is a representation of simulation results with more than 300 different combinations of traffic flow volumes. The value $p_{0,x}$ represents the impedance factor $p_{0,x}$, which has to be applied instead of the product $p_{0,r2} \cdot p_{0,r3}$. Again, this new factor is only necessary for the computation of Rank 4 capacities.

DELAYS AND LEVEL OF SERVICE

The existing procedures do not clearly distinguish between different level-of-service (LOS) values. For the German

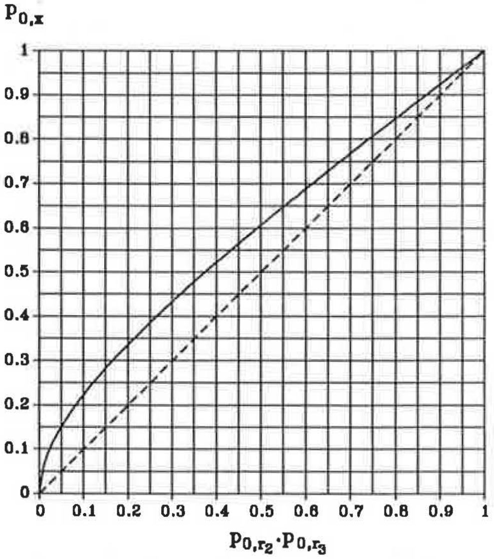


FIGURE 3 Corrected impedance factor $p_{0,x}$ for evaluating actual capacity of a Rank 4 stream.

guideline, this argument is of minor importance, because the German concept is more aimed at capacities and the direct evaluation of LOS parameters like delays.

To estimate delays, a series of formulas for the simple case of one priority stream and one minor stream has been developed. From the theoretical point of view, Kremser's formulas—as presented by Brilon (*3*)—can be used for steady state conditions. Assuming that t_g and t_f are constant values and that the headways are exponentially distributed both in the major and minor stream, Kremser (*13,14*) derived equations that can be simplified as

$$D = \frac{E(W_1)}{x} + \frac{n}{2} \cdot \frac{y \cdot E(W_1^2) + z \cdot E(W_2^2)}{xy} \qquad (5)$$

$$E(W_1) = \frac{1}{p} \cdot (e^{pt_g} - 1 - pt_g) + d$$

$$E(W_1^2) = \frac{2}{p} \cdot (e^{pt_g} - 1 - pt_g) \cdot \left(\frac{e^{pt_g}}{p} + d - t_g\right) + d^2 + t_g^2 \qquad (6)$$

$$E(W_2) = \frac{e^{pt_g}}{p} \cdot (1 - e^{pt_f})$$

$$E(W_2^2) = \frac{2e^{pt_g}}{p^2} \cdot [(e^{pt_g} - pt_g) \cdot (1 - e^{-pt_f}) - pt_f e^{pt_f}] \qquad (7)$$

where

D = average delay (s) for minor street vehicles,
$x = y + z$,
$y = 1 - nE(W_2)$,
$z = nE(W_1)$,
W_1 = delay for a minor vehicle arriving at the intersection when no other vehicle is queuing,
W_2 = waiting time in the first queuing position for those vehicles that have arrived in a higher position of the queue,
$E(W_i)$ = first moment of W_i,
$E(W_i^2)$ = second moment of W_i,
d = time needed for orientation after arriving at the first position of the queue $\approx t_f$,
$n = q_n/3{,}600$, and
q_n = volume of the minor stream.

With constant values for t_g and t_f, D is a function of the volumes q_p and q_n. Even this formula (Equation 5), which gives the impression of being rather complicated, is only an approximation. It is valid only for the unrealistic case of $t_g = t_f$. Poeschl (*15*) provides more exact but also more complicated formulas for $E(W_1)$ and $E(W_1^2)$. (His formulas are much too complicated for practical application.)

Therefore, the delay formula of Harders (*5*) can be recommended as a good approximation:

$$D = \frac{1 - \gamma}{L - q_n} \cdot 3{,}600 \qquad (8)$$

where

$$\gamma = e^{-(pt_g + nt_f)}$$

These formulas only apply for the simple queuing system with one major stream (priority stream) and one minor stream (e.g., Streams 1, 6, 7, and 12). A correct formula for other movements of Ranks 3 and 4 or for more realistic traffic flow assumptions is not available. For these cases, computer simulation techniques seem to be the only practicable tool.

Irrespective of the delay formula used, the so-called "Harders' trick" is true. This means that with constant reserve capacity R,

$$R = L - q_n \qquad (9)$$

[see Equation 10–2 in the HCM (*2*)], average delay D tends to remain on a nearly constant level (see Figure 4). Hence, with R greater than 100 pcu/hr the average delay of minor vehicles D is less than 35 sec. Therefore, the future German guideline will use this limit of $R \geq 100$ pcu/hr as the limit of practical capacity for minor traffic streams at unsignalized intersections. Recently, Brilon and Grossmann (*12*) also proved that Harders' trick—in principle—is also valid for traffic streams of Rank 3 and 4. In these cases, however, the average delay becomes insignificantly higher but remains below 45 sec (for R = 100 pcu/hr). In the final version of the new German guideline, figures similar to Figure 4 will be included for Rank 2, Rank 3, and Rank 4 streams.

On the basis of this fundamental idea, the LOS also in a future HCM could be determined from reserve capacities (Equation 9) similar to those in Table 10–3 in the HCM (*2*). The values of this table could be adjusted to the delays to be applied for different LOS values. However, the queue lengths,

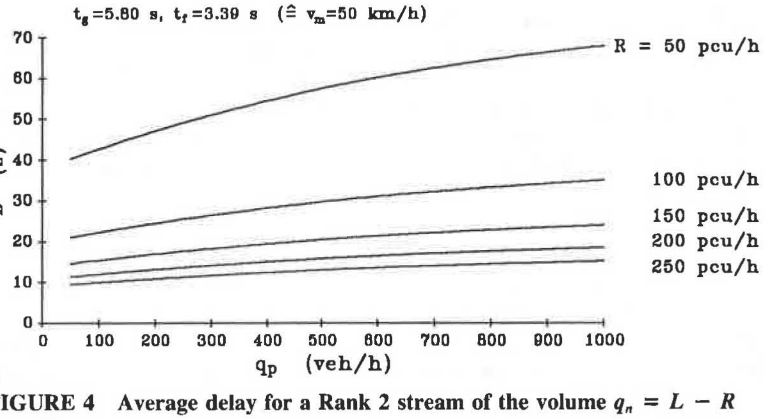


FIGURE 4 Average delay for a Rank 2 stream of the volume $q_n = L - R$ for different values of q_p.

aspects of fuel consumption, emission of pollutants, and difference between a specific traffic situation and total saturation should also be considered. Because each of these aspects differs for unsignalized and signalized intersections, it is not necessary that delays at each LOS correspond at both types of intersections.

TIME-DEPENDENT TRAFFIC VOLUMES

During recent years, traffic performance at intersections with time-dependent traffic volumes (peak-hour effect) has been studied by several authors. Results from Stamm (*16*) and Tonke (*17*) are also discussed in Brilon (*3*). Zhang (*7*) and Chodur et al. (*8*) also obtained more recent results. In qualitative terms, these results can be summarized as follows:

- The total capacity would increase considering time-dependent priority streams;
- If, however, the peaks of traffic volumes in priority streams and minor streets coincide, capacity decreases and delays increase (*3*).

Therefore, traffic performance depends on the offset between traffic peaks on the priority street and the minor street to a high degree. It is acknowledged that details of these relations cannot be expressed by a simple paper-and-pencil procedure.

PROPOSED PROCEDURE

The new procedure, which is going to be proposed to the German guideline committees in early 1991, can be characterized by the following features:

- Each of the movements at an intersection has to be evaluated by its traffic volume. There are 12 movements at a four-way intersection and 6 at a three-way intersection (see Figure 2). Heavy vehicles in the minor streams have to be converted to pcu with the help of special conversion factors [see Table 10–1 in the HCM (*2*)].
- The hierarchy of the traffic streams regarding priority brings about the order of evaluating the capacity for each minor stream *i*. First, the basic capacity G_i will be taken from a diagram like Figure 1. Parameters are the velocity on the main street and the volume of the priority stream q_p. This priority stream is composed of different streams that are placed hierarchically over the actual minor stream [see Figure 10–2 in the HCM (*2*)].
- The actual capacity L_i for Rank 2 movements (left turners from the major road and right turners from the minor road) is equivalent to the basic capacity. For the actual capacity of streams of Rank 3 or Rank 4 traffic, the queuing of privileged streams has to be considered by reducing the basic capacity with the impedance factor p_0 of the corresponding privileged streams (for Rank 3 streams, using Equations 3 and 4; for Rank 4 streams, using Equations 3 and 4 and Figure 3).
- If there is a shared lane for more than one movement, the common capacity is calculated from the individual capacities according to the individual volumes using Equation 10–1 in the HCM (*2*).
- Because the actual capacity describes the situation of very long queues and delays, the practical capacity is estimated by the actual capacity L_i minus a reserve capacity R_i (e.g., R_i = 100 pcu/hr). Depending on this reserve, a specific LOS characterized by average delay D can be guaranteed (see Figure 4).
- An unsignalized intersection can be called efficient for a chosen LOS value, if the traffic volumes of all minor streams are less than the corresponding practical capacities: $q_i < L_i - R_i$.
- In order to get more detailed information about traffic quality regarding queues and delays, simulation technique proves to be a useful tool. Therefore, the computer program KNOSIMO (*18*), which is already in practical use, has been developed. KNOSIMO estimates expectations, standard deviations, and total distributions of delays and queue lengths for each single traffic stream at an intersection also for time-dependent traffic volumes. KNOSIMO can be operated on a normal personal computer.

ROUNDABOUTS

Roundabouts are of increasing interest to traffic planners in Germany. Reasons for this are good experiences reported from other countries, chances of a well-designed adaptation of an intersection into an urban environment, and positive expectations regarding traffic capacity.

Therefore, a series of investigations of roundabout capacity has been conducted by the Ruhr University on behalf of the Federal Minister of Transport (*19*). Several methods of investigation have been studied and experimented with. Finally, an empirical method comparable to Kimber's approach (*20*) was used.

A group of 10 roundabouts was selected for the measurements out of more than 100 unsignalized circular intersections in West Germany. Each of the roundabouts provided situations with high traffic volumes of up to more than 3,000 veh/hr. In the course of the measurements, the streets entering into the roundabout were observed with video equipment that had a digital clock inserted into each frame. Measurements were carried out during times of continuous traffic jams on the entrance roads.

From the video recordings, the number of entering vehicles and the number of circulating vehicles was counted in 1-min intervals. It was presumed that the capacity of the entrance during this interval can be derived from these observations of continuous queuing on the approaching road during a 1-min interval.

On the basis of this assumption, the observed values of capacity (q_e = maximum volume of entering traffic in pcu/hr) were compared with those of circulating traffic (q_c in pcu/hr) observed on the circular road just upstream from the entrance. The measurement points for q_c and q_e were analyzed by regression techniques. Instead of a linear approximation used by, e.g., Kimber (*20*) and Louah (*21*), an exponential regression line was used because of a better agreement with capacity formulas for unsignalized intersections, e.g., Siegloch's formula (see Equation 1). Complementary simulations proved that with some minor bias—from a theoretical point of view—this technique represents the $q_e - q_c$ relation also for 1-hr intervals.

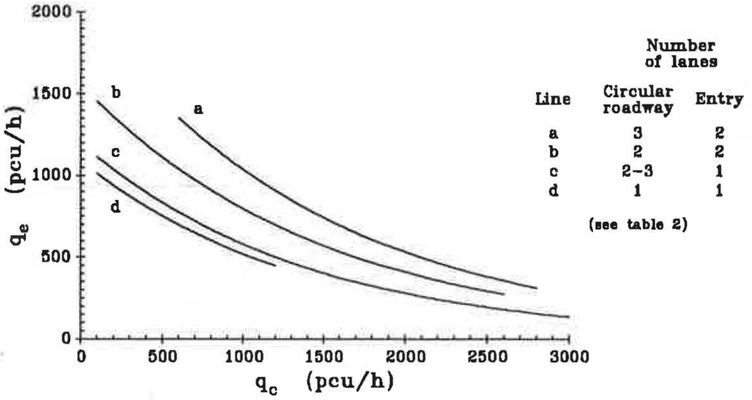


FIGURE 5 Regression curves for calculating the capacity of roundabouts.

Several types of roundabouts were investigated. The greatest samples could be observed at two-lane entrances into two-lane roundabouts. The total diameters of the circular roadways were between 40 and 142 m. Here the results seem to be a reliable representation of German conditions. At roundabouts with other combinations of entering and circulating lanes, the sample is not sufficient for final results. The reasons are that only a few of the one-lane roundabouts provided high traffic volumes and that only one roundabout with more than two circular lanes could be found.

Comprehensive results of the analysis are given in Table 2. The equations are illustrated by Figure 5. This figure indicates the range of q_c values for which empirical data could be obtained. These preliminary formulas are used in practice for capacity calculations of roundabouts in West Germany. They are implemented into the computer program KREISEL (*22*), which is also able to calculate capacities according to the English (*20*) and French (*21*) formulas and according to gap acceptance techniques.

On the whole, capacities of roundabouts in Germany seem to be considerably lower than the values predicted by the English formulas. The range of the German results is between 0.7 and 0.8 of the English values. However, there is good agreement between French and German results.

Reasons for differences may be explained by different driver behavior in the two countries; e.g., German drivers orientate by lanes, and the left lane of multilane approaches is hardly used. This is also a reason why variations in road widths—like in the English and French formula—do not seem to be good indicators for increased capacities under German conditions. Intersections with extremely high traffic volumes provided increased capacities compared with intersections with average traffic volumes. Because of different driver behavior, capacity formulas for roundabouts should not be transferred between different countries. Instead, each country has to find a solution of its own.

Of course, research on roundabout capacity should be continued in Germany as well. First, a greater data sample is needed for one-lane roundabouts. Furthermore, the influence of traffic volumes leaving the roundabout at the upstream and downstream exits (like in the French formula) should be determined. Finally, the impact of geometric features like the total diameter or the size of the center island on roundabout capacity is of considerable interest.

TABLE 2 FORMULA FOR CALCULATING ROUNDABOUT CAPACITY q_e IN RELATION TO CIRCULATING TRAFFIC VOLUME q_c

Capacity-formula		$q_e = A \cdot e^{-B \cdot \frac{q_c}{10,000}}$			
Number of lanes		Parameters		Sample size	Line
Circular roadway	Entry	A	B	(# 1-min.-intervals)	(in fig. 5)
1	1	1,089	7.42	275	d
2-3	1	1,200	7.30	455	c
2	2	1,553	6.69	3,873	b
3	2	2,018	6.68	295	a

REFERENCES

1. Berechnung der Leistungsfähigkeit von Knotenpunkten ohne Lichtsignalanlagen. *Anhang zum Merkblatt für Lichtsignalanlagen an Landstrassen*, Forschungsgesellschaft für das Strassenwesen, Köln, West Germany, 1972.
2. *Special Report 209: Highway Capacity Manual*, TRB, National Research Council, Washington, D.C., 1985.
3. W. Brilon. Recent Developments in Calculation Methods for Unsignalized Intersections in West Germany. In *Intersections Without Traffic Signals*, W. Brilon (ed.), Springer-Verlag, New York, 1988, pp. 111–153.
4. W. Brilon (ed.). *Intersections Without Traffic Signals*. Springer-Verlag, New York, 1988.
5. J. Harders. Die Leistungsfähigkeit nicht Signalgeregelter Städtischer Verkehrsknoten. *Schriftenreihe Strassenbau und Strassenverkehrstechnik*, Heft 76, Bonn, West Germany, 1968.
6. W. Siegloch. Die Leistungsermittlung an Knotenpunkten ohne Lichtsignalsteuerung. *Schriftenreihe Strassenbau und Strassenverkehrstechnik*, Heft 154, Bonn, West Germany, 1973.
7. X. Zhang. Einfluss des Grades der Teilgebundenheit auf die Wartezeit an Knotenpunkten ohne Lichtsignalanlagen. *Forschungsbericht*, Technische Hochschule, Darmstadt, 1986.
8. J. Chodur, S. Gaca, and M. Tracz. Störungen des Verkehrsablaufes an Höhengleichen Knotenpunkten ohne LSA-Steuerung bei Instationärem Zufluss. *Die Strasse*, Heft 7, 1989.
9. R. M. Kimber. Gap-Acceptance and Empiricism in Capacity Prediction. *Transportation Science*, Vol. 23, No. 2, May 1989.
10. J. Harders. Grenz- und Folgezeitlücken als Grundlage für die Leistungsfähigkeit von Landstrassen. *Schriftenreihe Strassenbau und Strassenverkehrstechnik*, Heft 216, Bonn, 1976.

11. R. Ashworth. The Capacity of Priority-Type Intersections with a Non-Uniform Distribution of Critical Acceptance Gaps. *Transportation Research*, Vol. 3, 1969.
12. W. Brilon and M. Grossmann. Aktualisiertes Berechnungsverfahren für Knotenpunkte ohne Lichtsignalanlagen. *Forschungsbericht*, Ruhr Universität, Bochum, Germany, 1990.
13. H. Kremser. Ein Einfaches Wartezeitproblem bei Einem Poisson'schen Verkehrsfluss. *Österreichisches Ingenieur-Archiv*, Band 16, Heft 1, 1961.
14. H. Kremser. Ein Zusammengesetztes Wartezeitproblem bei Poisson'schen Verkehrsströmen. *Österreichisches Ingenieur-Archiv*, Band 17, Heft 3, 1962.
15. F. J. Poeschl. Die Nicht Signalgesteuerte Nebenstrassenzufahrt als Verallgemeinertes M/G/1-Warteschlangensystem. *Zeitschrift für Operations Research*, Vol. 27B, 1983.
16. J. M. Stamm. Der Instationäre Verkehrsablauf an Knotenpunkten ohne Lichtsignalanlagen. Dissertation, Technische Hochschule, Darmstadt, 1974.
17. F. Tonke. Wartezeiten bei Instationärem Verkehr an Knotenpunkten ohne Lichtsignalanlagen. *Schriftenreihe Strassenbau und Strassenverkehrstechnik*, Heft 401, Bonn, West Germany, 1983.
18. M. Grossmann. KNOSIMO—A Practicable Simulation Model for Unsignalized Intersections. In *Intersections Without Traffic Signals*, W. Brilon (ed.). Springer-Verlag, Berlin, 1988, pp. 263–273.
19. W. Brilon and B. Stuwe. Einsatzmöglichkeiten von Kreisverkehrsplätzen und Aufgeweiteten Knotenpunkten unter Besonderer Berücksichtigung Ausländischer Erfahrungen. *Forschungsbericht*, Ruhr Universität, Bochum, West Germany, 1990.
20. R. M. Kimber. *The Traffic Capacity of Roundabouts*. Report LR 942. Transportation and Road Research Laboratory, Crowthorne, England, 1980.
21. G. Louah. Recent French Studies on Capacity and Waiting Times at Rural Unsignalized Intersections. In *Intersections Without Traffic Signals*, W. Brilon (ed.). Springer-Verlag, Berlin, 1988, pp. 248–262.
22. KREISEL. Computer Program for Calculating the Capacity of Roundabouts. *Lehrstuhl für Verkehrswesen*, Ruhr Universität, Bochum, West Germany, undated.

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

Integrated System of Freeway Corridor Simulation Models

LANNON LEIMAN, MOURAD BOUAOUINA, AND ADOLF D. MAY

The newly developed FREQ10 integrated system of freeway corridor simulation models is described with emphasis on traffic simulation, freeway improvement strategies, measures of effectiveness, and traveler responses. The major accomplishments have been to combine the previously developed entry control model, an extensively modified on-freeway priority model, and a newly developed common menu-driven interactive interface into an integrated system of models to extend the types of traffic management strategies that can be evaluated for a freeway corridor. The FREQ10 system of models enables the user to analyze design improvements, implementation of an HOV facility, implementation of normal and priority entry control, or implementation of time-varying capacity reduction situations such as reconstruction activities or freeway incidents. The extensive modifications that have been made to the on-freeway priority model include refining the procedure for traffic simulation, adding many new features such as user-supplied emission and fuel rates, and developing a new method for modeling the arterial and spatial response to reflect current policies that HOV facilities be considered as lanes added to the freeway. This system of models has received extensive testing and has been applied to freeway corridors in several urban areas in California.

During the past 20 years, the Institute of Transportation Studies at the University of California at Berkeley has developed and applied a sequence of freeway simulation models for the evaluation of various design and operational improvements. These models have been used by researchers and professionals both in the United States and abroad. Interactions with model users and experience in educational programs have led to continuous improvements in these models. Numerous reports are available describing these earlier developments (*1–20*).

A significant advancement in this modeling effort has been made as a result of a recently completed 2-year research and educational program sponsored by the California Department of Transportation and the FHWA. The major accomplishments have been to combine the previously developed entry control model, the extensively modified on-freeway priority lane model, and a newly developed common menu-driven interactive interface into an integrated system of models to extend the types of traffic management strategies that can be evaluated for a freeway corridor.

The purpose is to describe the newly developed FREQ10 integrated system of freeway corridor simulation models with emphasis on traffic simulation, freeway improvement strategies, measures of effectiveness, and traveler responses. A comprehensive final report of the research program (*21*) and two previous research reports (*9,17*) are available for in-depth coverage of each element of this newly developed system of models. Although this paper will not cover programming aspects nor present actual applications, these issues were of special importance and are covered in the final report. The computerized system of programs includes a menu-driven interactive graphical interface with comprehensive input checking, carefully selected default values, and user-selected output options including traffic performance contour maps. The extensive modifications that have been made to the on-freeway priority lane model include refining the procedure for traffic simulation, adding many new features such as user-supplied emission and fuel rates, and developing a new method for modeling the parallel arterial routes and spatial response to reflect current policies that HOV facilities be considered as lanes added to the freeway. Through a 6-month educational program for Caltrans professionals, 11 teams of engineers used FREQ10 to model existing freeway corridors in metropolitan areas throughout California.

MODEL OVERVIEW

The FREQ10 system contains two models: FREQ10PE, an entry control model for analyzing ramp metering; and FREQ10PL, an on-freeway priority model for analyzing HOV facilities. On entering the FREQ10 interface, the user may select either of the two models. Once the basic data have been entered for either FREQ10PE or FREQ10PL, the interface creates a data set that it can later access for either model. However, the models themselves do not interact; thus, entry control and on-freeway HOV facilities must be modeled independently. The two programs are similar in structure and use. The FREQ10PL model will be emphasized, but where major differences occur between the two models, the unique features in FREQ10PE will be identified.

The structure of the FREQ10PL priority lane model is shown in Figure 1. In the following general discussion, step numbers refer to numbers in Figure 1.

Step 1 The data are read from the input file. It includes freeway and parallel arterial design features, freeway ramp counts, arterial flows, and proposed design of freeway priority lanes.

Step 2 The model simulates the peak-period traffic operations for the (Day − 1) conditions before implementation of the priority lane. The results of this simulation serve as the basis of comparison for later simulations.

Step 3 If the freeway corridor has a parallel arterial route, the model simulates it for the (Day − 1) conditions.

Steps 4–5 If the user has requested only simulation of the (Day − 1) conditions before implementation, the model

Institute of Transportation Studies, University of California, 109 McLaughlin Hall, Berkeley, Calif. 94720.

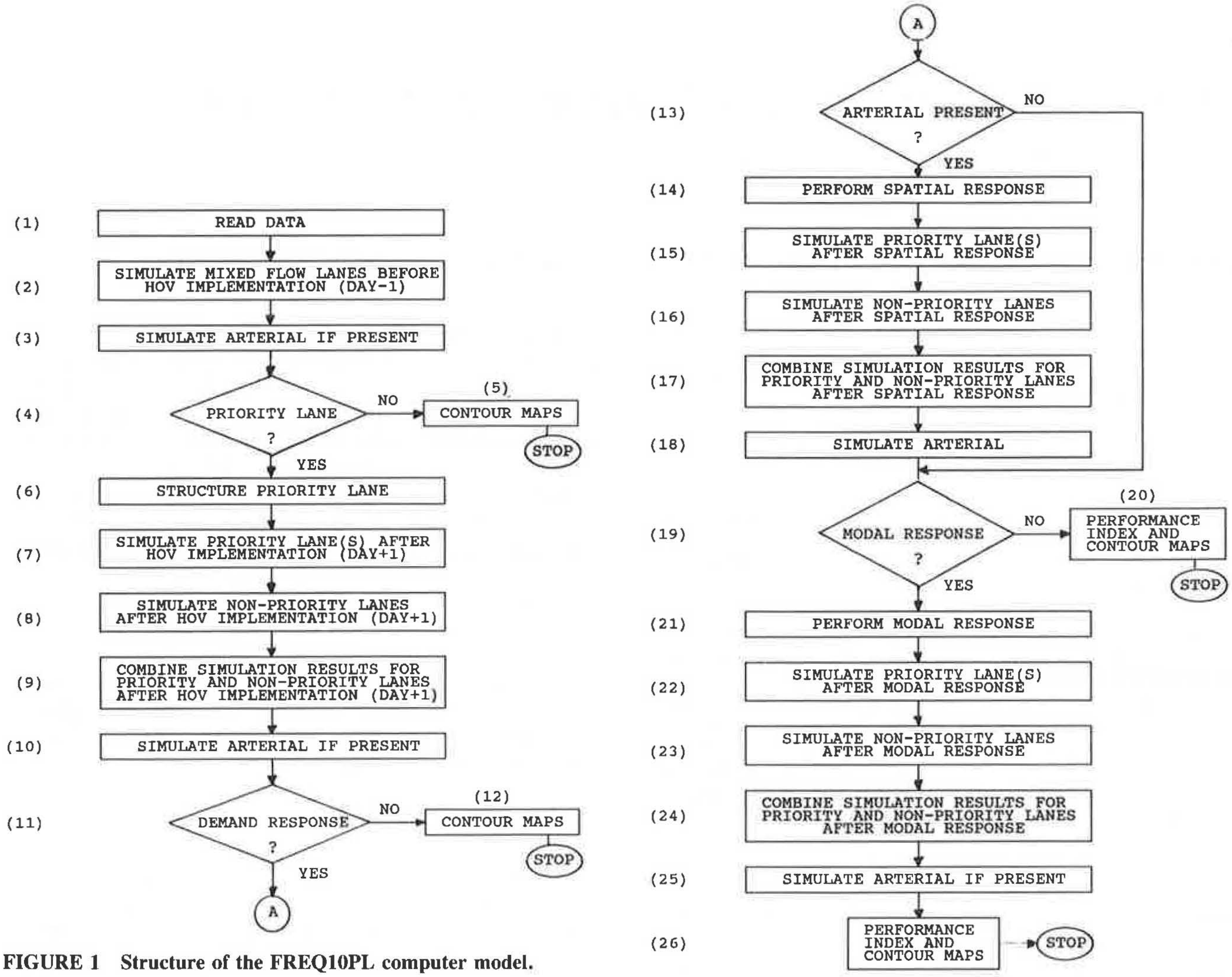

(1) READ DATA
(2) SIMULATE MIXED FLOW LANES BEFORE HOV IMPLEMENTATION (DAY-1)
(3) SIMULATE ARTERIAL IF PRESENT
(4) PRIORITY LANE ?
NO
(5) CONTOUR MAPS
STOP
YES
(6) STRUCTURE PRIORITY LANE
(7) SIMULATE PRIORITY LANE(S) AFTER HOV IMPLEMENTATION (DAY+1)
(8) SIMULATE NON-PRIORITY LANES AFTER HOV IMPLEMENTATION (DAY+1)
(9) COMBINE SIMULATION RESULTS FOR PRIORITY AND NON-PRIORITY LANES AFTER HOV IMPLEMENTATION (DAY+1)
(10) SIMULATE ARTERIAL IF PRESENT
(11) DEMAND RESPONSE ?
NO
(12) CONTOUR MAPS
STOP
YES
A
A
(13) ARTERIAL PRESENT ?
NO
YES
(14) PERFORM SPATIAL RESPONSE
(15) SIMULATE PRIORITY LANE(S) AFTER SPATIAL RESPONSE
(16) SIMULATE NON-PRIORITY LANES AFTER SPATIAL RESPONSE
(17) COMBINE SIMULATION RESULTS FOR PRIORITY AND NON-PRIORITY LANES AFTER SPATIAL RESPONSE
(18) SIMULATE ARTERIAL
(19) MODAL RESPONSE ?
NO
(20) PERFORMANCE INDEX AND CONTOUR MAPS
STOP
YES
(21) PERFORM MODAL RESPONSE
(22) SIMULATE PRIORITY LANE(S) AFTER MODAL RESPONSE
(23) SIMULATE NON-PRIORITY LANES AFTER MODAL RESPONSE
(24) COMBINE SIMULATION RESULTS FOR PRIORITY AND NON-PRIORITY LANES AFTER MODAL RESPONSE
(25) SIMULATE ARTERIAL IF PRESENT
(26) PERFORMANCE INDEX AND CONTOUR MAPS
STOP

FIGURE 1 Structure of the FREQ10PL computer model.

constructs any selected contour maps, provides the requested output, and stops.

Step 6 The model structures the HOV facility by adding the priority lanes to the freeway, changing the freeway capacities, splitting the origin-destination (O–D) tables into different occupancies, and making other manipulations necessary before the (Day + 1) short-term conditions after HOV implementation can be simulated. Note that all these changes are done by the program and are transparent to the user.

Steps 7–8 The model simulates the (Day + 1) conditions after implementation of the HOV facility in an effort to duplicate the first day of operations, before drivers have changed their driving behavior, i.e., all vehicles have the same time, space, and occupancy patterns as before. To do this, the model splits the freeway into two roadways, the priority lanes and the nonpriority lanes (including the upstream and downstream mixed-flow lanes); the model then simulates each roadway separately.

Step 9 The model combines the results of the simulations of the priority lanes and the nonpriority lanes for (Day + 1) to determine the performance of the entire freeway for (Day + 1). The results of this combined (Day + 1) simulation are used for comparison with the (Day − 1) simulation.

Step 10 If the freeway corridor has a parallel arterial route, the model simulates it for the (Day + 1) conditions. The results of simulation of the arterial for (Day − 1) and (Day + 1) are identical, because no spatial shift has occurred yet.

Steps 11–12 If the user has not requested any demand response, the model constructs any selected contour maps, provides the requested output, and stops.

Step 13 If there is no parallel arterial, the model will jump to Step 19. If there is an arterial, the model will proceed to Step 14.

Step 14 The model performs spatial response by shifting from the parallel arterial to the freeway those automobiles that save sufficient time. This includes priority as well as nonpriority automobiles.

Steps 15–16 The model simulates the corridor after spatial response in an effort to duplicate operations, after drivers have changed their driving behavior by changing routes. To do this, the model splits the freeway into two roadways, the priority lanes, and the nonpriority lanes (including the upstream and downstream mixed-flow lanes); the model then simulates each roadway separately, adding the vehicles that have shifted from the arterial to the appropriate roadway.

Step 17 The model combines the results of the simulations of the priority lanes and the nonpriority lanes after spatial response to determine the performance of the entire freeway after spatial response. The results of this combined simulation after spatial response are used for comparison with the (Day − 1) simulation.

Step 18 The model simulates the arterial for the conditions after spatial shift.

Steps 19–20 If modal response has not been requested, the model calculates the performance index, constructs any contour maps that have been selected, provides the requested output, and stops.

Step 21 The model performs modal response by shifting the passengers who would save sufficient time from nonpriority vehicles to priority vehicles. The actual number of passengers shifted is based on either program- or user-supplied elasticities.

Steps 22–23 The model simulates the corridor after modal response in an effort to duplicate operations, after drivers have changed their driving behavior by changing their mode of travel. To do this, the model again splits the freeway into two roadways, the priority lanes and the nonpriority lanes (including the upstream and downstream mixed-flow lanes); the model then simulates each roadway separately, adjusting the demand with respect to the passengers that have shifted from nonpriority vehicles to priority vehicles.

Step 24 The model combines the results of the simulations of the priority lanes and the nonpriority lanes after modal response to determine the performance of the entire freeway after modal response. The results of this combined simulation after modal response are used for comparison with the (Day − 1) simulation.

Step 25 If the freeway corridor has a parallel arterial route, the model simulates it for conditions after modal response. The results of simulation of the arterial for conditions after spatial response and after modal response are identical, because no additional spatial response has occurred.

Step 26 The model calculates the performance index, constructs any selected contour maps, provides requested output, and then stops.

The FREQ8PE research report (*9*) contains a description of the structure of the FREQ10PE model.

TRAFFIC SIMULATION

The same simulation module is the core of both the FREQ10PL model and the FREQ10PE model. The freeway corridor is simulated for existing conditions and after each change in those conditions. The simulation module analyzes the freeway corridor for each of a possible 24 time slices beginning with the first and continuing through the last. Within a time slice, each of a possible 38 subsections is processed sequentially from upstream to downstream. The simulation module consists of two parts: one for simulating conditions on the freeway, the other for simulating conditions on the parallel arterial.

The input for the simulation module initially consists of the freeway and arterial designs, the freeway demand in the form of synthetic O-D trip tables generated by the models from user-supplied ramp counts for each time slice, freeway and arterial occupancy distributions, and freeway and arterial subsection vehicle and passenger flows. Later in the run, additional input, consisting of those vehicles whose drivers have changed routes and those vehicles whose passengers have changed mode, is provided to the simulation module by the spatial response and modal response modules. Traffic demands, and subsection vehicle and passenger flows, can then be modified by the simulation module to reflect the conditions after traveler response in either the FREQ10PL or FREQ10PE model. Metered vehicles in the FREQ10PE model are handled similarly.

Simulation of conditions on the freeway can be divided into two parts. The first involves ramp queuing, ramp merging, and weaving analysis. The second includes mainline travel time and queuing analysis.

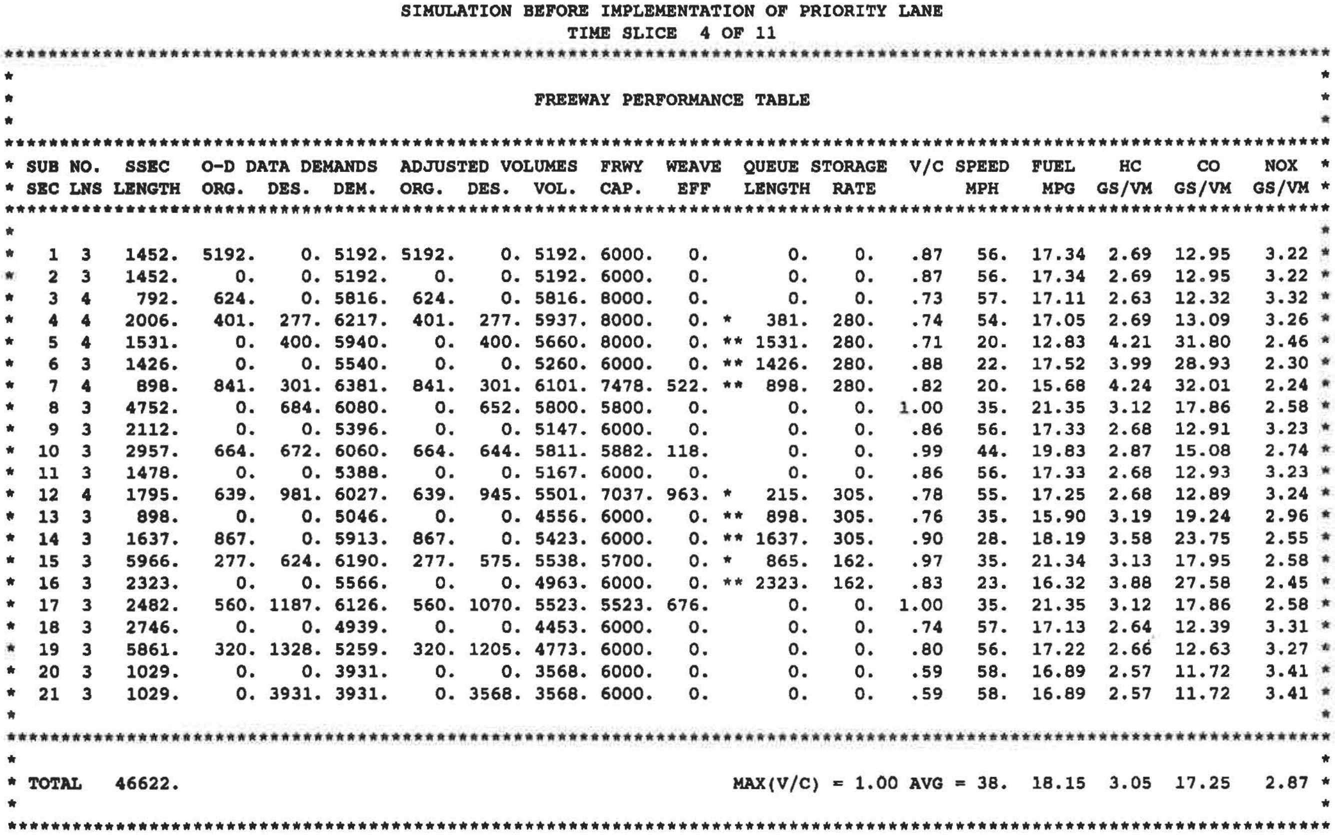
On- and off-ramp queues, which are modeled as vertical queues, form whenever ramp demand exceeds ramp capacity. Ramp merging analysis, at each on-ramp and at the merge point on the freeway, and weaving analysis, are performed in accordance with the 1965 Highway Capacity Manual (*22*). The model considers only simple and two-part compound weaving.

The average speed is calculated for each subsection as a function of the flow and the presence or absence of queuing by using speed versus volume-to-capacity ratio curves. A mainline queue develops whenever demand in any given subsection exceeds the subsection capacity. Shockwave analysis is used for determining queue evolution, queue collisions, queue splits, and queue discharge. In addition, the simulation module computes travel time, travel distance, fuel consumption, and vehicle emissions for each subsection during each time slice.

The simulation of conditions on the arterial are based on subsection flows. Travel times, travel distance, average speed, fuel consumption, and vehicle emissions are calculated for each subsection during each time slice.

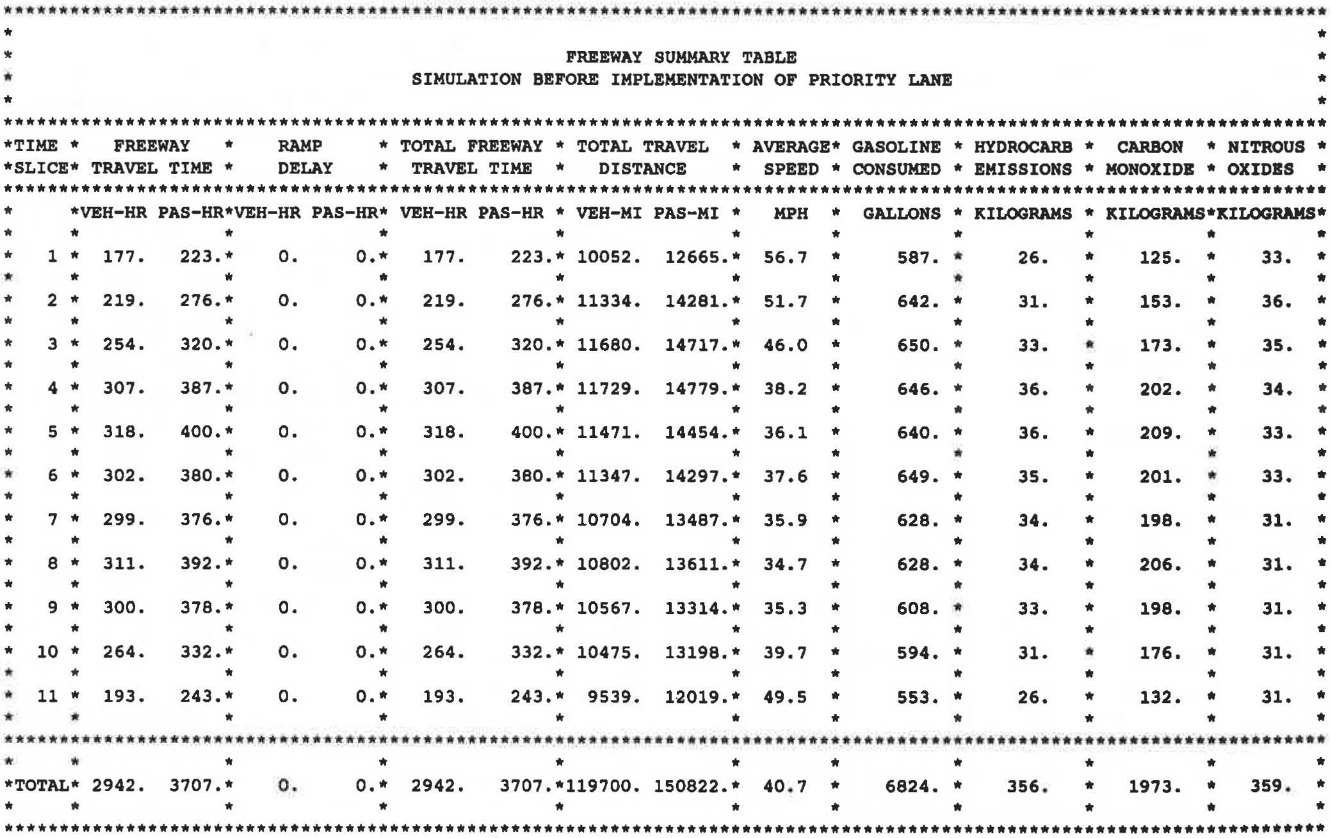
The simulation module produces results for each time slice as well as summaries over all time slices. Time slice output includes O–D tables as modified by the simulation module; O–D tables of freeway travel times, transferred, and diverted vehicles; freeway and arterial performance tables displaying the subsection characteristics and impacts; and detailed ramp queuing information. After all time slices have been processed, simulation summary tables are produced for the freeway and the arterial that display the impacts during each time slice and the totals over all time slices. For multiple simulation runs, a differential effects table is produced that compares the impact for the corridor totals from the current simulation with the corridor totals from the initial simulation. The final output can include distance-time contour maps of the first and last simulations for 10 different variables. The user selects which of these many outputs are to be provided for any particular run. Figure 2 shows an example of a freeway performance table for one time slice; Figure 3 shows an example of a freeway summary table.

Recent modifications have been made to the FREQ10PL simulation module to incorporate features of the more advanced FREQ10PE model. These include adding user-supplied speed versus volume-to-capacity ratio curves, user-supplied fuel con-



```
                         SIMULATION BEFORE IMPLEMENTATION OF PRIORITY LANE
                                       TIME SLICE  4 OF 11
*************************************************************************************************************
*                                                                                                           *
*                                       FREEWAY PERFORMANCE TABLE                                           *
*                                                                                                           *
*************************************************************************************************************
* SUB NO.  SSEC  O-D DATA DEMANDS  ADJUSTED VOLUMES  FRWY  WEAVE   QUEUE STORAGE  V/C SPEED  FUEL    HC     CO    NOX *
* SEC LNS LENGTH ORG.  DES.  DEM.  ORG.  DES.  VOL.  CAP.   EFF   LENGTH  RATE        MPH    MPG  GS/VM  GS/VM GS/VM *
*************************************************************************************************************
*                                                                                                           *
*   1  3  1452. 5192.    0. 5192. 5192.    0. 5192. 6000.   0.        0.    0.   .87  56.  17.34  2.69  12.95  3.22 *
*   2  3  1452.    0.    0. 5192.    0.    0. 5192. 6000.   0.        0.    0.   .87  56.  17.34  2.69  12.95  3.22 *
*   3  4   792.  624.    0. 5816.  624.    0. 5816. 8000.   0.        0.    0.   .73  57.  17.11  2.63  12.32  3.32 *
*   4  4  2006.  401.  277. 6217.  401.  277. 5937. 8000.   0. *    381.  280.   .74  54.  17.05  2.69  13.09  3.26 *
*   5  4  1531.    0.  400. 5940.    0.  400. 5660. 8000.   0. **  1531.  280.   .71  20.  12.83  4.21  31.80  2.46 *
*   6  3  1426.    0.    0. 5540.    0.    0. 5260. 6000.   0. **  1426.  280.   .88  22.  17.52  3.99  28.93  2.30 *
*   7  4   898.  841.  301. 6381.  841.  301. 6101. 7478. 522. **   898.  280.   .82  20.  15.68  4.24  32.01  2.24 *
*   8  3  4752.    0.  684. 6080.    0.  652. 5800. 5800.   0.        0.    0.  1.00  35.  21.35  3.12  17.86  2.58 *
*   9  3  2112.    0.    0. 5396.    0.    0. 5147. 6000.   0.        0.    0.   .86  56.  17.33  2.68  12.91  3.23 *
*  10  3  2957.  664.  672. 6060.  664.  644. 5811. 5882. 118.        0.    0.   .99  44.  19.83  2.87  15.08  2.74 *
*  11  3  1478.    0.    0. 5388.    0.    0. 5167. 6000.   0.        0.    0.   .86  56.  17.33  2.68  12.93  3.23 *
*  12  4  1795.  639.  981. 6027.  639.  945. 5501. 7037. 963. *    215.  305.   .78  55.  17.25  2.68  12.89  3.24 *
*  13  3   898.    0.    0. 5046.    0.    0. 4556. 6000.   0. **   898.  305.   .76  35.  15.90  3.19  19.24  2.96 *
*  14  3  1637.  867.    0. 5913.  867.    0. 5423. 6000.   0. **  1637.  305.   .90  28.  18.19  3.58  23.75  2.55 *
*  15  3  5966.  277.  624. 6190.  277.  575. 5538. 5700.   0. *    865.  162.   .97  35.  21.34  3.13  17.95  2.58 *
*  16  3  2323.    0.    0. 5566.    0.    0. 4963. 6000.   0. **  2323.  162.   .83  23.  16.32  3.88  27.58  2.45 *
*  17  3  2482.  560. 1187. 6126.  560. 1070. 5523. 5523. 676.        0.    0.  1.00  35.  21.35  3.12  17.86  2.58 *
*  18  3  2746.    0.    0. 4939.    0.    0. 4453. 6000.   0.        0.    0.   .74  57.  17.13  2.64  12.39  3.31 *
*  19  3  5861.  320. 1328. 5259.  320. 1205. 4773. 6000.   0.        0.    0.   .80  56.  17.22  2.66  12.63  3.27 *
*  20  3  1029.    0.    0. 3931.    0.    0. 3568. 6000.   0.        0.    0.   .59  58.  16.89  2.57  11.72  3.41 *
*  21  3  1029.    0. 3931. 3931.    0. 3568. 3568. 6000.   0.        0.    0.   .59  58.  16.89  2.57  11.72  3.41 *
*                                                                                                           *
*************************************************************************************************************
*                                                                                                           *
* TOTAL   46622.                                        MAX(V/C) = 1.00 AVG = 38.  18.15  3.05  17.25  2.87 *
*                                                                                                           *
*************************************************************************************************************
```

FIGURE 2 Sample performance table.



```
*****************************************************************************************************************
*                                                                                                               *
*                                         FREEWAY SUMMARY TABLE                                                 *
*                             SIMULATION BEFORE IMPLEMENTATION OF PRIORITY LANE                                 *
*                                                                                                               *
*****************************************************************************************************************
*TIME *   FREEWAY    *     RAMP      * TOTAL FREEWAY *  TOTAL TRAVEL  * AVERAGE* GASOLINE * HYDROCARB *  CARBON  * NITROUS *
*SLICE* TRAVEL TIME  *     DELAY     *  TRAVEL TIME  *    DISTANCE    *  SPEED * CONSUMED * EMISSIONS * MONOXIDE * OXIDES  *
*****************************************************************************************************************
*     *VEH-HR PAS-HR*VEH-HR PAS-HR* VEH-HR PAS-HR * VEH-MI  PAS-MI *   MPH  *  GALLONS * KILOGRAMS * KILOGRAMS*KILOGRAMS*
*     *              *               *               *                *        *          *           *          *         *
*   1 * 177.    223.*  0.       0.*  177.    223.* 10052.  12665.* 56.7   *    587. *     26.   *    125. *    33.  *
*     *              *               *               *                *        *          *           *          *         *
*   2 * 219.    276.*  0.       0.*  219.    276.* 11334.  14281.* 51.7   *    642. *     31.   *    153. *    36.  *
*     *              *               *               *                *        *          *           *          *         *
*   3 * 254.    320.*  0.       0.*  254.    320.* 11680.  14717.* 46.0   *    650. *     33.   *    173. *    35.  *
*     *              *               *               *                *        *          *           *          *         *
*   4 * 307.    387.*  0.       0.*  307.    387.* 11729.  14779.* 38.2   *    646. *     36.   *    202. *    34.  *
*     *              *               *               *                *        *          *           *          *         *
*   5 * 318.    400.*  0.       0.*  318.    400.* 11471.  14454.* 36.1   *    640. *     36.   *    209. *    33.  *
*     *              *               *               *                *        *          *           *          *         *
*   6 * 302.    380.*  0.       0.*  302.    380.* 11347.  14297.* 37.6   *    649. *     35.   *    201. *    33.  *
*     *              *               *               *                *        *          *           *          *         *
*   7 * 299.    376.*  0.       0.*  299.    376.* 10704.  13487.* 35.9   *    628. *     34.   *    198. *    31.  *
*     *              *               *               *                *        *          *           *          *         *
*   8 * 311.    392.*  0.       0.*  311.    392.* 10802.  13611.* 34.7   *    628. *     34.   *    206. *    31.  *
*     *              *               *               *                *        *          *           *          *         *
*   9 * 300.    378.*  0.       0.*  300.    378.* 10567.  13314.* 35.3   *    608. *     33.   *    198. *    31.  *
*     *              *               *               *                *        *          *           *          *         *
*  10 * 264.    332.*  0.       0.*  264.    332.* 10475.  13198.* 39.7   *    594. *     31.   *    176. *    31.  *
*     *              *               *               *                *        *          *           *          *         *
*  11 * 193.    243.*  0.       0.*  193.    243.*  9539.  12019.* 49.5   *    553. *     26.   *    132. *    31.  *
*     *              *               *               *                *        *          *           *          *         *
*****************************************************************************************************************
*     *              *               *               *                *        *          *           *          *         *
*TOTAL* 2942.  3707.*  0.       0.* 2942.   3707.*119700. 150822.* 40.7   *   6824. *    356.   *   1973. *   359.  *
*     *              *               *               *                *        *          *           *          *         *
*****************************************************************************************************************
```

FIGURE 3 Sample summary table.

sumption rates, user-supplied vehicle emission rates, and detailed user output specifications. In addition, FREQ10PL has been modified so that bus O–D tables for the freeway can be generated by the program based on user-supplied occupancy distributions. Several refinements have been made to the queuing analysis and the differential effects tables have been added. Finally, noise contour maps, a new fuel consumption module, a new vehicle emissions module, and a new flow-dependent simulation module for the arterial have been added to the FREQ10PL model. These modifications and additions to the FREQ10PL model ensure that both the FREQ10PL and the FREQ10PE models now produce identical results for simulation of existing conditions.

A more complete description of the simulation module can be found in the final report (*21*).

FREEWAY IMPROVEMENT STRATEGIES

Improvement strategies that can be modeled by the FREQ10 system of models fall into four categories: design improvements, implementation of an HOV facility, implementation of normal and priority entry control, and implementation of time-varying capacity reduction situations.

Design improvements can be evaluated by using either the FREQ10PL or FREQ10PE model. If a simulation run for existing conditions reveals congestion on the freeway that appears to be caused by isolated bottlenecks, the user should first try to improve the design of the freeway at the congested areas before considering other options. Such improvements might include adding an extra lane in one or two subsections, adding a lane to an on- or off-ramp, or making other improvements to the freeway design that would increase capacity at the congested areas. These design changes can be easily made to the original data set, and the new conditions simulated by either model. The user may discover after trying several possible design improvements, that the freeway congestion can be eliminated by making design improvements alone. However, if the elimination of the original bottlenecks merely results in new problems downstream, a more comprehensive solution such as the implementation of a priority lane, or the implementation of normal or priority entry control, should be explored.

The FREQ10PL model enables the user to evaluate the implementation of exclusive priority lanes. The user specifies the number of lanes in the HOV facility, where it begins, and where it ends. The occupancy cutoff level for vehicles that can use the HOV facility can be two or more passengers plus buses, three or more passengers plus buses, or buses only.

A major conceptual change needed to be made in the FREQ10PL model to reflect the policy that an HOV facility will always represent lanes added to the freeway. Because existing conditions are always simulated first during any run, the model needed the added capability of automatically modeling the additional HOV lanes after simulating existing conditions.

It is assumed that the HOV facility is continuous, that it is on the left side of the roadway, and that there is no barrier between it and the other freeway lanes. In addition, it is assumed that all HOV vehicles use the HOV facility, enter it immediately downstream of the freeway on-ramps, and exit it immediately upstream of the freeway off-ramps.

The effects of a particular HOV operational design can be evaluated for the day after implementation of the HOV lanes, after spatial response, which in the new FREQ10PL occurs from the parallel arterials to the freeway because of improved conditions on the freeway, and after modal response. The contour maps help the user visualize the effects of the improvement strategy. Figure 4 shows queue length contour maps for the existing conditions and for conditions after modal response for both the HOV facility and the non-HOV lanes.

The FREQ10PE model enables the user to consider improving freeway conditions by modeling normal or priority entry control at the on-ramps. The model can analyze a user supplied metering plan or provide an optimum computer-generated metering plan. The effects of either metering plan can be simulated after implementation, after spatial response (which in FREQ10PE is from the freeway to the arterial), and after modal response.

Although evaluating normal and priority entry control is the main function of the FREQ10PE model, it can also be used to simulate the operational effects of freeway maintenance and reconstruction activities by allowing the user to vary the subsection capacities over time (*17*). In this way, a schedule for maintenance or reconstruction can be developed that will cause the least disruption to traffic on the freeway. Another application of the time-varying capacity reduction capability of the FREQ10PE model would be to simulate incidents on the freeway.

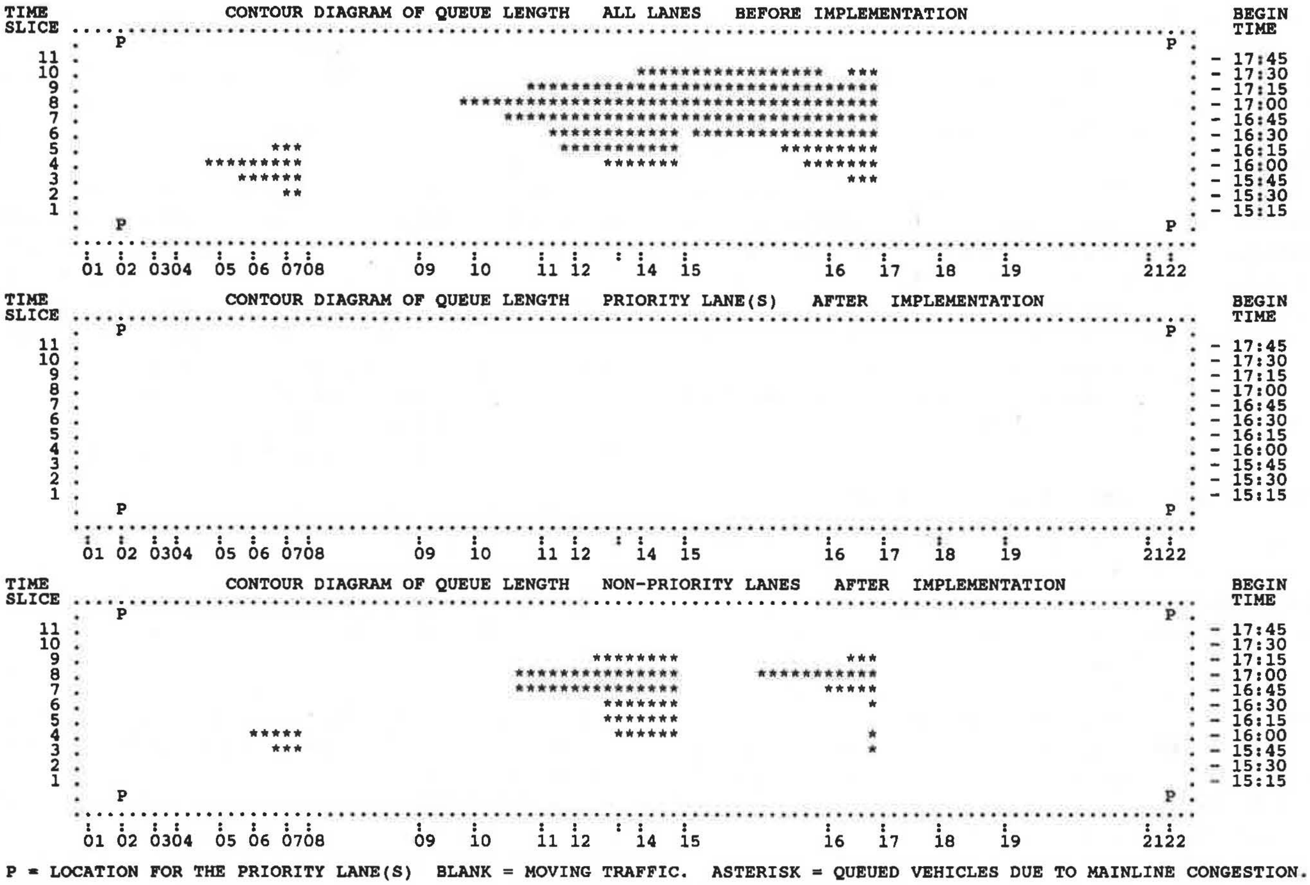
The FREQ10 interactive interface enables the user to model any of the many different improvement strategies with minimum effort. The data set is structured in such a way that the interface can access it for either a FREQ10PL or FREQ10PE simulation. Having entered the interface, the user selects the desired model, makes the necessary changes to the original data set, selects the desired output, and executes a run that simulates the improvement strategies under consideration.

MEASURES OF EFFECTIVENESS

In determining the effectiveness of any freeway improvement, it is important to evaluate traffic impacts on the freeway and parallel arterial routes. The FREQ10PL and FREQ10PE models provide information to help in this evaluation by calculating travel times, fuel consumption, vehicle emissions, traffic noise, and a cost-effectiveness performance index.

Travel Times

The travel times are computed for each subsection on the freeway and the arterial for each time slice. The freeway travel time for a given subsection and time slice is a function of the length of the subsection; the length of the time slice; and the speed, flow, and density of the subsection during the time slice. If the density of the subsection is not uniform throughout the time slice, travel time becomes the sum of the travel time during the congested period and the travel time during the noncongested period of the time slice. Ramp delays are added to the freeway travel times. The arterial travel time for a given subsection is a function of the flow, capacity, and signal properties of the arterial subsection.



```
TIME                CONTOUR DIAGRAM OF QUEUE LENGTH   ALL LANES   BEFORE IMPLEMENTATION                        BEGIN
SLICE ......................................................................................................   TIME
      .   P                                                                                              P .
   11 .                                                                                                    . - 17:45
   10 .                                                  *****************  ***                            . - 17:30
    9 .                                        ********************************                            . - 17:15
    8 .                                  **************************************                            . - 17:00
    7 .                                      **********************************                            . - 16:45
    6 .                                          ************ *****************                            . - 16:30
    5 .                 ***                       ***********         *********                            . - 16:15
    4 .           *********                           *******           *******                            . - 16:00
    3 .              ******                                                 ***                            . - 15:45
    2 .                  **                                                                                . - 15:30
    1 .                                                                                                    . - 15:15
      .   P                                                                                              P .
      ......................................................................................................
       :  :  : :   :  :  : :         :    :     :  :   : :   :            :    :    :     :            : :
       01 02 0304  05 06 0708        09   10    11 12    14  15           16   17   18    19           2122
TIME                CONTOUR DIAGRAM OF QUEUE LENGTH   PRIORITY LANE(S)   AFTER  IMPLEMENTATION                 BEGIN
SLICE ......................................................................................................   TIME
      .   P                                                                                              P .
   11 .                                                                                                    . - 17:45
   10 .                                                                                                    . - 17:30
    9 .                                                                                                    . - 17:15
    8 .                                                                                                    . - 17:00
    7 .                                                                                                    . - 16:45
    6 .                                                                                                    . - 16:30
    5 .                                                                                                    . - 16:15
    4 .                                                                                                    . - 16:00
    3 .                                                                                                    . - 15:45
    2 .                                                                                                    . - 15:30
    1 .                                                                                                    . - 15:15
      .   P                                                                                              P .
      ......................................................................................................
       :  :  : :   :  :  : :         :    :     :  :   : :   :            :    :    :     :            : :
       01 02 0304  05 06 0708        09   10    11 12    14  15           16   17   18    19           2122
TIME                CONTOUR DIAGRAM OF QUEUE LENGTH   NON-PRIORITY LANES   AFTER  IMPLEMENTATION               BEGIN
SLICE ......................................................................................................   TIME
      .   P                                                                                              P .
   11 .                                                                                                    . - 17:45
   10 .                                                                                                    . - 17:30
    9 .                                              ********               ***                            . - 17:15
    8 .                                       ***************       ***********                            . - 17:00
    7 .                                       ***************             *****                            . - 16:45
    6 .                                               *******                 *                            . - 16:30
    5 .                                               *******                                              . - 16:15
    4 .               *****                            ******                 *                            . - 16:00
    3 .                 ***                                                   *                            . - 15:45
    2 .                                                                                                    . - 15:30
    1 .                                                                                                    . - 15:15
      .   P                                                                                              P .
      ......................................................................................................
       :  :  : :   :  :  : :         :    :     :  :   : :   :            :    :    :     :            : :
       01 02 0304  05 06 0708        09   10    11 12    14  15           16   17   18    19           2122
P = LOCATION FOR THE PRIORITY LANE(S)  BLANK = MOVING TRAFFIC.  ASTERISK = QUEUED VEHICLES DUE TO MAINLINE CONGESTION.
```

FIGURE 4 Sample queue length contour maps.

Fuel Consumption

The fuel is calculated for each subsection during the mainline and arterial simulations for each time slice. Fuel consumption is a function of the vehicle-miles traveled on the subsection during the time slice, the class of vehicle, the fuel consumption rate of the class of vehicle traveling at the average speed of the subsection, the grade correction factor of the subsection, and the fraction of vehicles of each class on the subsection.

The vehicle fractions are supplied by the user. These fractions are constant over the entire simulation period. For the arterial, the vehicle mix is assumed to be the same throughout the entire length of the corridor section, while the vehicle mix on the freeway will usually vary from subsection to subsection because the vehicle distribution can be entered for each origin. The three vehicle classes that are simulated are automobiles, which include all of the four-wheeled vehicles; gasoline-powered trucks; and diesel-powered trucks. The grade correction factor is a function of the average speed, which is calculated by the program, and the mean subsection grade, which is provided by the user. Fuel consumption rate tables stored within the models give the rates at various average speeds for the three vehicle classes and the two facility types, freeway and arterial. A complete description of these and other tables mentioned in this section can be found in the final report (*21*). The user may supply substitute fuel consumption rate tables, if conditions are different from those assumed in the tables that are stored in the models.

Vehicle Emissions

The FREQ10PL and FREQ10PE models calculate the amounts of hydrocarbons (HC), carbon monoxide (CO), and oxides of nitrogen (NO_x) emitted by all vehicles in the modeled freeway corridor. The total emissions produced during the simulation are made up of two components: (a) emissions from vehicles traveling on the freeway and the arterial, and (b) emissions from vehicles delayed at on- and off-ramps. These calculations are performed during each simulation for each subsection and for each time slice.

The emissions generated from vehicles traveling on the freeway and on the arterial during a given time slice for a given subsection are a function of the type of pollutant, the class of vehicle, the vehicle emissions of the pollutant, the emission rate of the pollutant of the class of vehicle traveling at the average speed of the subsection, and the fraction of vehicles in each class.

The vehicle fractions, which are supplied by the user, and the vehicle classes are the same as those described for fuel consumption. Two emission rates tables are stored in the models, one for California vehicles and one for low-altitude non-California vehicles. These emission rate tables were obtained from the Environmental Protection Agency's MOBILE1 computer program (*23*) and give the emission rates at various average speeds for the three pollutants and the three classes of vehicles. The same tables are used for the freeway and the arterial. In addition, the models can read user-supplied emission rates that can then be used in place of the program-supplied rates.

The emissions generated by vehicles delayed at all freeway ramps during a given time slice are a function of the type of pollutant, the class of vehicle, the vehicle emission from delayed vehicles for the pollutant, the total vehicle delay at the ramps, the fraction of vehicles in each class, and the idling emission rate. The idling emission rates stored in the models were obtained from the MOBILE1 program (*23*). The user may supply different idling emission rates, which the model can then use instead of the program-supplied rates.

Traffic Noise

The level of noise generated by a freeway subsection during a time slice is computed for two different measurements. The L_{10} index expresses the noise level (in dBA) exceeded 10 percent of the time. This measure has been popular in the United States for many years. The equivalent noise level (L_{eq}) is an estimate of total noise. This measurement has been widely used in Europe. The output for both of these noise measurements consists of distance-time contour maps.

Cost-Effectiveness Performance Index

A new cost-effectiveness performance index has been added to the FREQ10PL model to aid the user in comparing the impacts of different HOV operational designs on a freeway corridor. The index is based on program- or user-supplied cost-benefit coefficients and the differential effects between the initial (Day − 1) conditions and the conditions after the last demand response. The differential effects that are considered are savings in travel time, savings in travel distance, savings in fuel consumption, and savings in vehicle emissions. The performance index is expressed as the savings in cost per peak period per mile of HOV lane added. It is recommended that the cost-effectiveness performance index only be used to compare different HOV operational designs for the same freeway, using identical cost-benefit coefficients. There is no cost-effectiveness performance index in the FREQ10PE model.

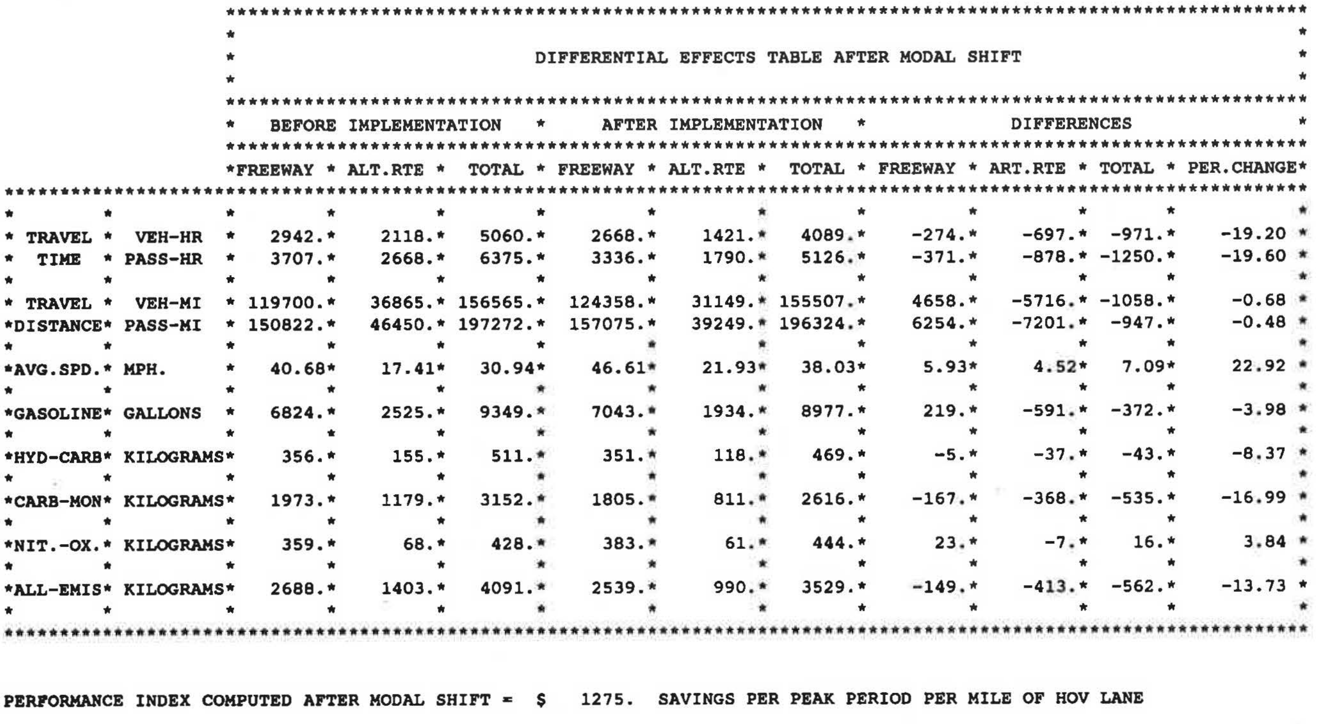
Figure 5 shows an example of a differential effects table that compares simulation of conditions after modal response to the initial conditions of the corridor. It contains most of the measures of effectiveness discussed in this section. Figure 5 also displays the cost-effectiveness performance index for the same run.

MODELING THE PARALLEL ARTERIAL ROUTES

The parallel arterial routes, if present, are aggregated and modeled as one. The freeway corridor can be modeled as having one of the following three arterial configurations:

1. No parallel arterial,
2. A parallel arterial that is continuous along all of the freeway, and
3. A parallel arterial that is discontinuous along the freeway, i.e., present at some locations and not at others.

The parallel arterial is incorporated into the analysis of the freeway corridor by a series of simplifying assumptions. The



```
                        **************************************************************************************************
                        *                                                                                                *
                        *                          DIFFERENTIAL EFFECTS TABLE AFTER MODAL SHIFT                          *
                        *                                                                                                *
                        **************************************************************************************************
                        *  BEFORE IMPLEMENTATION    *    AFTER IMPLEMENTATION    *              DIFFERENCES              *
                        **************************************************************************************************
                        *FREEWAY * ALT.RTE *  TOTAL * FREEWAY * ALT.RTE *  TOTAL * FREEWAY * ART.RTE * TOTAL * PER.CHANGE*
**************************************************************************************************************************
*        *              *        *         *        *         *         *        *         *         *       *           *
* TRAVEL *  VEH-HR      *   2942.*    2118.*   5060.*    2668.*    1421.*   4089.*    -274.*    -697.*  -971.*    -19.20 *
*  TIME  * PASS-HR      *   3707.*    2668.*   6375.*    3336.*    1790.*   5126.*    -371.*    -878.* -1250.*    -19.60 *
*        *              *        *         *        *         *         *        *         *         *       *           *
* TRAVEL *  VEH-MI      * 119700.*   36865.* 156565.*  124358.*   31149.* 155507.*    4658.*   -5716.* -1058.*     -0.68 *
*DISTANCE* PASS-MI      * 150822.*   46450.* 197272.*  157075.*   39249.* 196324.*    6254.*   -7201.*  -947.*     -0.48 *
*        *              *        *         *        *         *         *        *         *         *       *           *
*AVG.SPD.* MPH.         *   40.68*    17.41*   30.94*    46.61*    21.93*   38.03*     5.93*     4.52*   7.09*     22.92 *
*        *              *        *         *        *         *         *        *         *         *       *           *
*GASOLINE* GALLONS      *   6824.*    2525.*   9349.*    7043.*    1934.*   8977.*     219.*    -591.*  -372.*     -3.98 *
*        *              *        *         *        *         *         *        *         *         *       *           *
*HYD-CARB* KILOGRAMS    *    356.*     155.*    511.*     351.*     118.*    469.*      -5.*     -37.*   -43.*     -8.37 *
*        *              *        *         *        *         *         *        *         *         *       *           *
*CARB-MON* KILOGRAMS    *   1973.*    1179.*   3152.*    1805.*     811.*   2616.*    -167.*    -368.*  -535.*    -16.99 *
*        *              *        *         *        *         *         *        *         *         *       *           *
*NIT.-OX.* KILOGRAMS    *    359.*      68.*    428.*     383.*      61.*    444.*      23.*      -7.*    16.*      3.84 *
*        *              *        *         *        *         *         *        *         *         *       *           *
*ALL-EMIS* KILOGRAMS    *   2688.*    1403.*   4091.*    2539.*     990.*   3529.*    -149.*    -413.*  -562.*    -13.73 *
*        *              *        *         *        *         *         *        *         *         *       *           *
**************************************************************************************************************************


PERFORMANCE INDEX COMPUTED AFTER MODAL SHIFT =  $    1275.  SAVINGS PER PEAK PERIOD PER MILE OF HOV LANE

PERFORMANCE INDEX SHOULD BE USED TO COMPARE DIFFERENT HOV OPERATIONAL DESIGNS IN RUNS USING IDENTICAL COST COEFFICIENTS
```

FIGURE 5 Sample differential effects table and cost-effectiveness index.

arterial is divided into subsections that correspond to those of thc frccway. Thc boundarics of thcsc artcrial subsections are thus defined by the boundaries of the freeway subsections. Figure 6 shows an example of a modeled freeway corridor.

FREQ10PL is the first model in the FREQ family of programs that diverts vehicles from the arterial to the freeway. In order for the model to keep track of traffic patterns and the occupancy of vehicles shifting from the arterial to the freeway, FREQ10PL, unlike previous FREQ programs, must model the arterial in such a way that arterial O-D tables can be produced. Such tables are not needed in FREQ10PE, because in that model it is the freeway vehicles that change their routes to the arterial.

Because the arterial users of interest in FREQ10PL are those who could potentially change their routes to use the freeway if they save time, the origins and destinations on the parallel arterial are assumed to be located where there is a connection to the freeway (one-way or two-way traffic). This assumption is also shown in Figure 6.

The directional capacity of each arterial subsection is constant throughout the simulation run. The demand information for each arterial subsection consists of two variables: the average arterial flow and the percentage of vehicles leaving the arterial at the end of the subsection. The number of vehicles leaving the arterial at any destination is calculated by multiplying the percent turning off at the end of the subsection by the average subsection flow. The number of vehicles entering the arterial at any origin is calculated from the difference in flow volumes of the two consecutive arterial subsections. After using this method to calculate the demand at each arterial origin and destination, the model uses these demands to generate a synthetic arterial origin-destination trip table. This process is repeated for each of the time slices of the simulation period. The model then splits these arterial O–D matrices into priority and nonpriority vehicle matrices. The ability of the FREQ10PL model to generate synthetic O–D tables for the arterial from the origin and destination demands computed from arterial subsection flows and percent turning values eliminates the need for the user to collect detailed arterial O–D survey data.

Another important arterial modeling feature is the flow dependency of arterial travel times. Arterial travel times are calculated as a function of the characteristics of the subsections: flow, capacity, and control (arterial subsection with or without signalized intersections). For subsections with no signalized intersections, arterial vehicles are assumed to travel at a constant speed. For subsections with signalized intersections, an equation developed by Davidson (*24*) is used. This equation adjusts the speed according to the quality of traffic signal progression, the flow, and the capacity along the arterial subsection. This procedure has been added to the FREQ10PL model so that travel times on the arterial are now computed the same way in both the FREQ10PL and FREQ10PE models.

SPATIAL RESPONSE MODELING

In order to conform to current policies, FREQ10PL has been modified to model the on-freeway HOV facility as one or more add-on lanes. As such, the LOS on the freeway is always expected to improve, because additional capacity is provided. Therefore, spatial response, for the first time, is defined as a change in the travelers' route choice from the parallel directional arterials, to the freeway in response to the perceived changes in the freeway travel conditions. This change in direction of the spatial response required the development of a completely new spatial response module. The arterial vehicles that change their route to the freeway are assigned to the on-freeway HOV facility or general-purpose freeway lanes depending on the occupancy of the arterial vehicles. The new

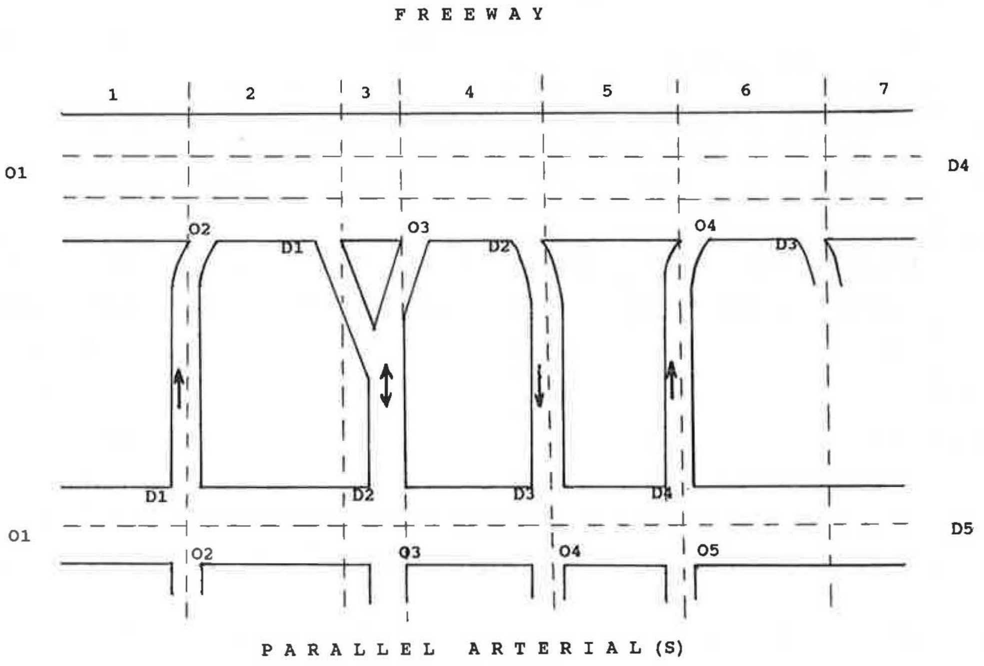


FIGURE 6 Sample freeway corridor.

model-generated arterial O–D tables, described in the previous section, enable the spatial response module to keep track of the traffic patterns and occupancy of vehicles that change their route from the arterial to the freeway after the addition of the HOV facility.

Spatial response is expected to occur within a few months of the implementation of the HOV lanes and before any significant modal response occurs. It is assumed that parallel arterial drivers reconsider their route choice (spatial response), at least to the extent of using or not using the freeway, before passengers in the nonpriority vehicles on the freeway consider changing modes (modal response). The reasons for making this assumption are as follows: mode choice generally represents a much more difficult decision than route choice, because drivers would need to adjust their trips to meet outside constraints, such as forming a carpool or adhering to a bus schedule. Route choice, on the other hand, typically has travel time or some other form of generalized cost as its only criterion. Therefore, the argument is that the modal split, existing before the freeway exclusive lanes implementation, reflects the difficult mode choice decision already made by commuters, and that changing mode would require more decision time than changing routes.

The two assumptions discussed earlier, that spatial shift occurs from the arterial to the freeway and spatial response occurs before modal response, are the most important ones in the development of the new spatial response algorithm. Additional assumptions and limitations are listed in the final report of the research program (*21*).

The basic rationale of the new FREQ10PL spatial response module is to set a travel time savings criterion for spatial response to occur, to calculate the freeway alternative route travel times for arterial travelers, and to divert to the freeway those arterial vehicles that meet the criterion of saving sufficient travel time. The minimum criterion for spatial diversion to occur is a travel time savings equal to a time penalty. This penalty is the sum of the average access time between the parallel arterial and the freeway, and a minimum perceived travel time savings.

The freeway corridor is assumed to be in equilibrium before the HOV facility is added; travelers use the arterial or the freeway depending on which better suits their needs. After the HOV facility is added and spatial response has occurred, the freeway corridor should once again be in equilibrium. It is extremely important to prevent the shifting of too many vehicles from the arterial to the freeway during spatial response, so that equilibrium can be achieved without the need to shift vehicles back to the arterial. In order to prevent over shifting from the arterial to the freeway and to better simulate the actual diversion process, the shifts are made in increments following a newly developed iterative scheme. Those arterial travelers saving the most time shift first, the arterial priority vehicles using the HOV facility wherever possible and the nonpriority vehicles using the general purpose lanes. The corridor is then resimulated to calculate the new travel costs before the next incremental group of vehicles is shifted from the arterial to the freeway. The user controls the number of incremental shifts that are performed by specifying the number of iterations. The model computes the criterion of travel time savings required for spatial response to occur during each iteration beginning with the maximum time that could be saved given the conditions in the corridor and decreasing the criterion after each iteration.

The structure of the new FREQ10PL spatial response module is shown in Figure 7. The following discussion refers to Figure 7.

Step 1. The maximum criterion for spatial diversion is calculated. This criterion defines the minimum travel time savings for spatial shift to occur during the first iteration.

Step 2. The potential travel time savings should arterial autos use the freeway to make their trips is calculated. Arterial vehicles are assumed to travel on the arterial until the first possible freeway on-ramp and come back to the arterial using the last off-ramp before their destination. Once on the freeway, arterial HOV vehicles are assumed to use the on-freeway HOV facility wherever possible and arterial non-HOV vehicles are assumed to use the general purpose lanes.

Step 3. Arterial priority automobiles meeting the criterion are diverted. For each arterial priority O–D pair for which vehicles have a travel time savings greater or equal to the criterion set for the current iteration, all of the vehicles are shifted to the appropriate corresponding freeway subsections.

Step 4. Arterial nonpriority automobiles meeting the criterion are diverted as in the previous step, except that vehicles are assigned to the general-purpose lanes of the freeway.

Step 5. After each arterial O–D pair for the current time slice is examined for possible diversion to the freeway, the program either proceeds with the next time slice and repeats Steps 2 to 4 above, or if the current time slice is the last one, it proceeds to the next step.

Step 6. After all of the time slices are analyzed for this iteration and potential shifters meeting the spatial shift criterion are diverted, the corridor is resimulated using the revised freeway and arterial demands that reflect the vehicles that have changed their routes during this iteration. The new travel times that are computed during this resimulation of the corridor are used for the next iteration.

Step 7. If all iterations are not completed, the model reduces the spatial response criterion and repeats Steps 2 to 6. After the last iteration, there are no more vehicles that would save time by changing their routes to the freeway and, thus, the corridor is once again in equilibrium. Control is returned to the main program that simulates the entire corridor after spatial response and provides any requested output for this simulation.

For more details on the FREQ10PL spatial response module, please see the final report (*21*).

Spatial response in FREQ10PE occurs from the freeway to the arterial. To avoid ramp delays caused by metering, vehicles divert to the arterial to enter the freeway at on-ramps downstream of their originally designated entry locations. For more information on the FREQ10PE spatial response module, please see the FREQ8PE research report (*9*).

MODAL RESPONSE MODELING

After spatial response has occurred, the arterial is in equilibrium with respect to the freeway, but the freeway itself is no

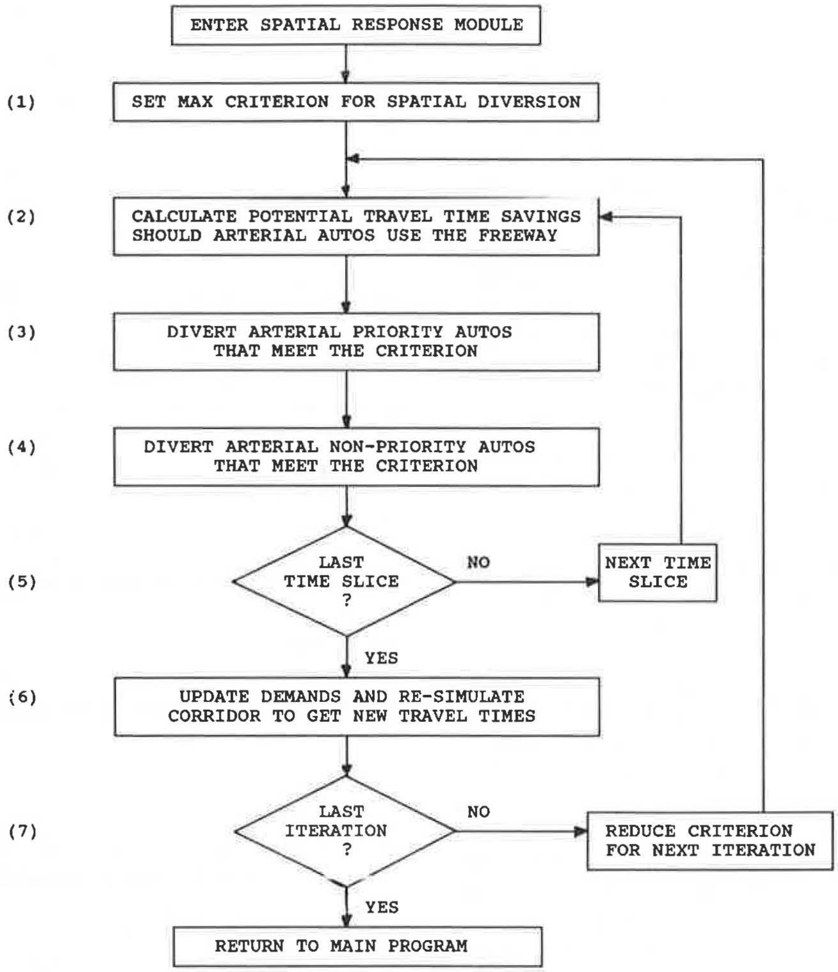

ENTER SPATIAL RESPONSE MODULE
(1) SET MAX CRITERION FOR SPATIAL DIVERSION
(2) CALCULATE POTENTIAL TRAVEL TIME SAVINGS SHOULD ARTERIAL AUTOS USE THE FREEWAY
(3) DIVERT ARTERIAL PRIORITY AUTOS THAT MEET THE CRITERION
(4) DIVERT ARTERIAL NON-PRIORITY AUTOS THAT MEET THE CRITERION
(5) LAST TIME SLICE ?
NO
NEXT TIME SLICE
YES
(6) UPDATE DEMANDS AND RE-SIMULATE CORRIDOR TO GET NEW TRAVEL TIMES
(7) LAST ITERATION ?
NO
REDUCE CRITERION FOR NEXT ITERATION
YES
RETURN TO MAIN PROGRAM

FIGURE 7 Overview of the spatial response module.

longer in equilibrium. The general-purpose lanes of the freeway are even more congested relative to the priority lanes than they were before spatial response because there are fewer HOV vehicles than non-HOV vehicles on the arterial that could potentially change their routes to the freeway. Thus, there could be an even stronger incentive for commuters to change their mode of travel to either buses or carpools to enjoy the increased savings in travel time offered by the HOV facility.

However, not all travelers who would save time change their mode of travel. The level of modal response depends on many variables. Some of these variables include socioeconomic characteristics of the commuters, perceived convenience, and attributes of the modes under investigation. The model predicts how many travelers change mode by applying elasticities that were derived from a multinomial logit model (*25*). These elasticities are expressed as a proportion of passengers shifting per minute of travel time saved. They are used within the program to predict how many freeway non-HOV passengers who would save time would actually change their mode of transportation to carpools or to buses. The values of elasticities depend on the priority cut-off level and the level of bus service. The user has the option to override these values by supplying the user's own elasticities. These user-supplied elasticities are also expressed as a proportion of passengers shifting per minute of on-freeway travel time saved.

The basic rationale of the modal response module is to set a criterion for sufficient time saved in order for modal response to occur, to calculate the difference in on-freeway travel times between priority and nonpriority vehicles, and to then shift to carpools and buses a percentage of those nonpriority passengers saving sufficient travel time as specified by the criterion. The percent shifted depends on the amount of time saved, the investigated priority strategy, and the LOS of the bus system. The priority strategy could be one of the following: buses only or buses and carpools (2^+ or 3^+).

In order to prevent overestimating the total number of passengers that change to carpools or buses and to better model the actual modal response, a new iterative scheme similar to that used in modeling spatial response was developed. Those nonpriority passengers saving the most time consider shifting first. The freeway is then resimulated before the next iteration to calculate the new travel costs. The number of iterations is provided by the user. The model computes the criterion of travel time savings required for modal response to occur for each iteration beginning with the maximum time that could be saved given the conditions on the freeway and decreasing the criterion appropriately for each successive iteration.

An overview of the structure of the modal response module along with its new iterative process is shown in Figure 8. The following discussion refers to Figure 8.

Step 1. The maximum criterion for modal shift is calculated. This criterion defines the minimum travel time savings required for modal response to occur during the first iteration.

Step 2. The potential travel time savings should passengers in nonpriority automobiles form carpools or ride buses is calculated.

Step 3. Passengers in nonpriority vehicles meeting the criterion are shifted to carpools and buses. For each nonpriority O–D pair for which vehicles have a travel time savings greater than or equal to the criterion set for the current iteration, a percentage of the passengers is shifted to the appropriate priority mode and thus to the priority lane subsections. The percent shift is determined by the elasticities provided by the program or by the user. Once an O–D pair for a given time slice has been partially shifted, it is never reconsidered during a subsequent iteration for the same time slice. This restriction prevents compounding the application of the elasticity values.

Step 4. After each nonpriority O–D pair for the current time slice is examined for possible modal response, the program either proceeds with the next time slice and repeats Steps 2 and 3, or if the current time slice is the last one, it proceeds to the next step.

Step 5. After all of the time slices are analyzed for this iteration and a portion of potential priority passengers meeting the time savings criterion are shifted, the corridor is resimulated using the revised freeway demands which reflect the reduction in nonpriority vehicles and the increase in priority vehicles because of modal response. The new travel times that are computed during this resimulation of the freeway are used for the next iteration.

Step 6. If all iterations are not completed, the model reduces the modal response criterion and repeats Steps 2 to 5. After the last iteration, no nonpriority O–D pairs exist that would save time with a modal response that has not already been considered. Control is then returned to the main program, which simulates the entire corridor after spatial response and provides any requested output for this simulation.

For further details on the modal response module, please refer to the final report (*21*).

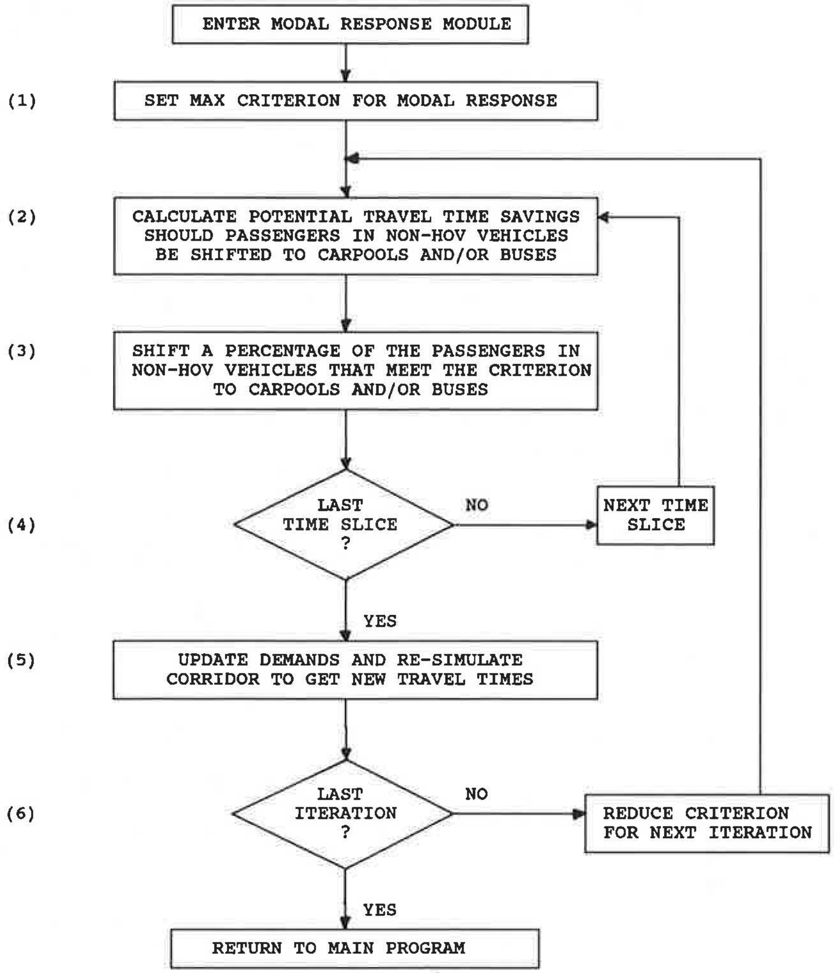

ENTER MODAL RESPONSE MODULE
(1)
SET MAX CRITERION FOR MODAL RESPONSE
(2)
CALCULATE POTENTIAL TRAVEL TIME SAVINGS SHOULD PASSENGERS IN NON-HOV VEHICLES BE SHIFTED TO CARPOOLS AND/OR BUSES
(3)
SHIFT A PERCENTAGE OF THE PASSENGERS IN NON-HOV VEHICLES THAT MEET THE CRITERION TO CARPOOLS AND/OR BUSES
(4)
LAST TIME SLICE ?
NO
NEXT TIME SLICE
YES
(5)
UPDATE DEMANDS AND RE-SIMULATE CORRIDOR TO GET NEW TRAVEL TIMES
(6)
LAST ITERATION ?
NO
REDUCE CRITERION FOR NEXT ITERATION
YES
RETURN TO MAIN PROGRAM

FIGURE 8 Overview of the modal response module.

SUMMARY

This paper has described the newly developed FREQ10 integrated system of freeway corridor simulation models with emphasis on traffic simulation, freeway improvement strategies, measures of effectiveness, and traveler responses. The major accomplishments have been to combine the previously developed entry control model, the extensively modified on-freeway priority lane model, and a newly developed common menu-driven interactive interface into an integrated system of models to extend the types of traffic management strategies that can be evaluated for a freeway corridor. This system of models has received extensive testing and, with the exception of the spatial response module, has been applied to freeway corridors in several urban areas in California. Application of the model to corridors with parallel arterials is currently in progress. Although outside the scope of this paper, a hypothetical case study of a freeway corridor analysis using various management strategies within the FREO10 system of models is presented in the final report (*21*). An application of the FREQ10PE model in a reconstruction study is also available (*17*).

Research in general, and research related to simulation models in particular, is never complete. There is always further research to be undertaken and related enhancements to be incorporated into the model. The next paragraphs identify such possible enhancements in regard to traffic simulation, freeway improvement strategies, measures of effectiveness, and traveler responses.

The traffic simulation portion of the model is currently limited to a corridor length of some 12 to 16 mi. Increasing the number of subsections and the number of on- and off-ramps would permit application to longer freeway corridors. Currently on-ramp, merge, and off-ramp queues are handled separately from freeway mainline queues. When these queues interact, the traffic performance results are approximate and there is a need to integrate them. In addition, a possible parallel freeway cannot be simulated.

Currently, the only improvement strategy that can be automatically optimized by the model is entry control. Incorporating optimization techniques within the model for on-freeway priority lanes and combined strategies such as design improvements with entry control would be desirable. The on-freeway priority lane improvement strategy in the model has several limitations that could be relaxed with further research. For example, no buffer is simulated to be present between the priority lanes and nonpriority lanes, and it is assumed that priority vehicles will enter the HOV lanes as soon as possible and stay in the HOV lanes as long as possible regardless of length of trip and with no regard to possible buffers or barriers.

The fuel, noise, and particularly the emission measures of effectiveness need to be updated to represent current and future vehicle fleet performance. The current cost-effectiveness calculations are in an early stage of development and a number of enhancements would be desirable.

The spatial and modal traveler response algorithms need to be tested against present real-life situations and calibrated when needed. The current model does not handle temporal or longer-term traveler responses.

After 20 years of almost continuous research and development, an integrated system of freeway corridor simulation models has emerged. Much has been accomplished and a comprehensive tool is available to researchers and professionals. Yet with the complexities of today's traffic situations, the comprehensiveness of available traffic management strategies, the wide variety of measures of effectiveness that require collective assessment, and the requirement for a time-stream system evaluation as travelers respond to system changes, further research is needed. Much has been accomplished—much has yet to be done.

ACKNOWLEDGMENTS

Many individuals assisted the project staff in this multiyear project. From the Berkeley campus, we wish to thank Dimitris Scapinakis for his significant contributions throughout this project. In addition, we appreciate the many contributions made by Joe Oliveira and Eric Freitag during the early phases of the project. We thankfully acknowledge that this research was sponsored by the California Department of Transportation and the FHWA. In particular, we wish to thank the following individuals who contributed to this research: Phil Jang and Mike Auslam from CALTRANS—we appreciate their support and cooperation throughout the life of this project—and Jeffrey Lindley and Frank Cechini from FHWA. In addition, we appreciate all the helpful feedback from members of our HOV professional panel.

REFERENCES

1. A. D. May. *Special Report 194: Models for Freeway Corridor Analysis*. TRB, National Research Council, Washington, D.C., 1981, pp. 23–32.
2. Parsons-Brinckerhoff, The TRANSPO Group, and A. D. May. *I-5 South HOV Project: Feasibility Report and Draft Environmental Assessment*. Washington State Department of Transportation, Nov. 1981.
3. V. Siu, J. Kahng, T. Imada, P. Jovanis, and A. D. May. *Freeway Priority Treatment Strategies Workshop Student Workbook*. Texas Department of Highways and Public Transportation, Austin, Dec. 1981.
4. A. D. May. *Introduction to Freeway Priority Lane Strategy Workshop Student Workbook*. Texas Department of Highways and Public Transportation, Austin, April 1982.
5. A. D. May and T. Imada. *Freeway Priority Treatment Strategies Program—Workshop I: FREQ7 Simulation/Calibration Student Workbook*. Texas Department of Highways and Public Transportation, Austin, April 1983.
6. A. D. May and T. Imada. *Freeway Priority Treatment Strategies Program—Workshop II: Ramp Metering Simulation Using FREQ7*. Texas Department of Highways and Public Transportation, Austin, July 1983.
7. A. D. May and T. Imada. *Freeway Priority Treatment Strategies Program—Workshop II: Priority Lane Simulation Using FREQ7*. Texas Department of Highways and Public Transportation, Austin, July 1983.
8. T. Imada and A. D. May. *FREQ8PL—Freeway Priority Lane Simulation Model*. Report UCB-ITS-TD-85-1. Institute of Transportation Studies, University of California, Berkeley, March 1985.
9. T. Imada and A. D. May. *FREQ8PE—A Freeway Corridor Simulation and Ramp Metering Optimization Model*. Report UCB-ITS-RR-85-10. Institute of Transportation Studies, University of California, Berkeley, June 1985.

10. A. D. May. Traffic Simulation on I-70/I-10: Implementations for System Control. *Proc., Arizona Symposium on Freeway Surveillance and Control*, Oct. 20–22, 1986, pp. 72–84.
11. A. D. May. Freeway Simulation Models—Revisited. In *Transportation Research Record 1132*, TRB, National Research Council, Washington, D.C., 1987, pp. 94–99.
12. A. D. May and V. Wong. *An Educational Program for Freeway Simulation and Optimization, Workshop I: Simulation and Calibration*. Texas Department of Highways and Public Transportation, Austin, Dec. 1987.
13. L. Newman, C. Nuworsoo, and A. D. May. Operational and Safety Experience with Freeway HOV Facilities in California. In *Transportation Research Record 1173*, TRB, National Research Council, Washington, D.C., 1988, pp. 18–24.
14. A. D. May and V. Wong. *An Educational Program for Freeway Simulation and Optimization, Workshop II: Optimization and Traveller Responses*. Texas Department of Highways and Public Transportation, Austin, March 1988.
15. C. K. Nurworsoo and A. D. May. *Planning HOV Lanes on Freeways: Final Report*. Report UCB-ITS-RR-88-11. Institute of Transportation Studies, University of California, Berkeley, Sept. 1988.
16. H. Al-Deek, M. Martello, A. D. May, and W. Sanders. *Potential Benefit of In-Vehicle Information Systems*. Report UCB-ITS-RR-88-2. Institute of Transportation Studies, University of California, Berkeley, Aug. 1988.
17. J. Zhang, L. Leiman, and A. D. May. Evaluation of Operational Effects of Freeway Reconstruction Activities. In *Transportation Research Record 1232*, TRB, National Research Council, Washington, D.C., 1989, pp. 27–39.
18. H. M. Al-Deek and A. D. May. *Potential Benefits of In-Vehicle Information Systems (IVIS): Demand and Incident Sensitivity Analysis*. Report UCB-ITS-PRR-89-1, Institute of Transportation Studies, University of California, Berkeley, May 1989.
19. A. D. May, L. Newman, and V. Wong. *An Educational Program for Freeway Operations Analysis, Workshop I: Basic Course*. California Department of Transportation, Sacramento, Jan. 1990.
20. A. D. May and V. Wong. *An Educational Program for Freeway Operations Analysis, Workshop II: Improvement Investigations*. California Department of Transportation, Sacramento, March 1990.
21. D. Scapinakis, L. Leiman, M. Bouaouina, and A. D. May. *Demand Estimation, Benefit Assessment, and Evaluation of On-freeway High Occupancy Vehicle Lanes*. Institute of Transportation Studies, University of California, Berkeley, Aug. 1990.
22. *Special Report 87: Highway Capacity Manual*, HRB, National Research Council, Washington, D.C., 1965.
23. *User's Guide to MOBILE 1: Mobile Source Emissions Model*. U.S. Environmental Protection Agency, Aug. 1978.
24. K. B. Davidson. *A Flow-Travel Time Relationship for Use in Transportation Planning, Proc.*, 3rd Conference of Australian Road Research Board, Vol. 3, Part 1.
25. D. McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. *Frontiers in Econometrics*, Academic Press, New York, 1973, pp. 105–142.

Publication of this paper sponsored by Committee on High-Occupancy Vehicle Systems.

Procedure for Validation of Microscopic Traffic Flow Simulation Models

RAHIM F. BENEKOHAL

Model verification and validation are two important tasks in developing a traffic simulation model. Traffic simulation models have unique characteristics because of the interaction among the drivers, vehicles, and roadway. The effects of the interaction on traffic flow should be considered in verification and validation of the models. If these two tasks are not properly performed, a traffic simulation model may not provide accurate results. A procedure for verification and validation of microscopic traffic simulation models is developed, and its application to a car-following simulation model, CARSIM, is demonstrated. The validation part of the procedure is emphasized. The validation efforts are performed at the microscopic and macroscopic levels. For validation at the microscopic level, the speed change patterns and trajectory plots obtained from simulation models are compared with those from a field data. For validation at the macroscopic level, the average speed, density, and volume for simulated platoons are compared with those of field data. Also, variation of these parameters when the platoons go through a disturbance and interrelationships between these variables computed from the simulation models and the field data are examined. Regression analysis and analysis of variance of the simulation results versus the field data are discussed. The procedure may be considered as a step toward development of a comprehensive systematic approach for verification and validation of traffic simulation models.

Two important tasks in developing a traffic simulation model are verification and validation of the model. Verification is to check if the model behaves as the experimenter assumes it does, and validation is to test whether the simulation model reasonably approximates a real system (*1*). Traffic simulation models have unique characteristics because of the interaction among the drivers, vehicles, and roadway environment. The effects of the interaction on traffic flow should be considered in verification and validation of the models. A traffic simulation model may not provide accurate results if these two tasks are not properly performed.

Often the users of traffic simulation models do not examine how well the verification and validation steps are carried out, or do not have access to such information. Consequently, the users rely on the model performance assuming that enough verification and validation have been done. Even when the information is available, it is difficult for most users to compare validation of one model to another model because they are validated differently. Thus, there is a need for developing a systematic approach for verification and validation of traffic simulation models to provide some degree of consistency and to increase the reliability of the models.

A procedure for verification and validation of traffic flow simulation models is suggested. The procedure is discussed and its application for verification and validation of a car-following simulation model, CARSIM (*2*), is demonstrated. Following this approach in validation of a model most likely will increase the reliability of the simulation results. However, using the suggested approach does not guarantee that this model will simulate the real-world conditions better than another model. The procedure may be considered as a step toward development of a comprehensive systematic approach for verification and validation of traffic simulation models.

BACKGROUND

Validation

A model should be validated under different experimental conditions to obtain a high model confidence. Validation is to see whether there is an adequate agreement between the model and the system being modeled. Annino and Russell (*3*) stated that using unverified models and lack of understanding the system were among most frequent causes of simulation failure. Sargent (*4*) suggested that model validation should consist of conceptual validation, computerized validation, operational validation, and use of adequate and correct data.

For the conceptual validation, the theories, the assumptions, and the relationships used are checked to ensure they are correct and proper for each submodel and for the overall model. Sargent (*4*) suggested using the tracing and face validation techniques. In tracing, the behavior of different entities (e.g., vehicles) is traced through each submodel and overall model to determine if the model's logic is correct and whether the necessary accuracy is obtained. In face validation, the experts in the subjects are asked to evaluate the logic of the submodel and the model and the input-output relationships.

In computerized model validation, it is checked to ensure that the conceptual model is implemented, and the computer program runs properly (error free). Each submodel is tested to see if it works properly and the overall model is executed under different conditions to investigate input and output relations. Operational validity is ensuring that the simulation model is a reasonable and accurate representation of the real system with certain levels of confidence. Here the validation can be done subjectively, such as graphical representation and examination of it, or can be done objectively such as using statistical techniques.

Various statistical techniques have been used for validation of simulation models. Torres et al. (*5*) developed a statistical

Department of Civil Engineering, University of Illinois, 205 N. Mathews Avenue, Urbana, Ill. 61801-2397.

guideline based on differences between the real world and traffic simulation results. Gafarian and Walsh (*6*) used travel time and velocity of vehicles as the measure of effectiveness of simulation model, and compared the simulation results with observed values using the Wilcoxon signed-rank test. Mihram (*7*) discussed the five stages in model building and the procedures for comparison of the results from several independent replications of a simulation model with that of one or more observed data.

Kleijnen (*8*) discussed techniques for validation of simulation models using chi squared, factor analysis, spectral analysis, and regression analysis between the actual and the simulation outputs. Naylor et al. (*9*) provided three alternative forms of analysis of variance for analysis of output from computer simulation experiments and for making a decision about their differences and ranking. Kleijnen (*10*) discussed how regression analysis is used to obtain a metamodel (a model explaining the simulation model) and how the effect of qualitative or quantitative factors can be investigated by using the weighted least squares technique. The four steps suggested by Sargent (*4*) were used to develop a procedure that will be discussed after review of the truncation and replication policies.

Truncation and Replication

The results of simulation studies should be collected after the system reaches a steady state condition. The data before the steady state conditions (initial transient state) may be eliminated to minimize the bias on the mean value of the response variable. There is not a definite rule on how the bias should be eliminated and how much of the simulation run should be truncated for this purpose. One way of finding out how much of the data should be truncated is by plotting the response variable versus time and locating the beginning of the region of steady state condition.

Start-up policies in simulation and reducing the effect of initial transient state were surveyed by Wilson and Pritsker (*11,12*). They found that deleting data from the beginning of the simulation output to reduce initial transient effect (bias reduction) causes loss of information and an increase in the variance. The net effect of deleting the initial observations was to increase the mean square error of the sample mean. Starting from empty and idle condition (no truncation of initial condition) provided a lower estimate of mean square error. Considering bias, variance reduction, and mean square error, they suggested starting as close to the steady state mode as possible and keeping all data. Furthermore, they stated, "The judicious selection of an initial condition appears to be more effective than truncation in improving the performance of the sample mean as an estimator of the steady-state mean." Also, the research has indicated that for small and well-behaved models truncation should not be performed, but for large models this may not be the case (*13*). Even after truncating the transient state, the simulation results will continue to fluctuate because of the stochastic variable used in the model.

Kleijnen (*14*) discussed that replication of simulation runs is the only alternative for gathering statistics about terminating systems. In a terminating system, the simulation run ends if a specific event occurs. In order to obtain independent replications, different random number seeds ought to be used. Initialization for replicated runs is performed by throwing away some observations from the beginning of each run. The rule-of-thumb is that the observations may be discarded as long as the response variable continues to increase or decrease. From several replications, the response can be obtained for each run and then the mean and the variance of the responses can be found. For independent results, a *t*-test can be run and the confidence level for the mean of observations can be found. However, in simulation the results do not always meet the requirements of a traditional *t*- or *f*-test. Kleijnen (*15*) provided techniques to compare means and variances of two simulations in which the outcomes from replications are pairwise correlated. To compare the means of autocorrelated observations of two simulation experiments, Fishman (*16*) suggested that the difference of the sample means be treated as a normal variate for a sufficiently long sample record. This procedure is not as good as comparing autocorrelation structure of the two experiments, but it suffices as an initial step for comparison (*15,16*).

When two operating conditions are to be compared, one can introduce a negative correlation between replications of runs under one operating condition to reduce the variance for within runs; and then introduce a positive correlation between runs under different operating conditions to reduce the variance for the difference between runs. This procedure is suggested as an efficient experiment design if there is not an initial transient phase (*17*). This variance reduction method uses antithetic variate and common random numbers jointly. Kleijnen (*18*) discusses possible undesirable effects of such a combination and indicates that, because of cross correlations, the results may even be worse than using each method alone. He compared the three methods for different conditions but could not determine which method was best for all systems.

SUGGESTED APPROACH

The suggested approach has two major tasks: verification and validation. Each task is performed at microscopic and macroscopic levels. At each level, several steps are suggested and application of some of them is discussed. The verification task includes efforts similar to those suggested by Sargent (*4*) for the conceptual and computerized validation, and the validation task is comparable to the operational validation discussed by Sargent (*4*).

Some forms of the steps suggested have been used in validation of other traffic simulation models. The author has greatly benefited from numerous studies in developing this approach. For instance, trajectory comparison and speed fluctuation were used in validation of INTRAS (*19,20*). In validation of other simulation models, comparable efforts have been made to ensure validity of the simulation results. The steps suggested here should be considered as the starting point for the efforts needed for verification and validation of a microscopic traffic simulation model, in general. For a particular simulation model, the developer should decide how suitable the suggested approach is, and take additional appropriate steps needed to obtain reliable results.

In the following sections, the verification task will be outlined, because of space limitation, and the validation task will be discussed in detail. Benekohal (*2,21*) provided further information on the verification and validation steps. The field data needed and the data used in validation of CARSIM are briefly described in the following sections.

Field Data

The data used for verification and validation should be accurate and appropriate for the model. A data set different than the one used for verification should be used for validation of the model. The data sets used for validation of CARSIM are the Ohio State University trajectory data collected using aerial photogrammetric techniques (*22*). The aerial photographs were taken in 1-sec time intervals from an elevation of about 3,000 ft. The location of a vehicle was determined with an accuracy of ±0.50 ft and the speed was determined with an accuracy of ±1.0 mph.

The field data provided a complete record of spacings, headways, longitudinal positions, and velocities for individual vehicles. The data were from median lane of I-71, near Columbus, Ohio, with normal traffic. Four platoons were used in validation of CARSIM. The platoon selected for presentation was Platoon 123, which went through a severe disturbancc. Thc platoon had 15 vchiclcs with no vehicle entering or leaving the platoon.

Model Verification

Verification ensures that the model behaves as intended. The program should be debugged first to eliminate any coding errors and programming problems. Then, the logic of different components of the model, such as car following, lane changing, merging, or diverging, should be carefully reviewed. Also, the acceleration and deceleration patterns, velocity change patterns, trajectory plots, and headways obtained from the simulation model should be examined. Sensitivity of these parameters to changes in the input variables should be studied. Model calibration, if needed, should be performed using field data for fine tuning of some of the variables in the model.

For verification of CARSIM, the following parameters were systematically varied and the sensitivity of the model outputs to the changes were carefully examined: maximum deceleration rate of a vehicle, compliance level of drivers, start-up delay of stopped vehicles, reaction time of drivers, buffer space between vehicles, and traffic mix. Then, different disturbances in traffic flow were induced to a platoon of 15 vehicles and their effects at microscopic and macroscopic levels were examined. A disturbance is induced when the leader of a platoon is required to decelerate, stop, and accelerate to a specified speed.

For the microscopic level verification, the effects of a regular disturbance, an emergency stop, and a stop-and-go operation on individual vehicle's trajectory, speed, and acceleration or deceleration were analyzed at high- and low-volume levels. For the macroscopic level verification, the effects of the regular disturbance and an emergency deceleration on average speed, density, volume, and average headway were examined at high- and low-volume levels. For example, the acceleration and deceleration patterns for a 15-car platoon in a stop-and-go condition are shown in Figure 1. The patterns are shown for the 1st, 4th, 10th, and last car in the platoon. Benekohal (*21*) provided further information on verification.

Model Validation

Two levels are proposed for validation of a microscopic traffic simulation model: (a) microscopic level, and (b) macroscopic level. At the microscopic level, the attributes of individual vehicles such as location, time, headway, and speed computed from the simulation model are compared with those obtained from the field data. At the macroscopic level, the aggregate parameters such as the average speed, density, and volume of a platoon of vehicles computed from the simulation model are compared with the results from field data. The steps for validation of CARSIM are shown in Figure 2.

For validation of CARSIM, four platoons covering a wide range of traffic conditions were used. The results for one of the platoons (Platoon 123) are presented here. The results for the other platoons are given elsewhere (*2,21*). Five independent replications were made for each traffic condition using different random number seeds. In each one of these runs, the attributes of all vehicles were generated randomly from the respective distributions. However, the location and speed of the leader of the simulated platoon were set equal to that of the leader of the field data. Thus, the leaders of the platoons had the same location and speed, but the followers may or may not.

It is recommended to use different platoons and for each platoon to make independent replications. The number of platoons and replications would depend on the system to be simulated and the range of variation of the response variable. The platoons should represent the real-world traffic conditions that the model is likely to simulate. When the range of the responses from different replications is narrow, a few runs might be enough. However, when a large variation is observed, more replications are needed. This topic will be discussed more in the macroscopic validation section.

Validation at Microscopic Level

The microscopic level validation is, perhaps, the most difficult task in validating a traffic simulation model. For microscopic validation, the variation of some or all of the following parameters should be examined: speed profile of an individual vehicle, location of the vehicle on the road, time headways between vehicles, and spacing between successive vehicles. For microscopic level validation of CARSIM, the speed change patterns and trajectory plots generated by CARSIM were compared with those obtained from the field data. The speed and location of every vehicle were determined at 1-sec time intervals in the simulation and field data. In the following sections, the changes on these variables will be examined.

Speed Profile Comparison

The speed of an individual vehicle was computed at 1-sec time intervals and a speed profile for each vehicle was generated by plotting speed versus time. The average speed of a vehicle

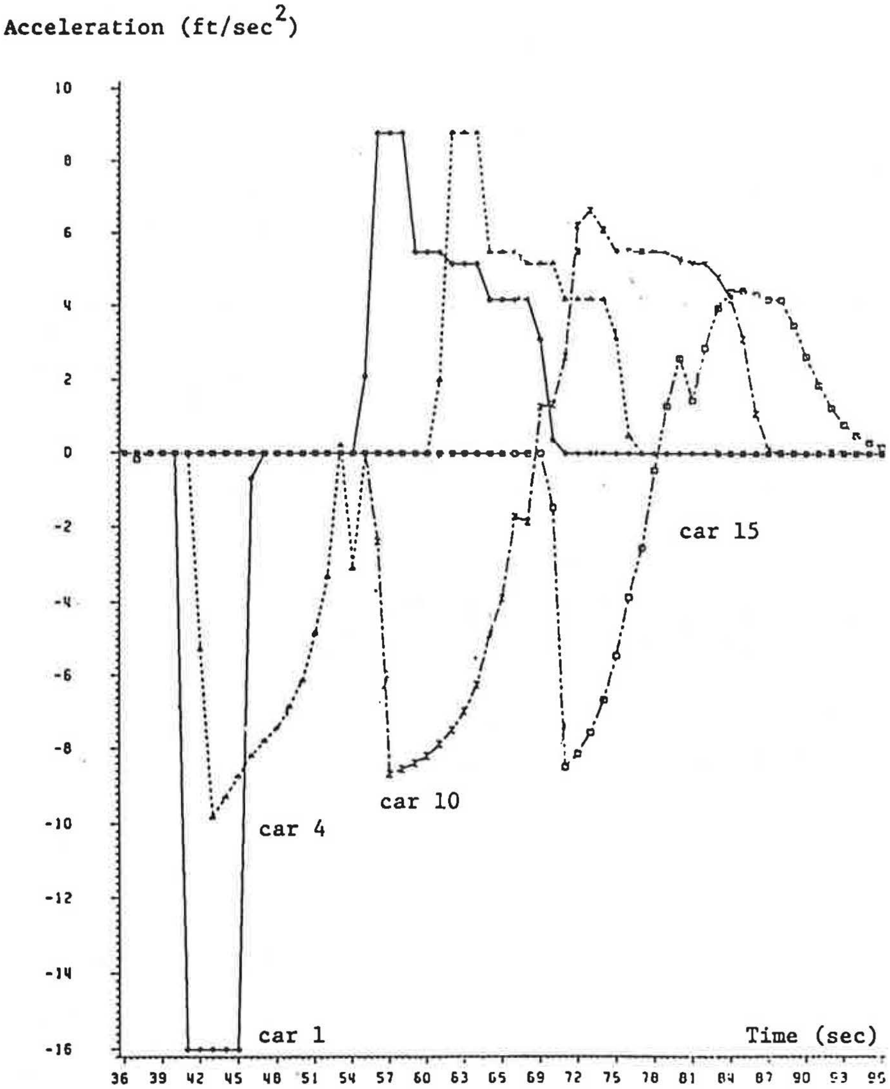


FIGURE 1 Acceleration-deceleration patterns for a platoon of vehicles in a stop-and-go condition. The leader of the platoon decelerates at 16 ft/sec^2, stops for 9 sec, and accelerates to a desired speed.

at a given time interval was computed as the mean of five speeds from the replications. Then, the average speed of an individual vehicle at a given time was compared to the observed speed from field data. Comparison of the speeds from simulation and field data are shown in Figure 3, which shows the speeds for the lead car, the 4th car, and the last (15th) car of Platoon 123. The other cars could not be shown because the plot became cluttered.

The simulation model generated speed change patterns similar to those from the field data. All vehicles came to a complete stop and then accelerated to reach their desired speed. The similarity of the speed change patterns indicates that the simulation model replicates the real-world traffic disturbance with an acceptable accuracy. An acceptable level of accuracy would depend on the model and the purpose the model is used for. Here, the similarities of speed change patterns were considered the important criteria in accepting accuracy of the model. Speed difference at each time interval may also be considered as the criteria; however, the difference may exhibit a large fluctuation because the speeds are updated at short time intervals (e.g., 1 sec). In the simulated platoon, the vehicles exhibited less speed fluctuation than the vehicles in the field data, because of the fact that the model was programmed to mimic only limited characteristics of real drivers.

From the comparison of speed change patterns, one may conclude how well the model duplicates the real-world speed change patterns in various traffic conditions. The criteria for the operational validation may be comparison of the shape of the speed profile for a vehicle generated by the model and the field data. The shift between the two profiles for a vehicle is not as important as the similarity of the shape of the profiles, because the shift would depend on the drivers' characteristics, but the profile would reflect the model's capability to simulate actual speed profiles. A driver with a longer reaction time would cause a larger shift than a driver with a shorter reaction time, but the profile for both drivers may look similar.

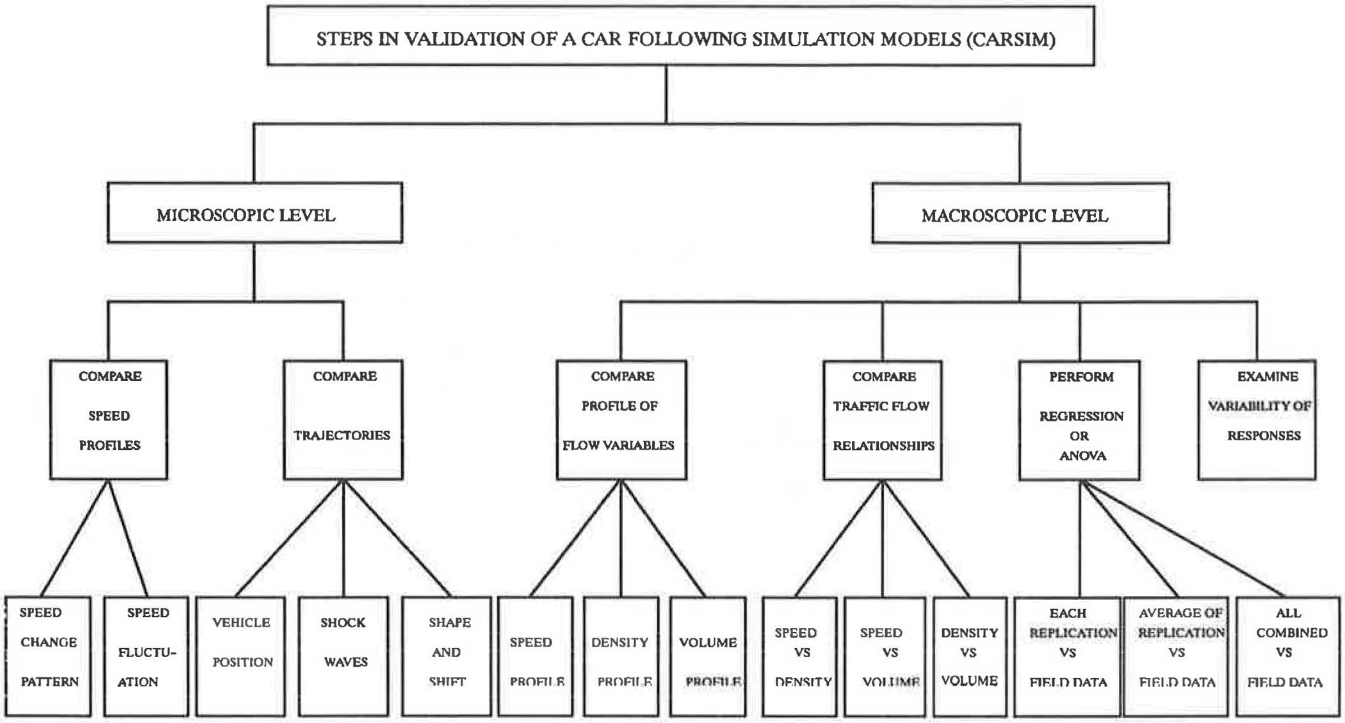

STEPS IN VALIDATION OF A CAR FOLLOWING SIMULATION MODELS (CARSIM)
MICROSCOPIC LEVEL
MACROSCOPIC LEVEL
COMPARE SPEED PROFILES
COMPARE TRAJECTORIES
COMPARE PROFILE OF FLOW VARIABLES
COMPARE TRAFFIC FLOW RELATIONSHIPS
PERFORM REGRESSION OR ANOVA
EXAMINE VARIABILITY OF RESPONSES
SPEED CHANGE PATTERN
SPEED FLUCTU- ATION
VEHICLE POSITION
SHOCK WAVES
SHAPE AND SHIFT
SPEED PROFILE
DENSITY PROFILE
VOLUME PROFILE
SPEED VS DENSITY
SPEED VS VOLUME
DENSITY VS VOLUME
EACH REPLICATION VS FIELD DATA
AVERAGE OF REPLICATION VS FIELD DATA
ALL COMBINED VS FIELD DATA

FIGURE 2 Steps in validation of the car-following simulation model (CARSIM).

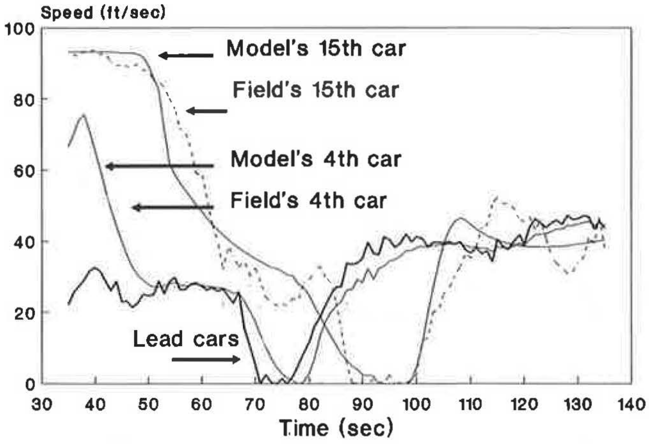


FIGURE 3 Comparison of speed change patterns from the simulation model versus field data for the first, fourth, and last cars in Platoon 123.

Trajectory Comparison

A vehicle trajectory was obtained by plotting the position of the vehicle versus the time in 1-sec time intervals. The average of five numbers obtained from the replications was used as the location of the vehicle at a given time. For Platoon 123, the trajectory plots for every third vehicle including the last vehicle are shown in Figure 4.

When there is a severe disturbance in the traffic flow, it is challenging for a simulation model to generate trajectories that are close enough to the actual trajectories. However, in normal traffic conditions it is not difficult to obtain trajectory plots that are close to the trajectory plots from field data. Thus, models validated only in free flow traffic conditions may not accurately simulate high-density traffic conditions. Also, models validated using data only either from the acceleration or from the deceleration phase of a traffic disturbance may not accurately simulate traffic flow in stop-and-go conditions.

The criteria for evaluating the similarity between the model and field data may include vehicle location, difference in location between the model and real platoon, shape of the plot for individual vehicles, general shift up or down, shift either before or after the disturbance, and location where a vehicle slows down, stops, starts, and recovers. It is important to use short time intervals (a few seconds) in generating trajectory plots, if the model will be used for detailed studies. Otherwise, the model may not accurately show the behavior of traffic within that interval. For instance, Figure 4 shows that the vehicles in the platoon stopped and moved in less than 20 sec. If the data had been collected every 30 sec, the stop-and-go behavior of the platoon would have been missed.

The microscopic validation may be conducted subjectively or objectively, as suggested by Sargent (*4*). The graphical comparison of (subjective validation) the speed change patterns and trajectory plots was found to be sufficient at this level. The objective validation (statistical technique) was not used because of a strong correlation between successive points on the speed profile or the trajectory plots for a vehicle. Once satisfactory results were obtained from the microscopic level validation, the macroscopic validation was started.

Validation at Macroscopic Level

For macroscopic validation, the overall performance of a platoon of vehicles should be evaluated rather than the perfor-

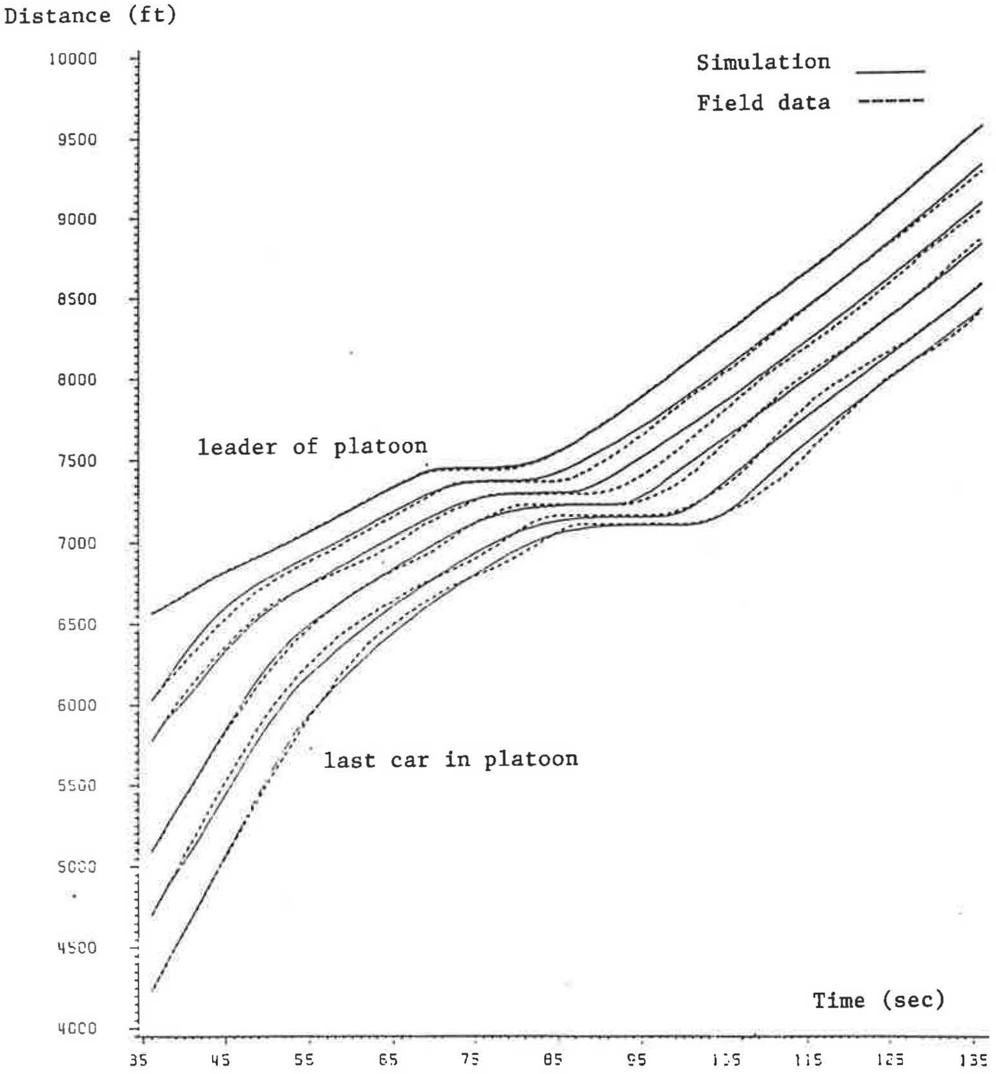


FIGURE 4 Comparison of trajectories of vehicles from the simulation model versus field data for Platoon 123. The trajectories are shown for every third car in a platoon of 15 cars.

mance of an individual vehicle. The macroscopic level validation may not reveal as much detailed information about the model capabilities as the microscopic level, because the variables are the average values for all vehicles in the platoon. For instance, the average speeds for a simulated and an actual platoon might be close, but the speed of individual vehicles may still be different. Likewise, the density of a simulated platoon might be close to that of a field platoon, but spacing between successive vehicles may be considerably different.

For the macroscopic level validation, the following comparisons are suggested:

1. Comparison of profile of traffic flow variables,
2. Comparison of fundamental relations of traffic flow, and
3. Comparison of simulation results versus field data.

In addition to the comparisons, the range of variation of the response variables should also be examined. Application of these concepts in validation of CARSIM is discussed in the following sections.

Comparison of Profile of Traffic Flow Variables

Traffic flow variables used for the comparison were the speed, density, and volume. These variables were computed at 1-sec time intervals and their variations over time (profile) were compared to the field data. The plot of the average speed from the simulation runs versus the speed from the field data for Platoon 123 is shown in Figure 5. The platoon suffered from a severe kinematic disturbance and recovered immediately. The platoon traveling at a speed of more than 80 ft/sec reached a speed of near zero in less than 1 min. The simulation results are close to the actual traffic speeds. The simulation curve exhibits less local fluctuation than the curve for the field data, as expected.

The plots of density versus time for Platoon 123 and the simulation counterparts are shown in Figure 6. The graphs show the same patterns and fluctuations for both simulated and actual platoons. The time a simulated platoon reaches the jam density is close to that of the actual platoon. The density of a platoon is computed from the distance between

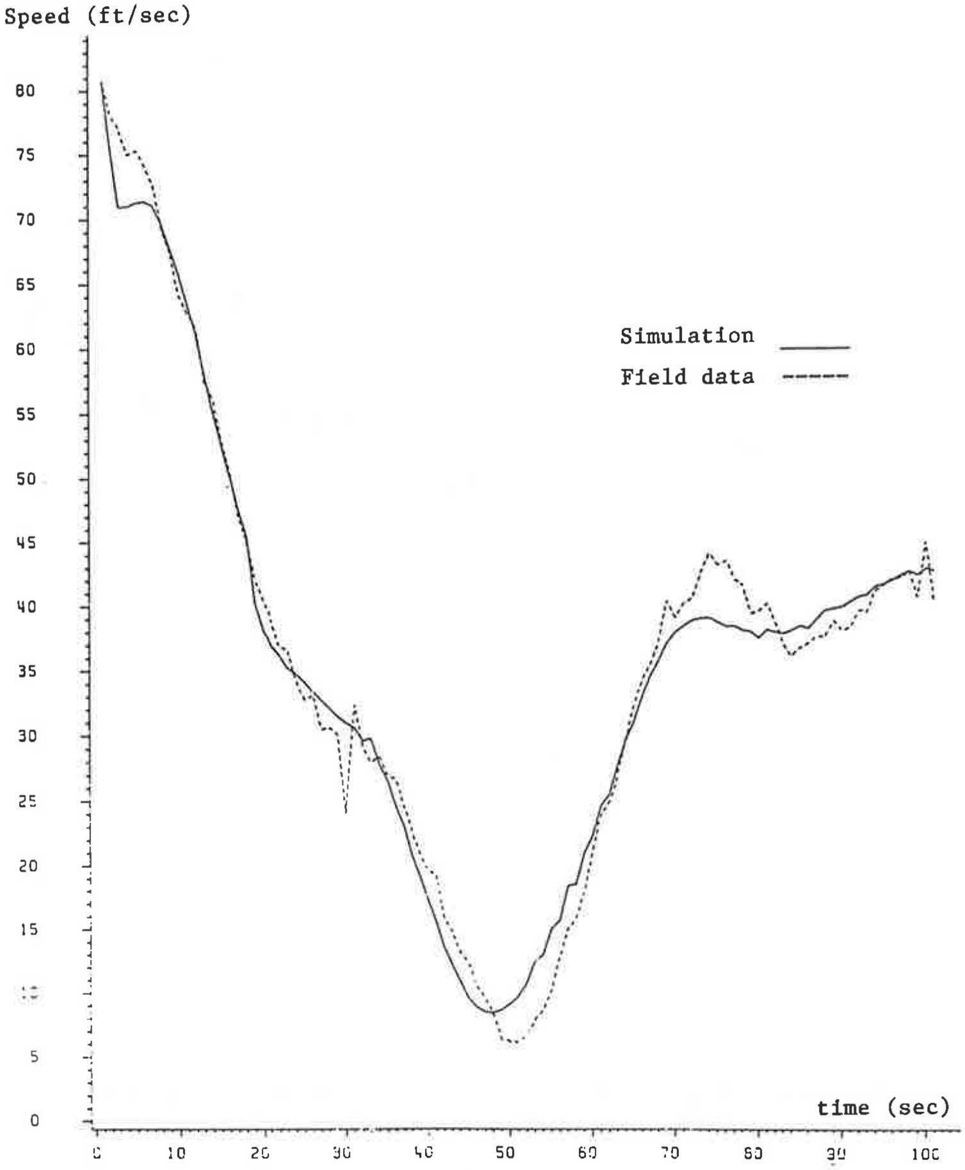


FIGURE 5 Comparison of the average speed from the simulation model versus field data for Platoon 123.

the first and the last car in the platoon. The distance is dependent on the spacing between these two cars. Therefore, one should be careful in using the density of a platoon for comparison of actual and simulated results.

Comparison of volume profile for the simulated and actual platoon is shown in Figure 7. The volume is computed as product of speed and density. The difference between the simulated and actual volume may be large because of the multiplication. Another reason for the large difference might be that the volume may not be equal to the product of speed and density when traffic flow breaks down (critical density). Thus, the volume comparison is not recommended when traffic density reaches its critical range.

Fundamental Relations of Traffic Flow

At this level, the fundamental relationships between traffic flow parameters (speed, density, and volume) obtained from the actual and simulated platoons should be examined. The speed-density, speed-volume, and density-volume relationships from the field data should be compared with those from the simulation model. The speed-density relationships for simulated and actual platoons are shown in Figure 8. The actual data exhibited a nonlinear relationship between speed and density, and a loop representing the hysteresis phenomenon (*19*) when the traffic flow breakdown occurs. The simulation results exhibit a similar relationship and the same phenom-

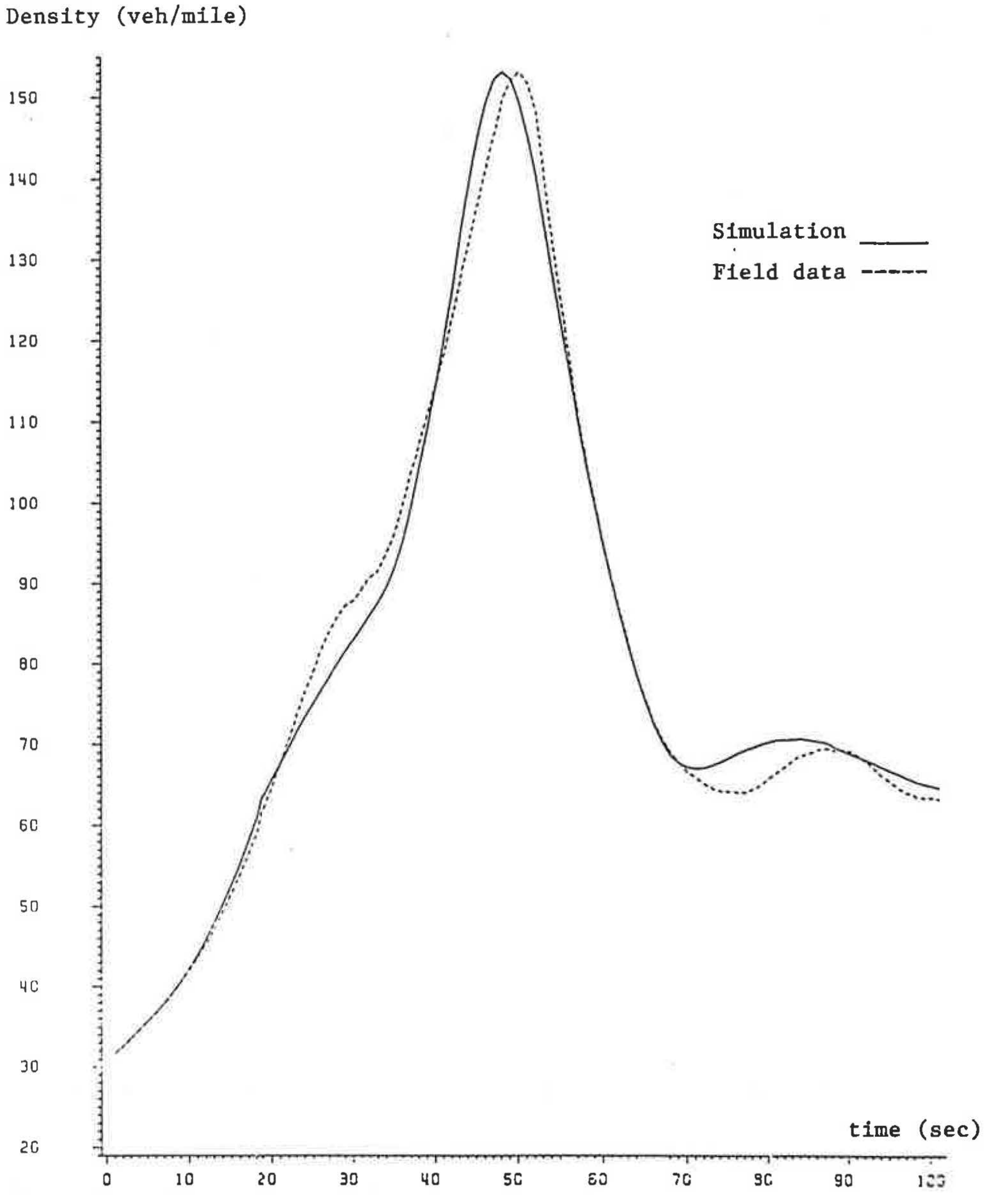


FIGURE 6 Comparison of density from the simulation model versus field data for Platoon 123.

enon. The loop from the simulation model is less distinct than that of the field data because the simulation results are the average of five replications.

Similar comparisons should be made for speed-volume and volume-density relationships. In addition to the graphical presentation of the results, the statistical analysis of the simulation results versus the actual data is also carried out. The statistical analyses will be discussed in the following section.

Comparison of Simulation Results Versus Field Data

Traffic parameters computed from the simulation model should also objectively (e.g., statistical analysis) be compared to the values from field data. Statistical techniques such as regression analysis, analysis of variance, or time series analysis may be used for comparison of the results. The appropriate method should be selected on the basis of factors such as type of data available, relationship with the other parameters, dependency to the other variables, etc. Application of regression analysis and analysis of variance (ANOVA) for model performance evaluation are discussed here. For comparison of the model performance, regression analysis of speed, density, and volume computed from the simulation model versus those from the field data is carried out. For each time interval, the average speed, density, and volume were computed from simulation and field data. The general form of the regression lines is

$$P_{\text{model}} = \text{b}0 + \text{b}1 * (P_{\text{field}})$$

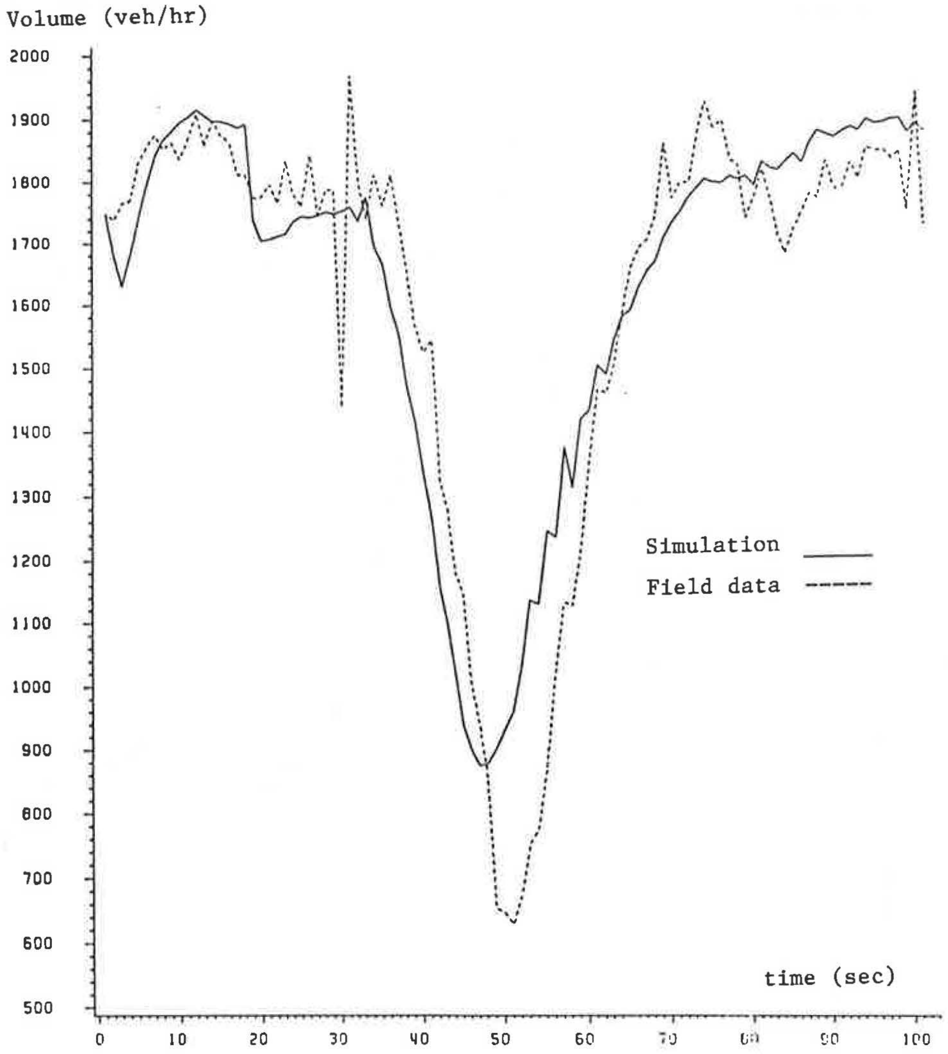


FIGURE 7 Comparison of volume from the simulation model versus field data for Platoon 123.

where

P_{model} = Speed (density or volume) from simulation model,
P_{field} = Speed (density or volume) from field data,
b0 = Y-intercept of the regression line, and
b1 = Slope of the regression line.

Independent replications must be made for any simulation model with stochastic parameter to account for variability of the parameters. However, replication of simulation runs creates more than one set of data for a given condition. One must be careful in interpreting the differences among the replications. The following comparisons may be considered among independent replications:

1. Comparison of an individual replication versus field data,
2. Comparison of average of replications versus field data,
3. Comparison of all replications combined versus field data.

For comparison of individual replication versus field data, speed, density, and volume computed from each simulation run are regressed over the values from field data. The coefficients of the regression lines, variance of the coefficients, and R^2 values are presented in Table 1. Note that, s(b0) and s(b1) are the variances of b0 and b1. The R^2 values and the coefficients of the regression lines indicate that there is a strong agreement between the simulation results and the field data. The slope and y-intercept of the regression lines as well as the R^2 values for a given parameter do not exhibit a large variation among the replications. When the variation is large, more replications should be made.

The advantage of using regression of individual runs over using the average of the five replications is that, one would get additional information about the variation of the response variables among the replications. When the average values are used, this information is no longer available; however, the results are easier to interpret.

For comparing the average of replications versus field data, the speed, density, or volume at a given time is computed as the average values from five replications. Then, the average of five replications is compared to field data. Table 2 presents some of the parameters obtained from the regression analysis. The slopes of the regression lines are close to 1 and the y-

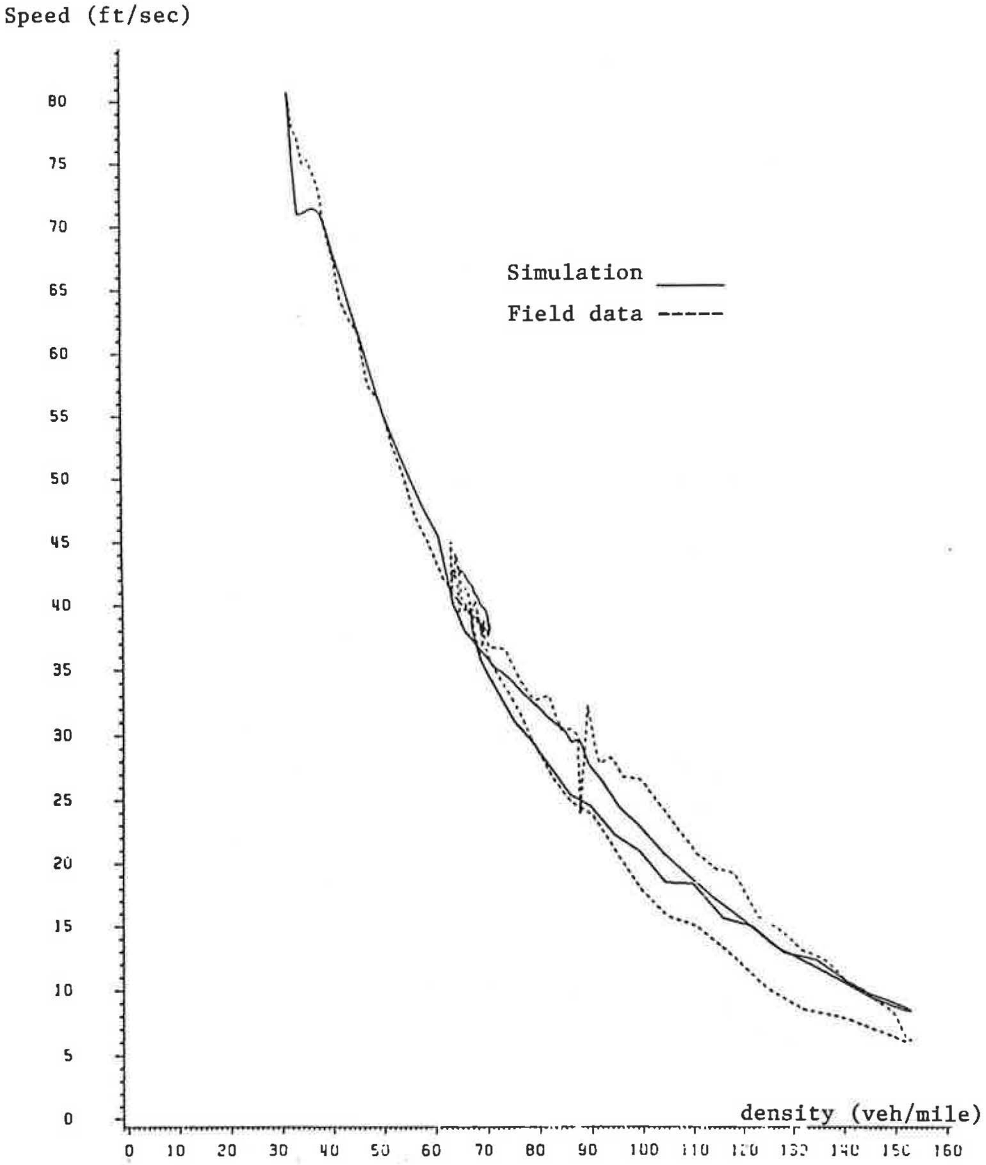


FIGURE 8 Comparison of speed-density relationships from the simulation model versus field data for Platoon 123.

intercepts are close to 0. The R^2 values for speed and density are 0.98 or higher, and for volume it is 0.80 or higher. The results from Table 2 indicate that there is a strong agreement between the speeds or densities computed from the simulation and the field data. As discussed before, the agreement between the traffic volumes is not expected to be as strong as that of speed or density.

When the average values of several replications were compared to field data, the parameters indicated less variation than the same parameter in the individual runs. Although using the average values of the replications makes the comparisons less complicated, finding average of several replications may conceal useful information about the sensitivity of a parameter to an input variable. For instance, the regression lines for speed in Table 1 indicate the range of variation of slope and intercept, but the regression line for speed in Table 2 does not have any variation. Because there are five values from the simulation model for each value from the field data, one might treat them as repeated observations at a given point. The consequence of this assumption is discussed in the following section.

For comparison of all replications combined versus field data, the values obtained from the individual simulation runs are combined to create one data set. Regression analysis of the combined data set versus field data yielded, as it was expected, slopes and y-intercepts equal to that of the average of five replications, see Table 2. However, the variance of slope and y-intercept decreased almost to one-half of that of the average of five replications. There was also a slight decrease on the R^2 values.

The assumption about the repeated data was proven to be incorrect. In regression analysis when there are repeated ob-

TABLE 1 PARAMETERS OF REGRESSION LINES FOR INDIVIDUAL SIMULATION REPLICATIONS VERSUS THE VALUES FROM FIELD DATA FOR PLATOON 123

Repli-cations	b0	b1	R^2	S(b0)	S(b1)
			SPEED		
1	1.09696	0.96812	0.98129	0.54144	0.01344
2	1.96750	0.94091	0.98012	0.54266	0.01348
3	1.91585	0.93815	0.98287	0.50152	0.01245
4	2.44674	0.91579	0.98478	0.46103	0.01144
5	1.10296	0.97226	0.97963	0.56778	0.01409
			DENSITY		
1	1.90606	0.98211	0.98744	0.95538	0.01113
2	2.05298	0.95981	0.98826	0.90241	0.01051
3	1.40735	0.97494	0.99118	0.79343	0.00924
4	2.32060	0.94958	0.99130	0.76726	0.00894
5	-0.49143	1.05514	0.97827	1.35641	0.01580
			VOLUME		
1	353.88146	0.79593	0.81033	64.38405	0.03870
2	277.90502	0.82216	0.82465	63.38825	0.03810
3	415.23342	0.74210	0.82827	56.49940	0.03396
4	383.34560	0.74492	0.88891	44.03049	0.02647
5	316.72237	0.85828	0.79982	71.79334	0.04315

TABLE 2 PARAMETERS OF REGRESSION LINES FOR THE AVERAGE OF FIVE REPLICATIONS AND ALL FIVE REPLICATIONS COMBINED VERSUS THE VALUES FROM FIELD DATA FOR PLATOON 123

Type of Analysis	Variable	b0	b1	R^2	s(b0)	s(b1)
Average of 5 Runs	Speed	1.70554	0.94705	0.98383	0.49177	0.01220
	Density	1.43827	0.98432	0.98922	0.88620	0.01033
	Volume	349.433	0.79265	0.84406	56.9654	0.03424
All 5 Runs Combined	Speed	1.70655	0.94705	0.98100	0.23682	0.00588
	Density	1.44075	0.98432	0.98123	0.52091	0.00607
	Volume	349.472	0.79267	0.80913	28.5566	0.01717

servations at a given point, the lack of fit of the regression line should be examined (*23*). The examination of ANOVA tables falsely indicated the lack of fit for the linear model.

The possibility of fitting a higher-order model was explored, and the residuals were plotted versus the time and predicted values. The plots did not exhibit any definite trend. The residuals were clustered around the line $y = 0$, and approximately made a horizontal band. The lack of the trend, presence of the band, and an R^2 value close to 1 indicated the adequacy of the model. Thus, there was no lack of fit.

The lack-of-fit test is not appropriate for this situation because there is a strong dependency between successive data points. For instance, the density at the current time interval depends on the density at the previous time interval. The false detection of lack of fit was not caused by the inadequacy of the linear regression model, but by using the lack-of-fit test when the data points were strongly correlated.

Variation of the Response Variables

The result obtained from a single simulation run may not be reliable when stochastic variables in the model affect the model's outcome. One method to increase reliability and confidence on the simulation responses is to make several independent replications. Then, the change in the response variable among the replications should be examined. Plots indicating the range of variation of speed, density, and volume among different replications were prepared. The range of variation of speed is shown in Figure 9. Figure 9 indicates that there was small variation from one simulation run to another. When a wider variation is observed, more runs should be made. The number of replications depends on the model and the range of the output parameter.

CONCLUSIONS AND RECOMMENDATIONS

A model verification and validation procedure is suggested and its application to CARSIM is presented. The procedure divides the tasks into microscopic and macroscopic levels. This paper concentrated on the validation part and briefly discussed the verification efforts.

For validation at the microscopic level, the speed change patterns and trajectories of vehicles obtained from the simulation models were compared with those from field data. For validation at the macroscopic level, the average speed, density, and volume computed from the simulation were compared to the field data. The variation of these parameters over time and the relationships between these parameters were examined. Furthermore, the regression of the simulation results versus the field data was discussed. Comparisons of the results indicated that using this procedure enabled CARSIM to provide results close to those from field data.

Some of the steps suggested here have been used in validation of other traffic simulation models. The steps suggested here should be considered as the starting point for the efforts needed for verification and validation of a microscopic traffic simulation model. For a particular simulation model, the developer should decide on appropriateness of the suggested approach and take additional steps needed to obtain reliable

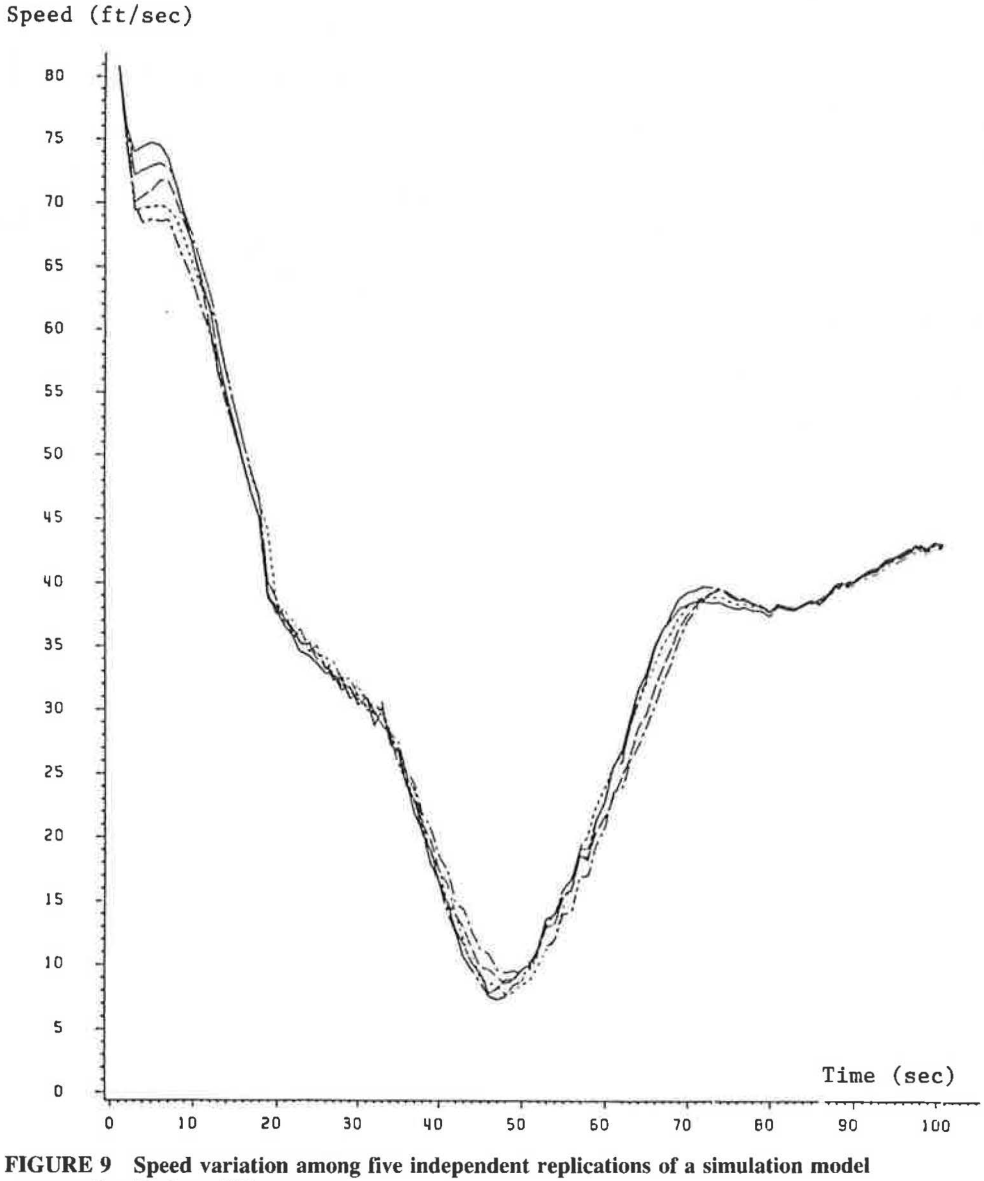


FIGURE 9 Speed variation among five independent replications of a simulation model representing Platoon 123.

results. Questions about number of replications, truncation policy, and traffic conditions to be covered were not discussed. The procedure is intended to be used for validation of microscopic models, although some parts of it may be used for validation of macroscopic models, as well. The approach should be considered as a step toward the development of a comprehensive guideline for validation of traffic simulation models.

REFERENCES

1. G. S. Fishman and P. J. Kiviat. *The Statistics of Discretion-Even Simulation*. SCI Simulation 10, 1968.
2. R. F. Benekohal and J. Treiterer. *A Car Following Model For Simulation of Traffic Flow in Normal and Stop-And-Go Conditions*. In *Transportation Research Record 1194*, TRB, National Research Council, Washington, D.C., 1989.
3. J. S. Annino and E. C. Russell. Ten Most Frequent Causes of Simulation Analysis Failure-and How to Avoid Them. *Simulation*, Vol. 32, No. 6, 1979.
4. R. S. Sargent. Verification and Validation of Simulation Models. *Progress in Modeling and Simulation*, F. E. Cellier (ed.), Academic Press, New York, 1982.
5. J. F. Torres, A. Halati, and A. Gafarian. *Statistical Guidelines for Simulation Experiments*. Executive Summary, JFT Associates, Culver City, 1983.
6. A. V. Gafarian and J. D. Walsh. Methods for Statistical Validation of a Simulation Model for Freeway Traffic Near an On-Ramp. *Transportation Research*, Vol. 4, 1970, pp. 379–324.
7. G. A. Mihram. *Simulation Statistical Foundations and Methodology*. Academic Press, New York, 1972.
8. J. P. C. Kleijnen. Statistical Design and Analysis of Simulation Experiments. *Informatie*, Vol. 17, No. 10, Netherlands, Oct. 1975, pp. 531–535.
9. T. H. Naylor, K. Wertz, and T. H. Wonnacott. Methods for Analyzing Data From Computer Simulation Experiments. *Communications of the ACM*, Vol. 10, No. 11, 1967.
10. J. P. C. Kleijnen. Experimentation with Models: Statistical Design and Analysis Techniques. In *Progress in Modeling and Simulation*, F. E. Cellier (ed.), Academic Press, New York, 1982.
11. J. R. Wilson and A. B. Pritsker. *A Survey of Research on the Simulation Start Up Problem. Simulation*, Vol. 31, No. 2, Aug. 1978.
12. J. R. Wilson and A. B. Pritsker. *Evaluation of Start Up Policies in Simulation Experiments. Simulation*, Vol. 31, No. 3, Sept. 1978.
13. A. A. B. Pritsker and C. D. Pegden. Introduction to Simulation and SLAM. John Wiley, New York, 1979.
14. J. P. C. Kleijnen. *Design and Analysis of Simulations: Practical Statistical Techniques. SCI Simulation*, Vol. 28, No. 3, March 1977.
15. J. P. C. Kleijnen. *Comparing Means and Variances of Two Simulations. SCI Simulation*, Vol. 26, No. 3, 1976.
16. G. S. Fishman. *Problems in the Statistical Analysis of Simulation Experiments: The Comparison of Means and the Length of Sample Records*. Communications of the ACM. Vol. 10, No. 2, 1967, pp. 94–99.
17. J. R. Emeshoff and R. L. Sisson. *Design and Use of Computer Simulation Models*. McMillan Co., New York, 1970.
18. J. P. C. Kleijnen. *Antithetic Variates, Common Random Numbers and Optimal Computer Time Allocation in Simulation. Management Science*, Vol. 21, No. 10, 1975.
19. D. Wicks et al. *Development and Testing of INTRAS, a Microscopic Freeway Simulation Model, Volume 1 and 2, Final Report*. FHWA/RD-80/106-107. FHWA, U.S. Department of Transportation, 1980.
20. R. B. Goldblatt. *Development and Testing of INTRAS, a Microscopic Freeway Simulation Model, Volume 3, Validation and Application, Final Report*. FHWA/RD-80/108. FHWA, U.S. Department of Transportation, 1980.
21. R. F. Benekohal. *Development and Validation of a Car Following Model for Simulation of Traffic Flow and Traffic Waves Studies*. Ph.D. dissertation, Ohio State University, Columbus, 1986.
22. J. Treiterer. *Investigation of Traffic Dynamics by Aerial Photogrammetry Techniques*. Final Report EES 278. Transportation Research Center Department of Civil Engineering, The Ohio State University, Feb. 1975.
23. N. Draper and H. Smith. *Applied Regression Analysis, 2nd ed.* John Wiley, New York, 1981.

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

Enhancement and Field Testing of a Dynamic Freeway Simulation Program

Panos Michalopoulos, Eil Kwon, and Jeong-Gyu Kang

An enhanced version of KRONOS, a dynamic, microcomputer-based freeway simulation program, is presented. KRONOS is based on mesoscopic continuum modeling and explicitly treats the coupling effects of ramps on the freeway flow in detail. The major improvements of the new version include the substantial reduction of computer execution time and the addition of new simulation modules that can describe the flow behavior under restricted capacity situations, such as incidents and short-term construction. Further, a newly developed output interface links KRONOS with other advanced data management and analysis software, so that more detailed and flexible output analysis can be possible. In order to test and validate the program in real traffic environments, field data collection was performed at 26 locations covering an 8-mi section of the I-35W freeway in Minneapolis, Minnesota. The data consist of 5-min volume and speed information for each lane during three morning and three afternoon peak periods. The test results with the available data indicate satisfactory performance of the enhanced KRONOS program in light-to-moderate traffic conditions, although the error increased during heavy congestion.

A major problem in freeway operations and traffic management practice is to assess the effectiveness of changes or improvements before implementation. This process includes comparing alternative geometric configurations, estimating the effects of improvements, determining the adequacy of traffic management schemes, assessing the impacts of control strategies, studying the formation and dissipation of congestion on the freeway and its ramps, etc.

Although simulation methods have long been recognized as the most powerful tool for such analysis, most existing freeway simulation programs are still either in the development stage (*1*) or lack the sophistication necessary for applications requiring reasonably high performance levels (*1,2*). Furthermore, despite recent advances in microcomputer technologies, the most widely known simulation programs developed to date require mainframe computers (*3–6*) that are not readily available to most practicing engineers.

The major recent enhancements and field testing results of KRONOS (*7*), a microcomputer-based, dynamic freeway simulation program, which was developed during the mid-1980s to address these problems, are described. KRONOS is based on continuum flow modeling, which treats traffic as a compressible fluid. Unlike other macroscopic simulation programs, KRONOS explicitly models interrupted traffic flow behavior such as merging, lane changing, weaving, and diverging (*7–11*). The coupling effects of ramps on the main freeway are considered in determining the actual entering and exiting flows as well as in following the simultaneous development of queues and propagation of congestion on both the freeway and the ramps.

Although KRONOS has been almost continuously enhanced since the inception of its initial version (*7*), the most notable enhancements were made during the past 2 years. This includes substantial reduction of computer execution time by reorganizing the program structure using the C programming language. Further, the input/output (I/O) modules were extensively modified to take advantage of recent hardware advancements, such as single-screen systems with high-resolution graphics for both text and graphics displays. In addition, an output interface was developed to create spreadsheet-formatted output files, which enable the user to use more advanced data management and analysis software including Lotus 1–2–3. Consequently, much more flexible and detailed output analysis is now possible. Perhaps the most important enhancement lies in the modeling adjustments through improved determination of the boundary conditions as well as the extension of the program capacity to simulate virtually all types of freeway geometric conditions, which include merging and diverging of freeways, two-lane entrance and exit ramps, lane addition and deletion, left side ramps, etc. In order to validate the simulation model, an extensive field data collection was conducted from the selected locations covering an 8-mi section of the I-35W freeway in Minneapolis, Minnesota, for six peak periods in November 1989. The collected data, which consisted of 5-min volume and speed information of the freeway main line as well as of entrance and exit ramps, has been used for the validation of the enhanced KRONOS program.

BACKGROUND

Existing macroscopic flow models fall into three general categories: (a) I/O, (b) simple continuum, and (c) high-order continuum. The models in the first category are rather simplistic in that they do not include space explicitly nor do they take compressibility into account. High-order models, on the other hand, are in principle more realistic as they include the effects of inertia and acceleration of the traffic mass; however, although the existing high-order models look promising, they have not as yet proved truly superior to the simple continuum alternative at least in medium-to-congested flow conditions (*7,8*). For this reason, the current version of KRONOS uses the simple continuum modeling that is based on the conservation equation and an equilibrium speed-density (or flow-density) relationship. However, the program's modeling

Department of Civil and Mineral Engineering, University of Minnesota, 500 Pillsbury Dr., S.E., Minneapolis, Minn. 55455.

methodology and structure are designed to allow use of high-order models if desired. The simple continuum modeling can be described by the partial differential equation

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = g(x,t) \qquad \text{with } q = ku = ku_e(k) \tag{1}$$

where

- k = density,
- q = flow,
- t = time,
- x = distance, and
- g = generation rate.

In order to implement this model, KRONOS uses a finite difference scheme specifically designed to solve one-dimensional, time-dependent, compressible flows containing strong shocks (*12*). With this scheme, the time and space domains are discretized in short increments, Δt and Δx, respectively, such that $\Delta x/\Delta t > u_f$, where u_f is the free-flow speed. Space discretization of a section including the most typical freeway components is shown in Figure 1. The relevant formulas are as follows:

$$k_j(n + 1) = \frac{1}{2}[k_{j+1}(n) + k_{j-1}(n)] - \frac{\Delta t}{2\Delta x}[q_{j+1}(n) - q_{j-1}(n)] - \frac{\Delta t}{2}[g_{j+1}(n) + g_{j-1}(n)] \tag{2}$$

$$u_j(n + 1) = u_e[k_j(n + 1)] \tag{3}$$

$$q_j(n + 1) = k_j(n + 1) * u_j(n + 1) \tag{4}$$

where

- $k_j(n)$, $u_j(n)$, $q_j(n)$ = traffic density, speed, and flow, respectively, of Node j at the nth time step (Δx = 100 ft, Δt = 1 sec);
- $u_e[k_j(n + 1)]$ = equilibrium speed-density relationship, which is location-specific and possibly discontinuous (*13*).

In KRONOS, the flow at the external boundaries (such as upstream and downstream ends of freeway, and the junctions of ramps with the adjacent surface streets) is specified from the arrival and departure pattern provided from input, whereas the flow at the internal boundaries (such as the metering stop-line and the beginning or ending node of deceleration and acceleration lanes, respectively) is determined from either the metering rates or other physical considerations.

In the following sections, the modeling methodology of the enhanced KRONOS is presented along with a description of the program, its capabilities, hardware and software requirements, testing and field validation results, and plans for future improvements.

MODELING METHODOLOGY OF THE ENHANCED KRONOS

The enhancements that have been made to the KRONOS simulation model since its introduction have been extensive and performed on an ongoing basis. So far, they include improving the treatment of internal boundaries and adding a spillback mechanism for the flow under capacity restraints, such as lane blockage and capacity reduction at the downstream off-ramp caused by construction or arterial traffic con-

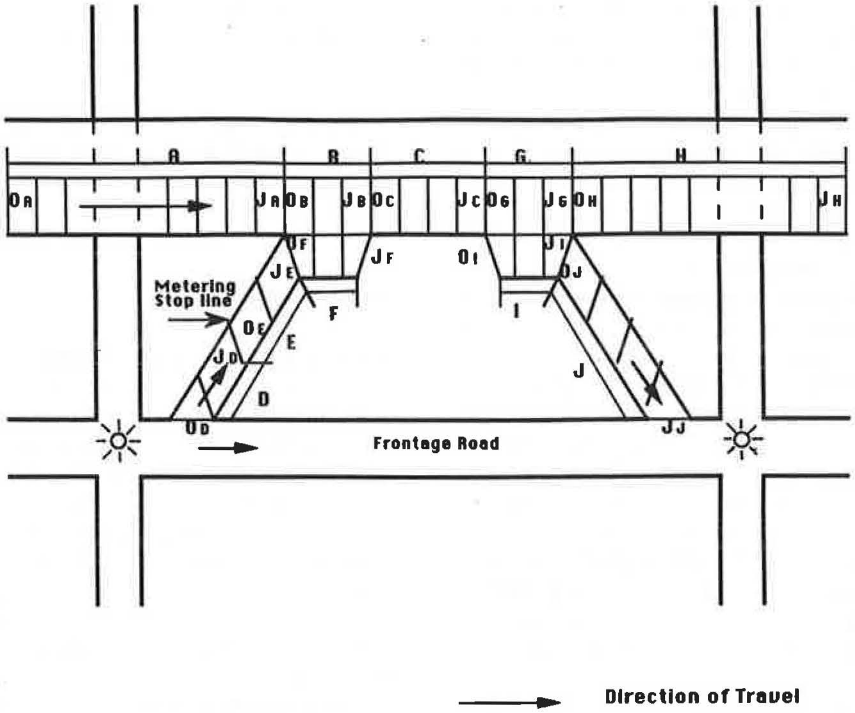


FIGURE 1 Space discretization of a typical freeway section.

ditions. In what follows, the basic modeling elements of the aforementioned enhancements are presented to illustrate the basic approach and allow the reader to assess the differences over the earlier modeling. The detailed KRONOS modeling and step-by-step algorithms can be found elsewhere (*10,14*); once again it is stressed that this section contains only the recent modeling improvements in KRONOS, which are summarized in the following sections.

Links *A*, *D*, *E*, *H*, and *J* (Figure 1)

1. Referring to Figure 1, these links are characterized by the absence of flow generation (dissipation). Calculations for the density of internal Nodes 1 to J_{M-1}, where $M = A, D, E, H$, or J, follow:

$$k_j(n + 1) = \frac{1}{2}[k_{j+1}(n) + k_{j-1}(n)] - \frac{\Delta t}{2\Delta x}[q_{j+1}(n) - q_{j-1}(n)] \qquad (5)$$

Speed and flow are calculated by using Equations 3 and 4, respectively.

2. Nodes O_A and O_D in Links A and D are external boundaries, so $q_{OA}(n + 1)$ and $q_{OD}(n + 1)$ are specified from the input demand patterns.

3. At the ending nodes in Link H and J, $q_{JH}(n + 1)$ and $q_{JJ}(n + 1)$ are specified from the downstream departure rates that can be either restricted with the specified input information or unrestricted. In case of unrestricted departure rates, $q_{JH}(n + 1) = q_{JH-1}(n)$, and $q_{JJ}(n + 1) = q_{JJ-1}(n)$. When there are capacity restrictions at the ending nodes, $q_{(JH \text{ or } J)}(n + 1) = q_{\text{cap, H or J}}(n)$, where $q_{\text{cap, H or J}}(n)$ is the restricted capacity at the last node of Link H or J.

4. Nodes J_D and O_E in Links D and E represent the stop line. If the freeway entering demand, $q_{JD-1}(n + 1)$, is less than the metering rate, node J_D is not treated as an internal boundary, and thus, the density, speed, and flow at each node in Links D and E are calculated by using Equations 2 to 4. If $q_{JD-1}(n + 1)$ is greater than the metering rate, then $q_{JD}(n + 1) = q_{OE}(n + 1)$ is the metering rate.

5. Following the definition of q, the densities of all the internal boundaries can be calculated from the equilibrium q-k relationship using the flow rates determined earlier.

6. Density, speed, and flow at the connecting Nodes J_A and J_E are calculated from Equations 2 to 4, respectively. Further, the ending and beginning nodes in consecutive links coincide, that is, J_A in Link A coincides with O_B in Link B, J_E with O_F in Link F.

Link F

1. This link represents the acceleration lane and therefore has dissipation terms. Density is calculated from

$$k_j(n + 1) = \frac{1}{2}[k_{j+1}(n) + k_{j-1}(n)] - \frac{\Delta t}{2\Delta x}[q_{j+1}(n) - q_{j-1}(n)] - \frac{\Delta t}{2}[g_{j+1}(n) + g_{j-1}(n)] \qquad (6)$$

where $g_j(n)$ is the dissipation term calculated from the merging flow in acceleration lane (Link F) and the flow in the freeway proper (Link B) according to

$$g_{j,F}(n) = \frac{q_{j-1,F}(n - 1)}{\Delta x} \frac{[k_{\text{jam}} - k_{j,B}(n - 1)]}{\sum_{i=j}^{J} [k_{\text{jam}} - k_{i,B}(n - 1)]} \qquad (7)$$

where $k_{\text{jam}} - k_{JB}(n - 1)$ represents the available storage space at Node j of Link B at the nth time step and k_{jam} is the maximum density. Because the dissipation and generation rates cannot be measured directly for each Δx, they are calculated by considering the remaining storage space on the freeway merging area as well as the flows in the ramp and freeway.

2. At the beginning node, $q_{OF}(n) = q_{JE}(n)$, $k_{OF}(n) = k_{JE}(n)$; whereas the flow rates at the ending Node J_F are determined by the traffic conditions of the adjacent freeway proper, that is, the Node J_B. If the Node J_B is not congested, then $k_{JF}(n + 1) = k_{JB}(n)$; otherwise $k_{JF}(n + 1) = k_{JF-1}(n)$.

3. With these dissipation rates and boundary conditions, density, speed, and flow at each internal node of Link F can be calculated by using Equation 6.

Link *B*

Link B represents the freeway merging area. Density is calculated from

$$k_j(n + 1) = \frac{1}{2}[k_{j+1}(n) + k_{j-1}(n)] - \frac{\Delta t}{2\Delta x}[q_{j+1}(n) - q_{j-1}(n)] + \frac{\Delta t}{2}[g_{j+1}(n) + g_{j-1}(n)] \qquad (8)$$

where $g_j(n)$ is given by Equation 7, but its sign is altered, i.e., $g_{j,F}(n) = -g_{j,B}(n)$. At the beginning and ending nodes, $q_{OB}(n) = q_{JA}(n)$ and $k_{OB}(n) = k_{JA}(n)$, $q_{JB+1}(n) = q_{OC}(n)$ and $k_{JB+1}(n) = k_{OC}(n)$, respectively.

Link *I*

This link represents the deceleration lane. All the exiting demand is assumed to diverge at the first segment of Link I, i.e., at Node O_I. The boundary condition of the beginning Node O_I is defined as follows:

$$q_{OI}(n + 1) = \min[q_{\text{exit}}(n), q_{\text{poss}}(n)] \qquad (9)$$

where

$q_{\text{exit}}(n)$ = exiting demand given by the input data, and
$q_{\text{poss}}(n)$ = possible exit flow rate, given by

$$q_{\text{poss}}(n) = [k_{\text{jam}} - k_{OI}(n)] * \Delta x * 3{,}600/(5{,}280 * \Delta t)$$

where Δx = 100 ft and Δt = 1 sec.

With these values of q_{OI} and the corresponding values of k_{OI}, calculated from either uncongested or congested region of the q-k curve depending on the downstream condition of Link I, Equations 2 to 4 can be applied to calculate the density, speed, and flow at each node of Link I.

Link C

When the capacity at the end of the exit ramp (J_J) is restricted, congestion on the off-ramp may spill back onto the freeway (Link C); in such case, the through capacity of Link C is reduced. The calculation of the density, speed, and flow values of Link C starts with the determination of any remaining exit demand on this link according to the following steps:

1. Determine the cumulative exiting demand at the Node J_C as follows:

$$R_{JC}(n + 1) = R_{JC}(n) + q_{\text{exit}}(n + 1) * \Delta t/3{,}600 - q_{OI}(n + 1) * \Delta t/3{,}600 \qquad (10)$$

where $R_{JC}(n)$ is the cumulative exiting demand remaining on the main line (vph).

If $R_{JC}(n) = 0$, then the Node J_C is not treated as an internal boundary, and Equations 2 to 4 are used for the calculation of k, u, and q at each internal node of Link C.

2. If $R_{JC}(n) > 0$, then the Node J_C is treated as an internal boundary and the through demand and through capacity are calculated as follows:

$$\text{TD}_{JC}(n) = k_{JC-1}(n) * u_{JC-1}(n) * \text{LN}_C \qquad (11)$$

$$\text{TC}_{JC}(n) = \text{CAP}_{JC}(n) * (\text{LN}_C - 1)/\text{LN}_C + q_{OI}(n) \qquad (12)$$

where

$\text{TD}_{JC}(n)$ = through demand,
$\text{TC}_{JC}(n)$ = through capacity,
LN_C = number of lanes in Link C, and
$\text{CAP}_{JC}(n)$ = capacity of Node J_C given by input data.

3. The boundary conditions of Node J_C are determined as follows: if $\text{TD}_{JC}(n) < \text{TC}_{JC}(n)$, then set

$$q_{JC}(n + 1) = \text{TD}_{JC}(n)$$

and calculate the corresponding $k_{JC}(n + 1)$ from the uncongested region of the q-k curve; otherwise, set

$$q_{JC}(n + 1) = \text{TC}_{JC}(n)$$

and calculate $k_{JC}(n + 1)$ from the congested region of the q-k curve. Finally, using Equations 2 to 4, k, u, and q at each internal node of the Link C can be calculated with the previous boundary conditions. At the beginning node,

$$q_{OC}(n) = q_{JB}(n)$$

and

$$k_{OC}(n) = k_{JB}(n)$$

Link G

The beginning node of Link G is treated as an internal boundary whose conditions are determined as follows:

$$q_{OG}(n + 1) = q_{JC-1}(n) - q_{OI}(n + 1) \qquad (13)$$

where $q_{OI}(n + 1)$ is given by Equation 9. If Node O_G is uncongested at the previous time step, i.e., $k_{OG}(n) < k_{\text{cr}}$, where k_{cr} is the density at the capacity flow, then $k_{OG}(n + 1)$ is calculated from the uncongested region of the q-k relationship using $q_{OG}(n + 1)$; otherwise, from the congested region. At the ending node J_G,

$$q_{JG}(n + 1) = q_{OH}(n + 1) \qquad (14)$$

$$k_{JG}(n + 1) = k_{OH}(n + 1) \qquad (15)$$

Finally, Equations 2 to 4 are used to calculate the k, u, and q values at each internal node of Link G.

Treatment of Lane Blockage Caused by Incidents

Figure 2 shows a discretized freeway section with part of its width closed because of an incident. When the reduced capacity of the incident area (Link B) is less than the through demand, congestion develops and propagates upstream. In the enhanced KRONOS program, Nodes J_A and J_B are treated as internal boundaries whose conditions are determined as follows:

1. If the through demand, $q_{JA-1}(n)$, is greater than CAP_B = the capacity of the incident area, then

For Link A, set

$$q_{JA}(n + 1) = \text{CAP}_B/\text{LN}_B \qquad (16)$$

and calculate the corresponding $k_{JA}(n + 1)$ from the congested region of the q-k curve.

For Link B,

$$q_{OB}(n + 1) = \text{CAP}_B/\text{LN}_B \qquad (17)$$

set

$$k_{OB}(n + 1) = k_{\text{cr}} \qquad (18)$$

If $q_{JA-1}(n) < \text{CAP}_B$, then $q_{JA}(n + 1) = q_{OB}(n + 1) =$

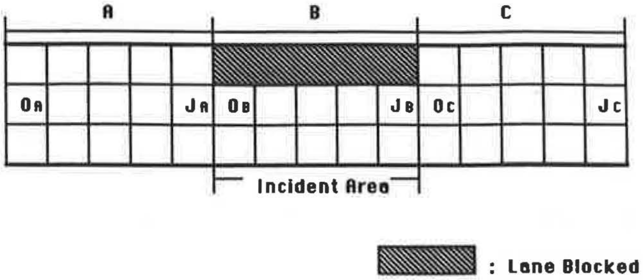


FIGURE 2 Space discretization of an incident area.

$q_{JA-1}(n)/\mathrm{LN}_B$, and the corresponding $k_{JA}(n + 1)$ and k_{OB} $(n + 1)$ can be calculated from the uncongested region of the q-k curve.

2. At the Node $J_B(O_C)$, the boundary flows are set as follows:

$$q_{JB}(n + 1) = q_{OC+1}(n)/\mathrm{LN}_B \qquad (19)$$

$$q_{OC}(n + 1) = q_{OC-1}(n)/\mathrm{LN}_C \qquad (20)$$

The corresponding $K_{JB}(n + 1)$ and $k_{OC}(n + 1)$ are calculated from either the uncongested or congested region of the q-k relationship, depending on the traffic condition of the downstream Link C.

With these boundary conditions, and by using Equations 2 to 4, the density, speed, and flow at each internal node of Links A, B, and C in Figure 2 can be calculated. The Nodes O_A and J_C can be considered either as external boundaries whose conditions are specified from the input data or as internal nodes when there exist other segments combined with the incident area.

The modeling presented allows simultaneous treatment of the freeway and its ramps in an integrated fashion, where the dynamic interaction between ramps and freeway is explicitly considered through the dissipation-generation terms in the state equations. The numerical scheme adopted in KRONOS provides a stable difference approximation of first-order accuracy with respect to Δt (*13*). Further, the internal boundary treatment developed in this study significantly improves flow conservation as well as realistic representation of queue propagation and dissipation through time. In this formulation, there are no restrictions with respect to arrival and departure patterns or the equilibrium u-k relationship. Thus, the former could assume any pattern including stochastic ones, and the latter could even be discontinuous.

From q, k, and u, other measures of effectiveness such as total travel, total travel time, delay, energy consumption, and pollution level can be derived. These details, along with the step-by-step algorithms to treat special geometric types such as merging and diverging of two freeways and weaving sections have been provided by Georgadakos et al. (*10*) and Kwon et al. (*14*).

PROGRAM SUMMARY

The current KRONOS program (Version VI) uses the previously summarized modeling for dynamically calculating speed, flow, and density on each node and segment Δx (100 ft). The simulated results of these basic flow parameters are stored in a spreadsheet-formatted output file and the measures of effectiveness (MOE) such as total travel, total travel time, delay, and the duration and extension of congestion (including queue length and size) are derived. These MOE values are summarized by zone (defined as a section between ramps or a merging, diverging and weaving area), as well as by ramp, and presented on the monitor screen. Two- and three-dimensional plots of k, u, and q are also produced for showing the evolution of these basic variables in space and time and, therefore, visualizing the propagation and dissipation of shock waves and congestion on both the freeway and its ramps. The spreadsheet file containing the detailed simulation results can be accessed by other data management and analysis software such as Lotus 1–2–3, and, thus, much more flexible analysis of the simulation results is possible. For a faster and easier presentation of the propagation of disturbances along the freeway as well as a quick review of its operation (after changes), the discretized form of the freeway is presented on the screen and the density variation of each segment is plotted continuously through time with different colors (from black to red) according to its level.

This output module operation, as well as the input data preparation, is performed interactively in a single-screen environment with on-line help, which minimizes the reference to the manual. For instance, the geometrics are entered by selecting the configuration of each segment from the available alternatives presented on the screen. Figures 3 and 4 show a total of 24 segment types available in the current version of KRONOS as well as an example data input screen. Following definition of the geometrics of each segment, the entire freeway can be plotted for verification. The other input requirements include the freeway and ramp demand patterns in 1- to 15-min time increments (called "time slices"), traffic composition, and departure patterns at exit points, that is, off-ramps and downstream ends of the freeway. Arrival and departure patterns are also plotted for verification and can be as complex as desired. The program allows use of user-specified speed-density models that are entered interactively. The changes to input already entered can be made at any stage during the data entry.

Currently the program can simulate a one-directional freeway section up to six lanes wide and approximately 10 mi long. The section can contain up to 20 entrance and 20 exit ramps, and auxiliary lanes between ramps are permitted. Following is a list of the program's output features:

1. A summary of the MOE indicators for each time slice, including
   a. Total travel, total travel time, and delay;
   b. Total arrivals and departures;
   c. Queue size and length on each ramp; and
   d. Energy consumption and pollution level;
2. Two- and three-dimensional plots of speed, flow, and density as functions of distance or time, or both;
3. Color graphics display of the density variation in the discretized freeway through time;
4. Spreadsheet-type files for flow, speed, and density for further analysis; and
5. Hard copies of two-dimensional plots, including freeway geometrics.

In its present form, the program requires an IBM Personal Computer (PC) (PC/XT, PC/AT, or PS/2) or fully compatible computer with a minimum of 640-kB RAM memory, keyboard, and one 1.2- or 1.44-MB floppy disk drive or a hard disk. It also requires an enhanced color graphics adaptor, a color monitor, and an 80-column (minimum) dot matrix printer with graphics capabilities. The only software required is PC–DOS or MS–DOS operating system, Version 3.0 or later. Simulation time can be substantially reduced if an IBM PC 8087/80287/80387 math coprocessor is added to the system. With an 80386 computer equipped with a 80387 math copro-

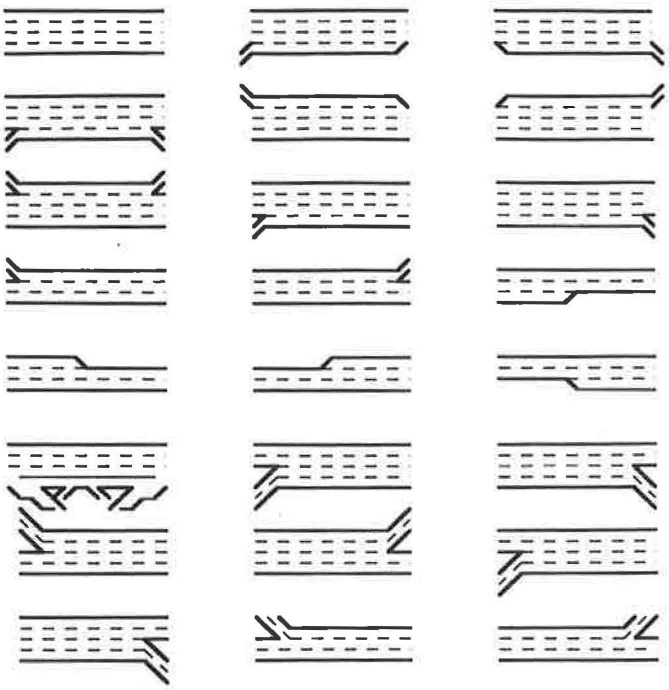
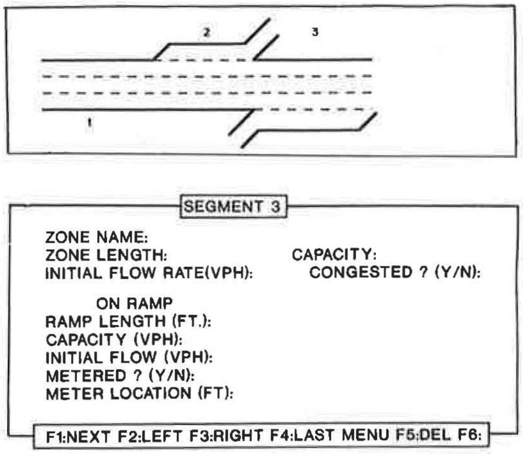


FIGURE 3 Freeway segment types for synthesizing the geometry of the section to be simulated.



```
SEGMENT 3
ZONE NAME:
ZONE LENGTH:                      CAPACITY:
INITIAL FLOW RATE(VPH):             CONGESTED ? (Y/N):

        ON RAMP
RAMP LENGTH (FT.):
CAPACITY (VPH):
INITIAL FLOW (VPH):
METERED ? (Y/N):
METER LOCATION (FT):

F1:NEXT F2:LEFT F3:RIGHT F4:LAST MENU F5:DEL F6:
```

FIGURE 4 Sample input screen.

cessor, the program requires approximately 12 min of execution time to simulate a 6-mi freeway section with three lanes for a 1-hr time period.

PRELIMINARY TESTING

Testing of the modeling was done first by implementing the models to hypothetical situations on an individual basis, that is, by including only one of the merging, diverging, weaving, or other sections at a time rather than a combination of these components. All the 24 potential freeway components (shown in Figure 3) incorporated in the KRONOS-VI program were tested. To illustrate the nature of this testing, which is both intuitive and quantitative, the test results of the spillback mechanism in the diverging area where the off-ramp is closed for 5 min are presented in this section. The detailed test results of other components were provided by Kwon et al. (*14*). Figure 5 shows the geometrics and demand-capacity variation of a hypothetical diverging area, where the upstream demand of the main line and the exiting demand are kept constant at 4,800 and 900 vph, respectively, for 20 min, while the capacity of the downstream off-ramp is reduced from 1,500 to 0 vph for the second 5-min time interval. The simulation results of this hypothetical case are indicated in Figure 6, which shows the density trajectory of the freeway main line at every 3-min time interval. As illustrated in these figures, the model realistically represents the queue build-up process after the ramp is closed as well as the dissipation when the capacity is restored.

Following testing of the model components, several longer freeway sections were tested using the real freeways in Minneapolis, Minnesota. For these sections, only volume data were available from the existing loop detectors. Figure 7 shows the test sections considered. The detailed test results of this preliminary testing were provided by Kwon et al. (*14*); briefly, the volume estimation error, defined as

$$\frac{\sum \left| \begin{array}{c}\text{(observed volume} - \\ \text{estimated volume)/observed volume}\end{array} \right|}{\text{number of observations}} * 100 \text{ (percent)}$$

ranged from 3 to 8 percent for all short segment types shown in Figure 7. Because of the lack of speed data, a simple *u-k* relationship estimated from the occupancy information was used in this testing.

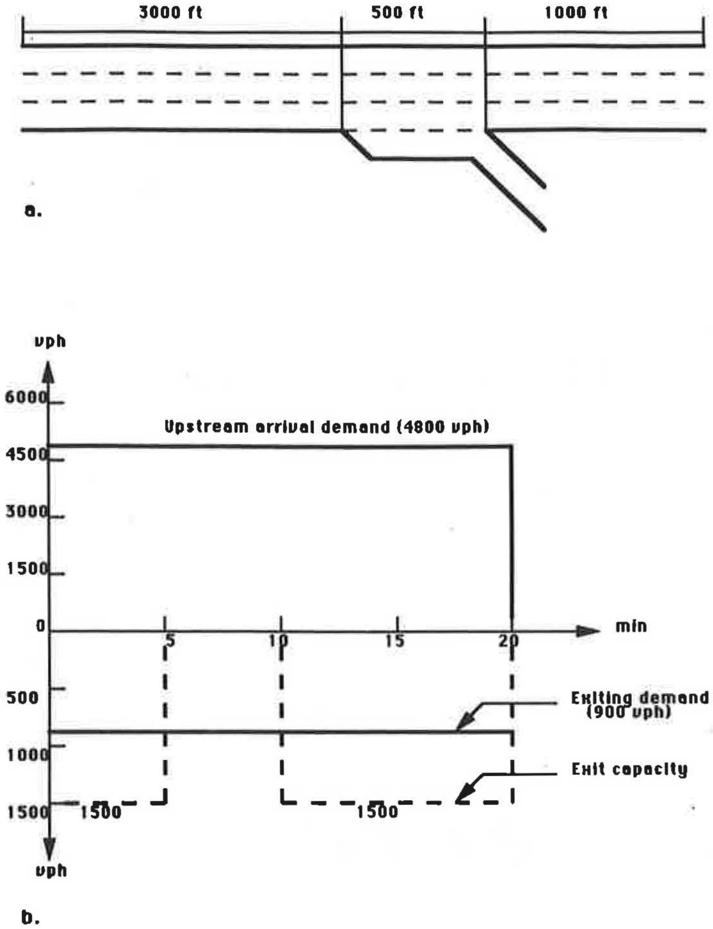


FIGURE 5 Geometrics and demand pattern of an example diverging area: (a) geometrics, and (b) demand position.

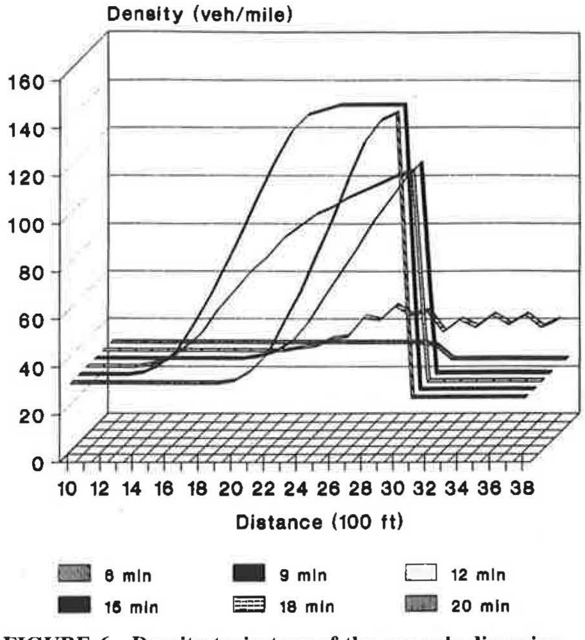


FIGURE 6 Density trajectory of the example diverging area.

FIELD VALIDATION

Although the preliminary test results indicate satisfactory model performance with reasonable accuracy for individual segments, robustness and reliability were not ensured because of the lack of speed data and volume information from all segments. To address this problem, field data collection was recently performed for future validation through a research project funded by the Minnesota Department of Transportation. An 8-mi section of the I-35W freeway was selected as the test site. This section, connecting the Minneapolis downtown to the southern suburban areas, contained 21 entrance-exit ramps as well as a variety of geometric types, such as merging, diverging, weaving, lane addition or deletion at ramps, and so forth. Figures 8 and 9 show schematically both directions of the test freeway section and the data collection sites, which coincided with the locations of in-place loop detectors. Although the loop detectors provided volume and occupancy information, accurate speed information was not directly available from the current Minneapolis detection system. In this study, the speed data were collected manually using stop watches, whereas the volume data were obtained from loop detectors, with the cooperation of the Traffic Management

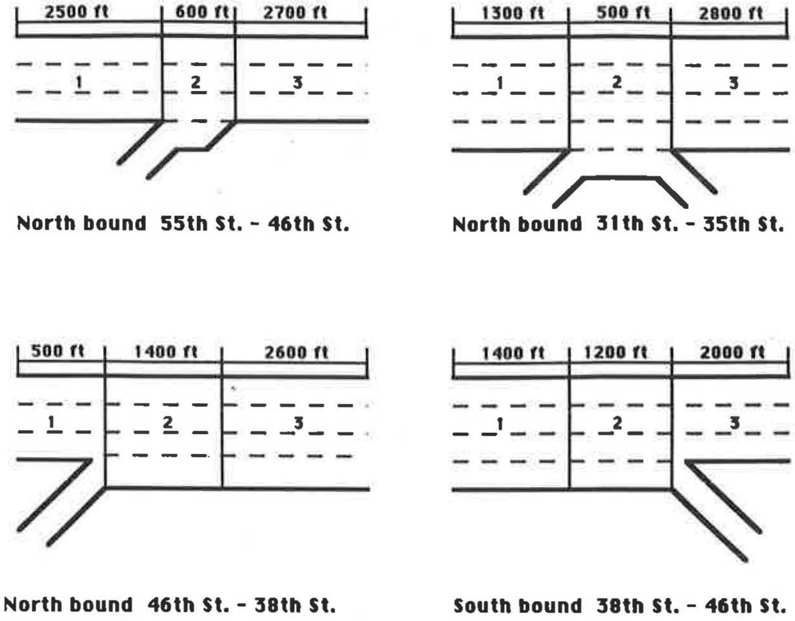

2500 ft
600 ft
2700 ft
1
2
3
North bound 55th St. - 46th St.
1300 ft
500 ft
2800 ft
1
2
3
North bound 31th St. - 35th St.
500 ft
1400 ft
2600 ft
1
2
3
North bound 46th St. - 38th St.
1400 ft
1200 ft
2000 ft
1
2
3
South bound 38th St. - 46th St.

FIGURE 7 Layout of the test sites at I-35W used in the preliminary testing.

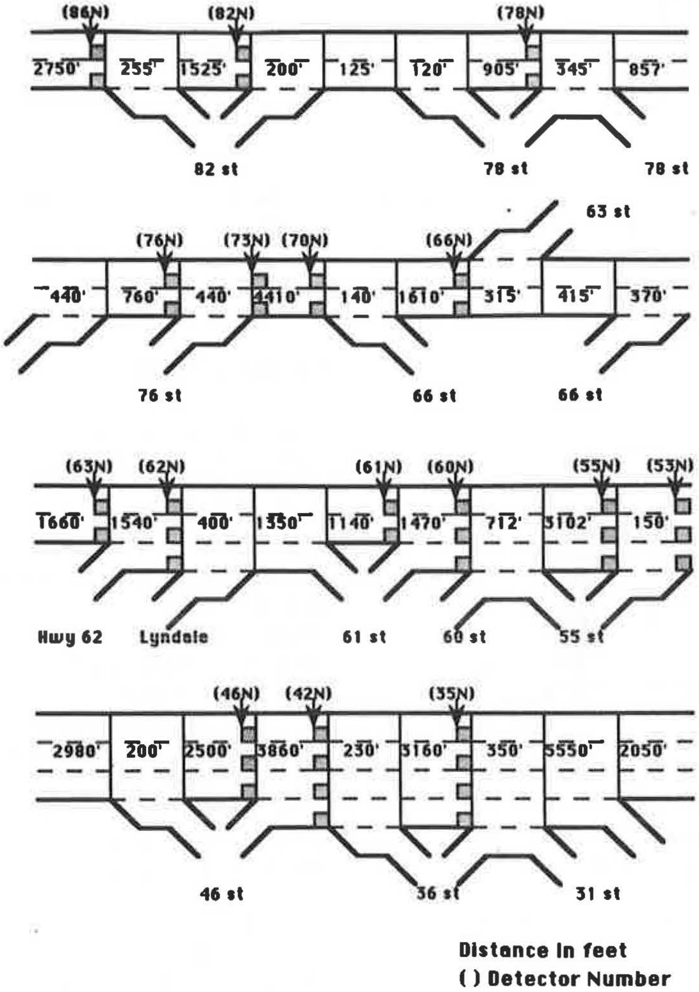


FIGURE 8 Geometrics of the test section (I-35W, Northbound).

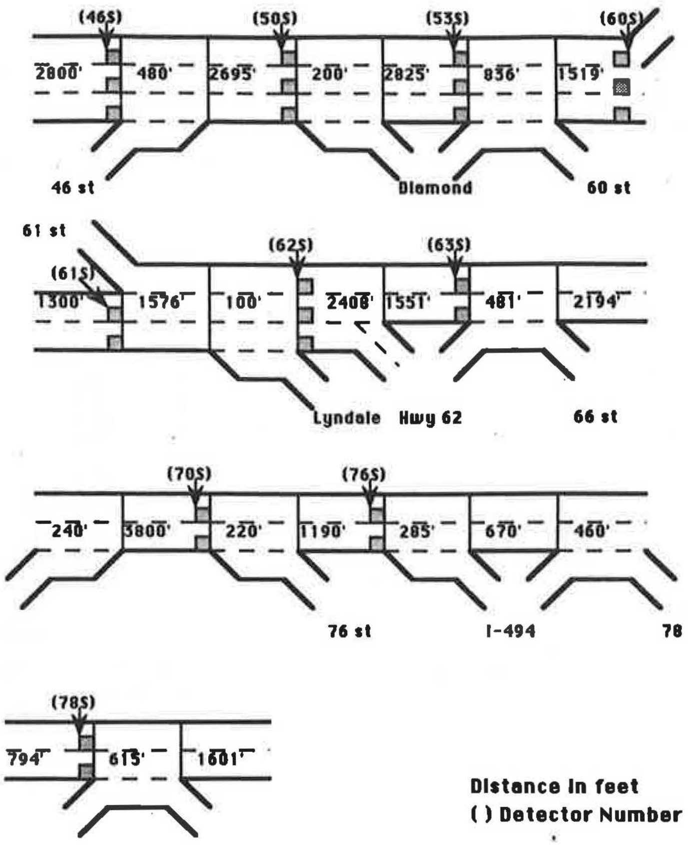


FIGURE 9 Geometrics of the test section (I-35W, Southbound).

Center of the Minnesota Department of Transportation. For those sites where the loop detectors were not functioning properly, video recordings or tube detectors were used. Approximately 70 persons were involved to cover 35 collection sites for 6 days during the morning and afternoon peak periods in November 1989.

The data collected from each site include

1. 5-min travel time measurements of vehicles passing between two observation posts for each lane,
2. 5-min volume-occupancy passing the observation posts on the main line (from loop detectors), and
3. 5-min freeway merging-exiting volume from the entrance ramps and to the exit ramps (from either loop or tube detectors).

Although the morning periods do not include any severe congested situation, snow and heavy congestion occurred in all three afternoons when the data were collected. Approximately 6,500 data sheets were gathered from the 6-day measurements. Each data sheet contained the sample travel time measurements for each lane for every 5-min interval. The information on each data sheet has been stored in computer files and processed to eliminate measurement errors. Because of detector malfunctions, only 4-day measurements were usable for the testing. The remainder of this section summarizes the test results on the basis of the available data.

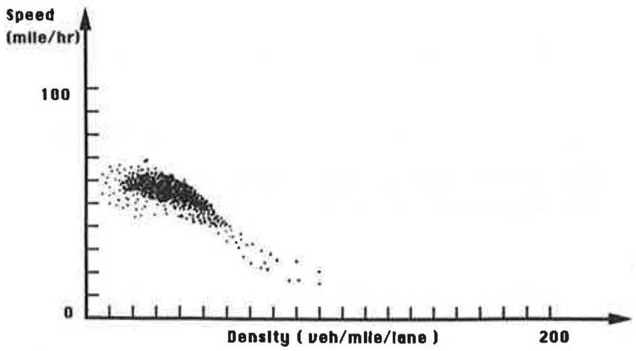


FIGURE 10 Speed-density relationship derived from the morning measurements.

First, a speed-density relationship was derived from the collected data. Figure 10 shows the speed-density plots derived from measurements for two peak periods (i.e., November 7 and 9 morning data). From these plots, a representative speed-density (u-k) relationship was derived by fitting a curve to the data points. This curve was entered numerically to KRONOS. Further, the capacity of each individual segment of the sample freeway section was estimated by using the *Highway Capacity Manual* (*15*) and field observations. Using the real volume data collected from the upstream boundary and entrance-exit ramps as the input demand patterns, the 8-mi sample freeway section was simulated with the current geometric conditions and the new u-k curve. The resulting comparisons of volume and speed generally indicated close agreement between simulated and actual data. Results at one typical location where the actual data were collected are shown in Figures 11 and 12, which indicate that speed estimation error is higher than the error in volume estimation.

In order to evaluate the model performance quantitatively, the following error measurements are calculated:

- Percentage difference (PD: %):

$100 * (\text{measured} - \text{estimated})_k/\text{measured}_k$

- Mean percentage difference (MPD: %):

$\Sigma(100 * (\text{measured} - \text{estimated})_k/\text{measured}_k)/N$

- Mean absolute error (MAE):

$\Sigma(\text{measured} - \text{estimated})_k/N$

where N is the number of measured points.

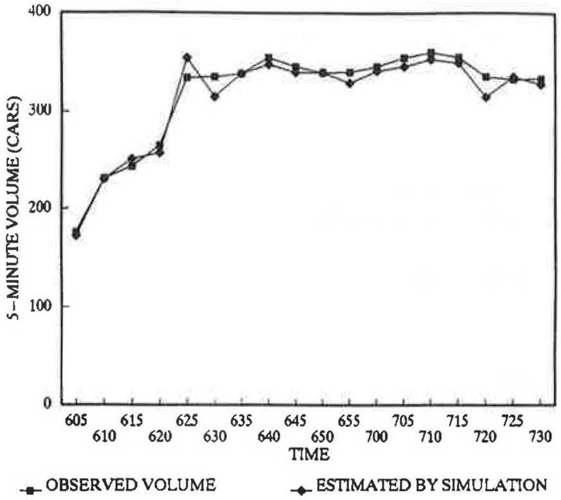


FIGURE 11 Volume comparison results (Nov. 9, morning period, location: 82N).

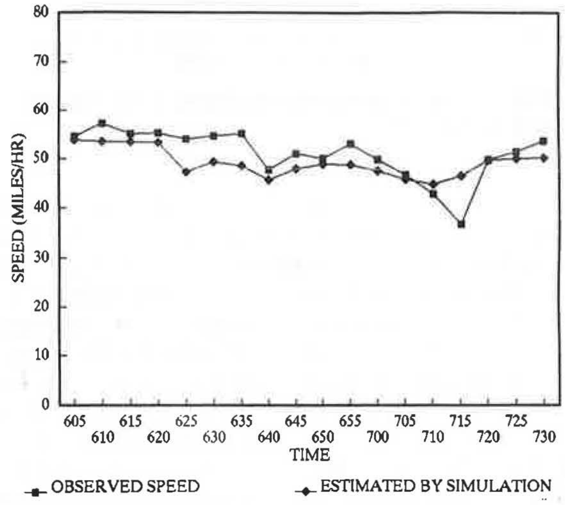


FIGURE 12 Speed comparison results (Nov. 9, morning period, location: 82N).

Figures 13–16 show the error (percentage difference) distribution of the 5-min volume-speed estimation results for one morning and one afternoon period, and Tables 1 and 2 present the mean error (MPD and MAE) of each test site for 4 days. As indicated in these figures and tables, the model exhibited satisfactory performance with the morning data set where

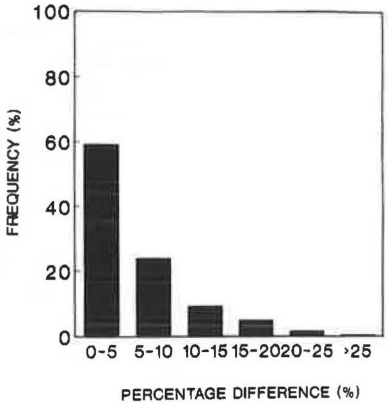


FIGURE 13 Volume error (PD) distribution during the morning period on Nov. 7.

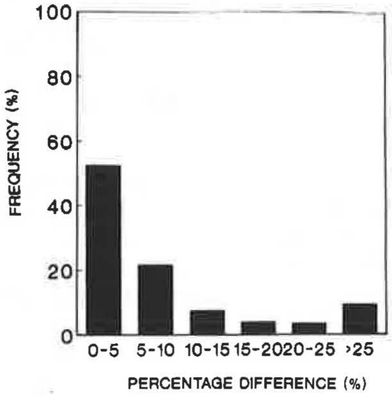


FIGURE 14 Speed error (PD) distribution during the morning period on Nov. 7.

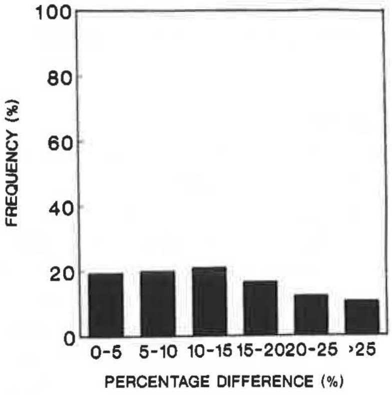


FIGURE 15 Volume error (PD) distribution during the afternoon period on Nov. 20.

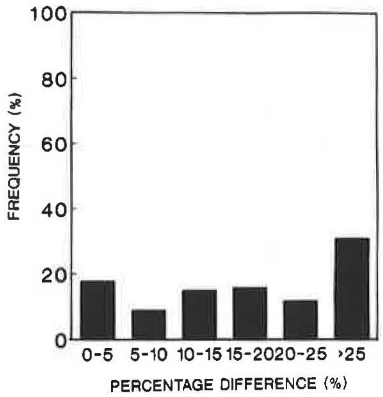


FIGURE 16 Speed error (PD) distribution during the afternoon period on Nov. 20.

traffic was relatively uncongested, whereas the estimation error increased in two afternoon peak periods in which traffic was heavily congested because of inclement weather conditions. For example, the comparisons for the November 7, 1989, morning data suggest that, out of 576 data points (16 stations * 36 5-min time intervals) estimated by KRONOS, 92 percent of volume and 82 percent of speed estimation results were within 15 percent of the actual measurements. However, during the afternoon of the November 20 peak period, only 62 percent of volumes and 51 percent of speeds were within 15 percent of the actual measured data. Further, the MPD of volume for two mornings ranges from 0.6 to 15.9 percent, whereas the speed MPD for the same two mornings was between 2.5 and 28.0 percent. For two afternoons, the volume estimation MPD ranges from 7.1 to 27.6 percent and the speed MPD between 2.8 and 31.7 percent.

The higher error in speed estimation indicates the need to use multiple speed-density relationships reflecting weather conditions rather than a single curve throughout. Further, the speed estimation error indicates the need for more effort in calibrating speed-density curves and estimating the capacities. The latter, which is an important issue common to all macroscopic simulation models, is currently being investigated.

Overall, the test results using four peak periods indicate satisfactory performance of the enhanced KRONOS program in light-to-moderate traffic conditions, whereas they indicate further investigation is needed to explain the higher differences during heavy congestion. One possible explanation is the lack of sufficient stable data for deriving the speed-density relationship during such conditions. Clearly, either a new methodology for this procedure or more advanced state equations that do not require use of such relationships have to be developed. Development of such models is currently being considered along with use of high-order continuum models, which, by their nature are truly dynamic as they use a momentum equation.

CONCLUSIONS AND FUTURE IMPROVEMENTS

The enhanced KRONOS program provides a practical tool for traffic engineers in evaluating various freeway design and

TABLE 1 MEAN ERROR OF THE VOLUME-SPEED ESTIMATION RESULTS (MORNING, 6:00 TO 9:00 a.m.)

Site	Nov. 7				Nov. 9			
	Volume/5-min.		Speed/5-min.		Volume/5-min.		Speed/5-min.	
	MPD	MAE*	MPD	MAE**	MPD	MAE	MPD	MAE
86N	0.7	13.6	4.6	2.3	0.6	12.3	6.9	3.8
82N	2.5	42.6	3.3	1.9	2.3	41.9	3.8	1.9
78N	6.6	96.6	7.8	3.7	7.8	116.8	5.6	3.2
76N	4.0	57.3	3.7	2.5	4.4	64.6	4.8	2.5
73N	3.5	54.3	N/A	N/A	5.0	79.1	6.1	3.4
70N	3.5	53.7	3.3	1.9	4.5	72.8	4.2	2.4
66N	4.7	65.2	2.5	1.4	5.9	82.9	4.5	2.3
63N	14.4	196.0	5.9	2.4	15.9	221.5	16.4	5.1
62N	8.4	133.7	N/A	N/A	9.6	155.1	N/A	N/A
61N	6.9	139.8	24.2	10.5	5.9	125.3	28.0	13.3
60N	7.1	128.0	21.8	10.5	6.9	129.7	17.3	7.3
55N	4.6	88.2	15.4	8.2	3.9	77.4	16.5	6.8
53N	7.1	134.2	20.5	9.3	6.5	127.1	26.1	9.8
46N	4.9	95.7	7.9	4.3	4.1	82.1	6.1	2.4
42N	4.8	73.8	4.4	2.6	3.4	65.0	4.0	2.4
35N	4.2	73.3	3.8	2.1	3.6	67.8	4.9	2.7

*: veh/hr/lane ** : miles/hr

TABLE 2 MEAN ERROR OF THE VOLUME-SPEED ESTIMATION RESULTS (AFTERNOON, 3:30 TO 6:30 p.m.)

Site	Nov. 14				Nov. 20			
	Volume/5-min.		Speed/5-min.		Volume/5-min.		Speed/5-min.	
	MPD	MAE*	MPD	MAE**	MPD	MAE	MPD	MAE
46S	5.5	83.0	27.0	8.6	13.1	223.9	26.2	11.5
50S	7.1	114.5	27.1	9.8	14.4	255.7	31.6	11.1
53S	10.2	156.1	N/A	N/A	11.5	198.1	30.3	11.5
60S	9.2	137.3	31.4	7.1	20.8	359.2	30.8	6.0
61S	13.9	211.8	21.8	6.6	15.7	276.7	31.7	8.3
62S	10.4	145.2	20.7	8.7	8.2	131.8	2.8	1.3
63S	27.6	273.6	30.2	12.4	20.0	238.5	15.6	7.1
70S	16.1	217.8	13.6	7.3	8.4	141.3	20.0	11.2
76S	17.1	215.4	14.6	7.5	7.9	123.3	16.4	9.4
78S	24.1	246.2	N/A	N/A	17.1	224.7	7.1	3.2

*: veh/hr/lane **: miles/hr

management alternatives, including decisions frequently made in freeway operations, such as ramp metering, lane blockages, construction and maintenance operations, and others. The program, taking advantage of the recent advancements in the microcomputer hardware technology, has the user-friendly menu-driven I/O modules that significantly reduce the time-consuming effort to prepare input and analyze the results. Furthermore, the improvements presented in the earlier sections reduced execution time substantially. In addition, by integrating traffic simulation with advanced data management and analysis software, more flexible and detailed analysis of the simulation results is possible. For example, using the spreadsheet-formatted output file, contour maps of speed, flow, and density that greatly assist the user to follow the generation and propagation of congestion as well as its dissipation can be easily produced. These contour maps are useful for visual (qualitative) assessment of the effects of improvements or incidents, on the freeway. The test results with the field data collected from an 8-mi section of the I-35W freeway indicate promising model performance in estimating volume and speed through time.

Although the data collected in this study provide sufficient information to evaluate the overall performance of the model, more detailed measurements of speed, density, and volume at shorter time and space increments are required to develop accurate models that can be applicable in a real-time environment. Future plans include the collection of such detailed data using the video detection system currently being developed at the University of Minnesota (*16*). A total of 38 video detectors are planned to be installed by 1993 along a 2.5-mi section of the I-394 freeway in Minneapolis, Minnesota (*16*). This section will serve as a live laboratory for collecting and studying traffic characteristics as well as for testing and validating the KRONOS program in a more rigorous fashion by fully analyzing the traffic dynamics in merging, diverging, and weaving areas. Further enhancements to KRONOS currently under consideration also include optimal calibration of multiple speed-density relationships for each section of freeway, high-occupancy-vehicle (HOV) priority treatment, corridor modeling with demand-diversion, estimation of optimal ramp metering rates, and development of high-order continuum models that do not require equilibrium speed-density relationships, and so forth. Finally, an input interface that can automatically extract the necessary input information (geometric and demand) from available data bases will be developed, so that time-consuming manual efforts in preparing the input data can be minimized.

ACKNOWLEDGMENTS

Funding for this project was provided by the Minnesota Department of Transportation. Jim Aswegan greatly assisted the research team in his role as a project monitor.

REFERENCES

1. A. D. May. Freeway Simulation Models Revisited. In *Transportation Research Record 1132*, TRB, National Research Council, Washington, D.C., 1988, pp. 94–99.
2. T. Imada and A. D. May. *FREQ8PE: A Freeway Corridor Simulation and Ramp Metering Optimizing Model*. Report UCB-ITS-RR-85-10. University of California, Berkeley, 1985.
3. H. J. Payne. FREFLO: A Macroscopic Simulation Model of Freeway Traffic. In *Transportation Research Record 772*, TRB, National Research Council, Washington, D.C., 1979, pp. 68–75.
4. D. A. Wicks and E. B. Lieberman. *Development and Testing of INTRAS, a Microscopic Freeway Simulation Program*. FHWA-RD-76-75. FHWA, U.S. Department of Transportation, 1980.
5. M. Van Aerde, J. Voss, A. Ugge, and E. R. Case. Managing Traffic Congestion in Combined Freeway and Traffic Signal Networks. *ITE Journal*, Vol. 59, No. 2, Feb. 1989, pp. 36–42.
6. A. K. Rathi and Z. A. Nemeth. FREESIM: A Microscopic Simulation Model of Freeway Lane Closures. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986, pp. 21–24.
7. P. G. Michalopoulos. A Dynamic Freeway Simulation Program for Personal Computers. In *Transportation Research Record 971*,

TRB, National Research Council, Washington, D.C., 1984, pp. 68–79.

8. P. G. Michalopoulos, D. E. Beskos, and J. K. Lin. Analysis of Interrupted Traffic Flow by Finite Difference Methods. *Transportation Research*, Vol. 18B, 1984, pp. 409–421.
9. P. G. Michalopoulos. Integrated Modelling of Freeway Flow and Application to Microcomputers. *Traffic Engineering Control*, Vol. 27, No. 4, 1986, pp. 198–204.
10. G. Georgadakos, R. Plum, and P. G. Michalopoulos. KRONOS-V: An Interactive Freeway Simulation Program for Personal Computers. Minnesota Department of Transportation, St. Paul, 1988.
11. J. Lin. Numerical Analysis of Traffic Flow in Complex Transportation Networks. Ph.D. dissertation, University of Minnesota, St. Paul, 1985.
12. F. L. Hall and M. A. Gunter. Further Analysis of the Flow-Concentration Relationship. In *Transportation Research Record 1089*. TRB, National Research Council, Washington, D.C., 1988, pp. 1–9.
13. P. D. Lax. Weak Solution of Non-Linear Hyperbolic Equations and Their Numerical Computations. *Communications of Pure and Applied Mathematics*, Vol. 7, 1954, pp. 159–173.
14. E. Kwon, C. F. Lee, J. Kang, et al. Traffic Management Strategies for Freeway Corridors Phase II Extension. Minnesota Department of Transportation, St. Paul, 1990.
15. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
16. P. G. Michalopoulos, B. Wolf, and R. Benke. Testing and Field Implementation of the Minnesota Video Detection System. In *Transportation Research Record 1287*, 1990, pp. 176–184.

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

Assessment of the Traffic Experimental and Analysis Simulation Computer Model Using Field Data

A. Essam Radwan, Farid Naguib, and Jonathan Upchurch

Computer simulation techniques and their applications to traffic operations have gained popularity in recent years. The traffic experimental and analysis simulation (TEXAS) computer model is a prime example. This model was developed to be used as a tool to evaluate traffic performance at isolated intersections operating under various types of intersection control. The objective of this study was to assess the ability of the TEXAS model in simulating isolated signalized intersections. This assessment was conducted using field data. Eight intersection conditions were used for data collection. Existing field data were obtained from previous studies and supplemented with more data collection where needed. Critical 10-min time periods were chosen for the simulation runs, and the average stopped delays obtained from simulation were compared with the observed stopped delay data. Statistical tests were conducted; the results indicated that in most of the intersection conditions no significant difference was observed for through traffic. However, significant differences between field and simulated results were achieved for left-turning traffic. These findings were limited to eight intersection conditions studied in Arizona, and it suggests that the user of the model is advised to carefully assess the default variables embedded in TEXAS.

Traffic simulation models are computer programs that are designed to represent realistically the behavior of the physical system. The major advantage of using a simulation model is its ability to compare and evaluate alternate solutions of a single problem by changing the variables in the model without physically going out to the field to make the change. This analysis tool is effective in comparing different scenarios before deciding on one to implement, thus reducing costs that would otherwise be incurred if an unsuccessful scenario is implemented. Another advantage is that a simulation model can give results for variables that would otherwise be difficult to obtain from field measurements.

Painting a rosy picture about computer simulation could be deceiving. The user of computer simulation models has to be careful about how the model is being used and how much faith the user has in the results. A well-validated simulation model can be a powerful tool. Validation involves large-scale field data collection for different values of selected variables followed by using the model to create the same circumstances in the computer. The produced results are then compared against the field data to decide whether the model realistically replicates the real world. If the comparison indicated that the field data are significantly different from the simulated results, refinement of the model variables is then pursued.

This study focused on traffic operations at isolated signalized intersections. Intersections are usually the critical component of a traffic network in the urban area. One of the traffic simulation models developed for isolated intersections is the traffic experimental and analysis simulation (TEXAS) computer model.

The main goal of this research was to compare delay data collected in previous field studies with the delay predicted by the TEXAS computer model. The main purpose of this exercise was to attempt to assess the simulation model in terms of its ability to replicate the real world without changing the default values. The field data used for the assessment purpose came from three related studies. Most of the data for this effort were provided by Matthias and Upchurch (*1*). The remainder of the data were taken from two theses (*2,3*). Although some of the data are close to 5 years old and were collected for the purpose of developing left-turn signal warrants, the data collection method provided comprehensive information that could be used with high confidence to assess a traffic simulation model.

THE TEXAS SIMULATION MODEL

The TEXAS model is a microscopic simulation model. It was developed by the University of Texas at Austin for the Texas State Department of Highways and Public Transportation (*4–6*). In a microscopic model, each vehicle in the traffic stream is monitored individually. The model can be used as a tool by transportation engineers to evaluate traffic performance at isolated intersections operating under various types of intersection control.

The model is divided into three main parts, a geometry processor called GEOPRO, a driver-vehicle processor called DVPRO, and a traffic simulation processor called SIMPRO.

The geometry processor accepts data describing the physical configuration of the intersection such as the number of legs and the number of lanes and their widths. The processor calculates the geometric path of vehicles on the approaches and within the intersection, identifies points of conflict between intersection paths, and determines the minimum available sight distance along each inbound approach.

The input requirements for this processor include approach information such as the number of inbound and outbound approaches, the speed limit for each approach, the number

A. E. Radwan, Department of Civil and Environmental Engineering, University of Central Florida, Orlando, Fla. 32816-0450. F. Naguib, Wilbur Smith Associates, 6053 W. Century Blvd., Suite 780, Los Angeles, Calif. 90045. J. Upchurch, Center for Advanced Research in Transportation, Arizona State University, Tempe, Ariz. 85287-6306.

of lanes for each approach, and the maximum angular deviation of through movement and U-turn movement for each approach. Lane information, such as the width of each lane, the geometry of each lane, and the turning movements generated from each inbound lane and accepted by each outbound lane, is also required.

The driver-vehicle processor stores information related to the driver characteristics and the vehicle characteristics in the traffic stream. The processor generates driver-vehicle units for use by the traffic simulation processor. Each one of these units is randomly assigned a driver class, a vehicle class, a lane, a turning movement, and a desired speed using a discrete empirical distribution. The total number of driver-vehicle units generated depends on the vehicular volume and on the simulation time. The processor then orders these units sequentially by queue-in time.

The input data required for this processor, for each intersection approach, includes the headway distribution and its parameter, the minimum headway, the hourly traffic volume, the mean and 85th percentile speeds, the turning distribution (percent of vehicles going to each outbound approach), the lane occupancy, the time for generating traffic, and the number of driver and vehicle classes (*2*). The model has seven different types of headway distributions to choose from. They are the constant, the Erlang, the gamma, the log-normal, the negative exponential, the shifted negative exponential, and the uniform distributions. The time for generating traffic is made up of the start-up time plus the simulation time. In the start-up time period, if the flow through the system has not attained a steady state condition (*3*) then the performance statistics are unreliable. After the specified start-up time, flow is assumed to have reached steady state and it is followed by the simulation time period. Another input requirement for this processor is the percent of left-turning vehicles to enter in the median lane and the percent of right-turning vehicles to enter in the curb lane.

The traffic simulation processor uses the output from the previous two processors and processes each driver-vehicle unit through the intersection system. This processor simulates the traffic behavior of each driver-vehicle unit depending on its surrounding conditions. The driver-vehicle unit is monitored second-by-second from the time it enters an inbound approach until it leaves the system from an outbound approach. The processor adjusts the movement of the driver-vehicle unit depending on various elements, such as the indication of the traffic control device, the presence of a vehicle ahead, and whether the driver-vehicle unit is in a car-following situation.

The input data requirements for a simulation run are the type of intersection control, the start-up and simulation time, the time step increment for simulation, the maximum clear distance for being in a queue, the speed for delay below XX mph, the time for lead and lag safety zone for conflict checking, and the parameter values for the car-following equation. The available intersection control options to choose from are traffic signals with pretimed, semiactuated, or fully actuated controllers, all-way stop sign, two-way stop sign, yield sign, and uncontrolled.

Once SIMPRO has been successfully executed, the model can be prepared to display the simulation in graphics format. Two programs are used for this activity: DISPRE, which is a display preparation program, and DISPRO, which is the display processor in which the animated graphics are shown.

After running DISPRE, the program DISPRO is executed. This displays the intersection layout on the screen and then simulates the position of the vehicles from the time they enter the system to the time they leave it. This information can be displayed in real time or in a stop-and-go mode that is manually run. The graphic display is useful in detecting errors in the intersection geometry that may occur during the input of the offset of each leg centerline from the intersection center. It is also used to detect errors in the input of the phasing sequence for signalized intersections.

OBJECTIVE

The principal objective of this research study was to assess the ability of the TEXAS simulation model to simulate traffic movements at isolated signalized intersections using existing field data. This effort is needed to determine whether the model behaves in a way similar to the real world when run under the same conditions as the field data. The assessment process was attempted by comparing the results of the measures of effectiveness obtained from the simulation runs with those obtained from field data. Statistical methods were used to test the significance between the simulation and field results.

DESCRIPTIONS OF THE INTERSECTIONS

Eight intersection conditions, each having available field data, were used in this study. Six intersections located in the greater Phoenix area were used for the previous left-turn warrants study (*1*). The intersections were

- University and Alma School,
- Alma School and Broadway,
- Priest and Broadway,
- Thomas and 44th Street,
- Scottsdale and Thomas, and
- Dobson and Main.

These were chosen because they represented a wide variety of intersection characteristics. The intersections cover a range of values for the geometry such as the number of opposing lanes, the type of left-turn phasing, left-turn volume, and volume of opposing traffic. Table 1 presents selected data items for the six intersections. The near-side approach means the intersection approach at which the time lapse camera was placed.

Subsequent to the left-turn warrants study (*1*), the intersection of Thomas and 44th Street was operated under two other different scenarios. The original scenario had a pretimed signal with permissive left-turn phasing. The second scenario involved changing the left-turn phasing from permissive to exclusive-permissive (*2*). The signal was also changed to a fully actuated signal but with a variable cycle length that disrupted the vehicular progression. The third scenario had a phasing similar to the previous scenario but with a fixed cycle length that helped vehicular progression (*3*).

Therefore, a total of eight different intersection conditions were available for use. Two of the conditions address permissive left-turn phasing, two address leading exclusive left-turn phasing, and the remaining four conditions address lead-

TABLE 1 INTERSECTION DATA

Intersection Name	Major Street Near Side Approach			Major Street Far SideApproach			Control Type	Left Turn Control	No. of Phases	Major Street ADT (1984)
	No.of Inbound Lanes	No. of Outbound Lanes	Left Turn Pocket Lane	No. of Inbound Lanes	No. of Outbound Lanes	Left Turn Pocket Lane				
UNIVERSITY AND ALMA SCHOOL	3	2	YES	3	2	YES	PRETIMED	PERMISSIVE	2	26860
ALMA SCHOOL AND BROADWAY	3	2	YES	3	2	YES	ACTUATED	EXCLUSIVE/ PERMISSIVE	8	27210
BROADWAY AND PRIEST	4	2	YES	3	2	YES	ACTUATED	EXCLUSIVE	8	35400
44TH STREET AND THOMAS	4	3	YES	4	3	YES	PRETIMED	PERMISSIVE	2	43500
THOMAS AND SCOTTSDALE	3	3	YES	4	2	YES	ACTUATED	EXCLUSIVE/ PERMISSIVE	8	27860
MAIN AND DOBSON	4	3	YES	4	3	YES	ACTUATED	EXCLUSIVE	8	23100

ing exclusive-permissive left-turn phasing. Data collected for the eight different intersection conditions were used in this study.

Of the six different intersections being studied, three intersections have two opposing lanes that the left-turning vehicles have to cross, whereas the other three intersections have three opposing lanes to be crossed. This variety was intentional in the selection of the intersections because previous studies have shown that drivers making a left turn have greater difficulty in identifying acceptable gaps when there are three lanes of opposing traffic than when there are only two opposing lanes (*1*).

DATA COLLECTION

The TEXAS simulation model requires a considerable amount of input data. Extensive data collection is important for properly re-creating the field environment on the computer. These requirements help in minimizing errors that could develop in the later stages because of lack of data.

The data collected from the previous three studies (*1–3*) were used as the basis for this study and were complemented with more data collection. The extra data were needed because the three studies collected data for only two of the four approaches at each intersection. The data were collected in each of the previous studies by filming each one of the intersection scenarios using a time-lapse camera. Each scenario was filmed continuously for 7 or 8 hr during the day. This filming procedure captured morning and evening peaks as well as off-peak periods. The camera was situated about 300 ft in advance of the intersection. The location and orientation of the camera provided a good view of the two intersection approaches parallel to the camera's direction and of the middle of the intersection, but it was occasionally difficult to identify the green arrow indication and the green ball indication of the signal head. Nevertheless, side street volumes classified by turning movements were observed in the films. These volumes were used in simulating the actuated signal intersections. The filming was conducted on weekdays.

The data collected from the previous studies were for the two approaches parallel to the camera's field of view. The data included the number of vehicles stopped and the number of vehicles not stopping for each of the two approaches. Within each approach, these data were collected for the left-turn movement and for the through movement separately. For those studies, the definition of the through movement incorporated both the through and right turns (*3*). For each hour of filming, the data were set up in convenient 5-min intervals and were used to calculate stopped delay. Additional data collected for those studies included the signal timing plans and intersection geometry. Intersection configuration for the approaches perpendicular to the camera's field of view was obtained from city drawings that indicated the number of lanes, the lane widths, and the number and position of the loop detectors.

Additional data, not extracted from the film in the previous studies, were extracted. One kind of additional data was vehicle type. The TEXAS model has 12 different vehicle classes. For practical purposes, these were reduced to seven different classes, which were a sports car, a compact, a medium car, a large car, a single-unit vehicle using gas, a single-unit vehicle using diesel fuel, and a semitrailer. Viewer judgment was required in categorizing the vehicles viewed to their appropriate class. Trucks were usually categorized as single-unit vehicles using gas, whereas buses were categorized as single-unit vehicles using diesel fuel.

A second kind of additional data was headway distribution data. It was collected for the near-side approach parallel to the camera's field of view. The time between two successive free flowing vehicles was recorded for each open lane of the approach separately. The data were recorded for the vehicles traveling freely through the intersection during a green period and without slowing down because of a queue. When this condition occurred, the time when a lead vehicle passed a fixed point on the screen was recorded and the time that subsequent vehicles passed the same point was also recorded until the free flow state was interrupted. The difference between these recorded times gave the headway in seconds.

A third kind of additional data was a continuous record of signal timing. This data was collected randomly in 5-min intervals in each hour.

After acquisition of the additional data for the eight intersection conditions and the selection of the default values for the variables that had no data collected, the TEXAS model was used to simulate the intersection operation. For each set of conditions that were simulated, 15 replications were conducted, each using a different random seed. Mean values of the average stopped delay from the 15 replications produced by TEXAS were used to compare with the field stopped delay data acquired in the three previous studies.

Each of the eight intersection conditions was analyzed separately. The volume data for each intersection were divided into 10-min intervals. For each hour of film, the first five 10-min intervals were considered. The last 10-min interval was omitted because each film was not exactly 1 hr long and varied from 52 min in some cases to 59 min in others.

Preliminary Trial Runs

Another task that took place parallel to the data collection stage was using the TEXAS model to perform preliminary trial runs and a crude sensitivity analysis for some of the variables in the model. The purpose of this analysis was to have a better comprehension of the default values assumed in the model and to study how they affect the delay results. These trial runs were performed on the variables that would not have available data and so the default values of the TEXAS model were to be used. The variables to be considered were the parameters of the car-following equation, the different types of the headway frequency distribution, the percentage of left-turning vehicles in the median lane, the percentage of right-turning vehicles in the curb lane, the warm-up period, and other calibration variables used in the TEXAS model.

The trial runs were conducted on the intersection of Thomas and 44th Street with the pretimed signal condition, using the average volume data for the 2:00 p.m. hour, the observed vehicle classification, and using the default-shifted negative exponential as the initial headway distribution with a parameter of 1.0. For these runs, a warm-up time of 5 min was used, with a simulation time of 30 min. All other geometry and signal data required for the model were taken from the observed and previously documented data. This condition was chosen because the intersection was the first to be completed in the data collection phase. The runs were conducted to evaluate the effect of some of the model variables on the delay results.

Two of the variables were the percentage of left-turning vehicles in the median lane and the percentage of right-turning vehicles in the curb lane. The default values in the model were 80 percent for each variable; these were both changed to 100 percent and the model was run. The average stopped delay for all movements on both near-side and far-side approaches was recorded. The percentage difference was observed to range between −8.5 and +7.5 percent. Although this experiment only considered just one run with no statistical analysis, it appears that the two variables in question do not significantly affect the results. Because of lack of data availability concerning these two variables, it was decided to use the default values of 80 percent both for the percentage of left turners in the median lane and for the percentage of right turners in the curb lane.

The next variables considered were the lambda and alpha parameters of the car-following equation. Three runs were conducted with (a) changing the lambda value from the default of 2.8 to 2.4, and (b) changing the alpha value from the default of 4,000 to 6,000 and from 4,000 to 2,500. When each change was implemented, all other conditions and variables remained constant. The results from these three runs and the percentage difference that occurred with respect to the initial condition are presented in Table 2.

Closer examination of the results with lambda changed to 2.4 indicated that the change in the delay ranged from −16.0 to +129.8 percent. The range was between −16.0 and +12.9 percent for the right-turn movement and between −0.7 and +7.5 percent for the through movement, which did not seem significant. The major changes occurred for the left turns, which ranged from +129.8 to +11.5 percent. Changing the alpha parameter to 2,500 made the values of the percentage difference range from −44.2 to +2.8 percent. Again the wider range occurred for the left turns, ranging from −44.2 to −1.3 percent. The ranges for the through movement and right turns are closer together and are between −7.5 and +2.8 percent and −9.6 and −1.2 percent, respectively. The same observations are valid when the alpha parameter was changed to 6,000. To be able to conduct a statistical analysis on the attained results and draw some conclusions, a com-

TABLE 2 EFFECT OF CHANGING THE CAR-FOLLOWING EQUATION PARAMETERS ON THE AVERAGE STOPPED DELAY

Conditions	AVERAGE STOPPED DELAY (SEC)					
	Approach 1 (Southbound)			*Approach 2 (Westbound)*		
	Left	Thru	Right	Left	Thru	Right
Initial Condition (alpha = 4000, lambda = 2.8)	179.9	13.5	9.4	37.9	12.6	8.0
Change lambda to 2.4	228.1	13.7	7.9	87.1	12.8	9.0
Percentage Difference %	26.8	1.5	-16.0	129.8	1.6	12.5
Change alpha to 6000	200.4	14.4	8.8	49.6	12.8	8.1
Percentage Difference %	11.4	6.7	-6.4	30.8	1.6	1.3
Change alpha to 2500	100.3	13.2	8.5	22.5	12.5	7.8
Percentage Difference %	-44.2	-2.2	-9.6	-40.6	-0.8	-2.5
	Approach 3 (Northbound)			*Approach 4 (Eastbound)*		
	Left	Thru	Right	Left	Thru	Right
Initial Condition (alpha = 4000, lambda = 2.8)	186.9	14.1	8.4	22.3	12.0	9.3
Change lambda to 2.4	208.4	14.0	7.2	45.2	12.9	10.5
Percentage Difference %	11.5	-0.7	-14.3	102.7	7.5	12.9
Change alpha to 6000	227.9	13.7	6.9	28.3	12.3	8.6
Percentage Difference %	21.9	-2.8	-17.9	26.9	2.5	-7.5
Change alpha to 2500	154.2	14.5	8.3	22.0	11.1	9.0
Percentage Difference %	-17.5	2.8	-1.2	-1.3	-7.5	-3.2

prehensive sensitivity analysis is deemed essential, a task which is considered beyond the scope of this study. Nevertheless, the results indicate that the alpha and lambda parameters significantly affect the average stopped delay for the left-turn movements.

The fifth variable considered in these trial runs was the headway distribution. The default value used by the TEXAS model is the shifted negative exponential distribution with a parameter of 2.0. The parameter is defined as the minimum headway in seconds; it should be less than or equal to the mean headway. From the field headway data, the minimum headway was 1.0 sec, which was used for the parameter of the headway frequency distribution.

This distribution was compared with three other runs representing three different distributions. The other distributions considered were the gamma, the Erlang, and the negative exponential. The parameter required for the gamma was the square of the mean divided by the variance for the headway observations with the limitation that it must be greater than 1.0. From the field data, the calculated gamma parameter was 1.06. The Erlang parameter uses the same formula, but the limitation is that it must be an integer value greater than 1, so the parameter used was 2.0. The final distribution, which was the negative exponential, did not require a parameter.

The same pattern of results was noticed, namely that the range for the left turn was much wider than that for the through or right turn. The value of the percentage difference for the gamma distribution was observed between +66.4 and −40.5 percent, between −41.3 and +14.3 percent for the Erlang distribution, and between +40.0 and −25.1 percent for the negative exponential distribution. Without going into a detailed sensitivity analysis, the results indicate that the headway frequency distribution affects the average stopped delay. It is important to point out that the results of this analysis are limited to one intersection; including more sites in the analysis could possibly produce different conclusions.

For this study, the default shifted negative exponential headway frequency distribution was used, but with a parameter of 1.0 sec, which corresponds to the minimum headway observed. Again this decision was for practical reasons, because it is difficult, costly, and time consuming for a traffic engineer to conduct detailed headway data studies. Accuracy is an important factor because, for example, a change in grouping the data from 1- to 2-sec intervals could lead to a change from a gamma distribution to a shifted negative exponential, depending on the number of observations present in each interval.

The TEXAS model requires two time intervals before conducting the simulation, a warm-up time period and a simulation time period. The warm-up time is required so that the system can reach a steady or stable state before statistics are collected.

An experiment was designed to determine the sensitivity of the model to the length of the warm-up period and the length of the simulation time. In this experiment, both warm-up time periods and simulation time periods were changed. The volume for a 10-min period and the random number seed combination were kept constant. The simulation period was set at 10 min and then changed by being increased in 5-min increments up to 30 min. The time was not increased further because of the approach volume limitation of the TEXAS model. For each simulation period, the warm-up time was increased from 2 to 10 min in 1-min increments and the results were documented.

The ratio of the warm-up time to the simulation time was compared with the ratio of the number of vehicles processed during the warm-up period to the number of vehicles processed during the simulation period. Steady or stable state was assumed to occur when the difference between the two ratios was small. The definition of a processed vehicle in the TEXAS model is a vehicle that both enters and leaves the intersection system. When it leaves that system, it is considered to have been processed and statistical results for the vehicle are recorded.

The warm-up time runs were conducted for the fifth 10-min interval in the 4:00 p.m. film of the intersection of Thomas and 44th Street, which had a two-phase pretimed signal with a cycle length of 50 sec. This intersection was used because all the required data were collected for that intersection and the time interval chosen was the most critical one in the afternoon peak period. The results presented in Table 3 indicate that a warm-up period of 8 min is the optimum case for shorter simulation periods of 10 to 15 min, whereas a 9- to 10-min warm-up period is better for longer simulation periods of 20 to 30 min. No significant trends were observed and the percentage change value was not stabilizing as simulation periods became longer.

The warm-up time was further increased from 10 to 15 min using 1-min increments but still no trend was observed. Further investigation of the warm-up experiment was beyond the scope of this study. The stability of the warm-up period is certainly sensitive to the left-turn phasing. The user manual of the model recommends a warm-up period of 2 to 5 min depending on the traffic volumes. Because for most of the intersection conditions studied the volumes were usually above 1,000 vehicles per hour, a 5-min warm-up period was used.

The TEXAS model automatically produces a random number seed if the user failed to input one. The problem with using the default seed is that when a new run is initiated, the model will produce the same random number seed. This process results in the generation of the same arrival pattern. To change this default seed, the random number seed was entered as part of the data and changed each time a new run was done. To ensure that the seed selection was random, the numbers were generated from an external random number generator using the GWBASIC program found in MS-DOS. A subroutine was written to generate a random number table of 5 columns and 15 rows. The first column was used as the random number seed for the first approach and the second column for the second approach up to the fourth column for the fourth approach. The 15 rows corresponded to the random number seed for the 15 runs. When the initial 15 runs were completed, a few runs gave errors because of the random seed combination. To remedy this problem, the random number columns used were shifted one space for that particular run.

The last set of test runs was conducted to test the impact of selected calibration variables on the model output. Three variables were identified: the driver operational factor for different driver classes (IDCHAR), the operating characteristics for different vehicle classes (IVCHAR), and the perception-reaction time for different driver classes (PIJR). Several computer runs were conducted using the extreme val-

TABLE 3 RESULTS OF THE WARM-UP PERIOD EXPERIMENT

THOMAS AND 44TH STREET

WARM UP TIME (MIN.)

SIM TIME (MIN.)	2	3	4	5	6	7	8	9	10	
10	.2	.3	.4	.5	.6	.7	.8	.9	1	Warm up/simulation (time)
	.11529	.24540	.32380	.45802	.57641	.63882	.78645	.81119	.93772	Warm up/simulation (vehicles)
%Change	42.354	18.2	19.050	8.3910	3.9351	8.7399	1.6944	9.8679	6.2275	% Change
15	.13333	.2	.26667	.33333	.4	.46667	.53333	.6	.66667	Warm up/simulation (time)
	.09434	.16223	.22846	.28413	.34852	.45891	.52573	.58632	.624	Warm up/simulation (vehicles)
% Change	29.245	18.885	14.328	14.762	12.870	1.6622	1.4258	2.2797	6.4	% Change
20	.1	.15	.2	.25	.3	.35	.4	.45	.5	Warm up/simulation (time)
	.06790	.12666	.16687	.20977	.28633	.31075	.38168	.43075	.48635	Warm up/simulation (vehicles)
% Change	32.102	15.558	16.564	16.091	4.5558	11.215	4.5790	4.2777	2.7295	% Change
25	.08	.12	.16	.2	.24	.28	.32	.36	.4	Warm up/simulation (time)
	.05699	.09339	.13780	.17336	.21384	.25208	.29217	.33446	.38204	Warm up/simulation (vehicles)
% Change	28.756	22.179	13.877	13.320	10.901	9.9699	8.6977	7.0955	4.4890	% Change
30	.06667	.1	.13333	.16667	.2	.23333	.26667	.3	.33333	Warm up/simulation (time)
	.04631	.07905	.11142	.14932	.18881	.20974	.24202	.29031	.31167	Warm up/simulation (vehicles)
& Change	30.533	20.947	16.433	10.405	5.5944	10.113	9.2424	3.2302	6.4998	% Change
35	.05714	.08571	.11429	.14286						Warm up/simulation (time)
	.04084	.06625	.09789	.12617						Warm up/simulation (veh. proc.)
% Change	28.534	22.708	14.342	11.681						% Change

ues suggested for these variables and little or no changes were observed in the left-turn delay statistics. The reason for the little changes in the results is that all six intersections are heavily traveled and changes in the driver behavior or vehicle characteristics have limited impact on the intersection capacity. Closer examination of the left-turning maneuvers, using the animation option of TEXAS, indicated that the model is incapable of simulating the left-turn laggers.

Determining the Critical Time Periods

Each intersection was divided into 35 to 40 ten-minute time intervals depending on the number of hours of film (7 or 8) for each location. The process of taking every interval (approximately 35 intervals) and running it 15 times (15 replications) for each intersection condition would have been extremely time consuming. Therefore it was decided to choose representative intervals that would be the most critical for each intersection.

The procedure undertaken was to arrange the 35 to 40 ten-minute approach volumes for the near-side approach of an intersection in ascending order. The emphasis was on the two approaches parallel to the camera's field of view because they were the ones for which average stopped delay was calculated in the previous studies. Table 4 indicates this procedure. The table contains a column for the left-turn volume of the near approach, the percent of left turns, the volume of the opposing approach, the cycle length, and the product parameter.

The product parameter is reached by multiplying the approach volume by the percentage of left turns in decimal form and by the opposing volume. The range of the approach volume is divided into 6 to 8 intervals and a frequency table is generated as indicated in Table 4. Out of each frequency interval, the critical reading is selected on the basis of the greater value of the product parameter. This was the basis for the selection procedure of the critical 10-min time periods. The actual selection differed somewhat depending on the type of left turn phasing at that particular intersection.

For permissive left-turn phasing, the most critical parameter is the volume on the opposing approach. Therefore, for this condition the critical 10-min time period chosen in each interval is the one satisfying the greater value of the product parameter. If one of the intervals has more than one cycle

TABLE 4 SELECTION OF THE CRITICAL 10-min PERIODS

Intersection: Alma School & Broadway (Exclusive/Permissive Left Turns)

Approach: Northbound (leg 3)

Hourly Volume (VPH)	*Left Turn Volume*	*Percent of Left Turns*	*Opposing Volume (VPH)*	*Product Parameter*		*Frequency*	*Table*
765	107	14	900	96390		750-850	2
786	165	21	1128	186188	----1	850-950	7
858	129	15	858	110425		950-1050	10
864	147	17	774	113685		1050-1150	10
888	186	21	1050	195804	---2	1150-1250	7
912	128	14	792	101123		1250-1350	4
930	140	15	990	138105			
942	170	18	930	157691		Maximum Cycle	
948	142	15	864	122861		Length = 148.5 seconds	
954	153	16	930	141955			
954	181	19	924	167484			
954	105	11	954	100113			
960	154	16	1128	173261			
978	137	14	888	121585			
996	100	10	804	80078			
1002	160	16	1434	229899	---3		
1014	172	17	930	160313			
1044	188	18	810	152215	---4		
1044	146	14	870	127159			
1068	171	16	1158	197879			
1068	182	17	1344	244017			
1080	173	16	978	168998			
1086	185	17	1134	209359			
1098	285	26	1236	352853	----5		
1104	188	17	1098	206073			
1116	246	22	1116	274000			
1134	91	8	1446	131181			
1140	160	14	1266	202054			
1146	218	19	954	207724			
1152	207	18	1446	299843			
1158	151	13	1572	236649			
1158	139	12	1512	210108			
1164	128	11	1140	145966			
1170	140	12	1350	189540			
1194	96	8	1452	138695			
1194	215	18	1308	281115	---7		
1272	191	15	1140	217512			
1290	232	18	930	215946	---8		
1320	198	15	984	194832			
1332	173	13	918	158961			

length, the selection process is conducted for each value of the cycle length separately.

The critical parameter for the exclusive-permissive left-turn phasing is also the volume on the opposing approach for the permissive part and the left-turn volume on the near approach for the exclusive part. The selection process for this case involves finding, within each interval, the 10-min period that has the greater value of the product parameter and at the same time the greater value of the left-turn volume on the near approach. The best selection is a 10-min period that satisfies both criteria. If this is not possible, then two 10-min periods are selected—one for each criteria. Again, if there is more than one cycle length in the interval, the selection is conducted for each cycle length separately.

For the exclusive left-turn phasing, the critical parameter is the left-turn volume on the near approach. The selection criterion for this condition is the greater value of the left-turn volume on the near approach. Usually this 10-minute period would have a high value for the product parameter but not necessarily the greater value. When there is more than one cycle length in each interval, the selection is conducted for each cycle length separately.

Table 5 presents the findings of the selection process for all eight intersection conditions. A total of 87 critical 10-min periods were selected and each one of the periods was run 15 times using the TEXAS model resulting in 1,305 individual runs.

TABLE 5 SUMMARY OF THE CRITICAL TIME PERIODS SELECTED

Intersection Condition	Number of Intervals	Number of 10 min Periods Selected
University and Alma School	9	11
Alma School and Broadway	6	8
Priest and Broadway	10	8
Thomas and 44th Street	7	10
Stonex's Scenario	8	12
Warne's Scenario	9	15
Scottsdale and Thomas	7	14
Dobson and Main	7	9

Running The Model

Once this stage was completed for each intersection condition, it was time to run the model and document the results. For each one of the selected 10-min periods, the corresponding 10-min volume for the approaches parallel to the camera's field of view was adjusted to an hourly volume. The data relevant to each 10-min period were keyed into the TEXAS model, i.e., the geometry, driver-vehicle data, and simulation data. The model was run 15 times for each particular case with the replacement of the random number seed for each approach being the only change conducted from one run to the next. For each run, a warm-up period of 5 min and a simulation period of 10 min were used. After the model had completed execution, the results for the required variables were documented and keyed into a spreadsheet. The results that were of importance to this study were the average stopped delay for each movement on the approaches that were parallel to the camera's field of view. The output of the TEXAS model gives extensive results for each movement on each approach as well as summary results for each approach and for the whole intersection. Only the two approaches parallel to the camera's field of view were considered for comparison because the field data from the previous studies were only for those two approaches.

The decision of using a warm-up period of 5 min and a simulation period of 10 min was achieved from the sensitivity analysis conducted for the Thomas and 44th Street intersection. The signal cycle length at that intersection is 50 sec. Five-minute warm-up period and 10-min simulation period would result in 6 and 12 complete cycles of simulation, respectively. It was felt that these two periods are suitable to produce steady state conditions at most of the intersections. Increasing the simulation period would have required much longer computer time, and it was not clear how much improvement in the results would have been achieved.

The only intersection that should be treated with some caution is Alma School and Broadway. The traffic signal system for this intersection is composed of a fully actuated eight-phase controller with a maximum cycle length of 148.5 sec. If the cycle reaches the maximum value consistently during the warm-up and simulation period, then it would result in two and four complete cycles, respectively. The results of this analysis may become questionable. The type of left-turn phasing at this intersection is exclusive-permissive. There are three other intersection conditions for exclusive-permissive left-turn phasing and they all have much shorter cycle length.

RESULTS AND ANALYSES

In order to display the results in a logical form, the eight intersection conditions are grouped by the type of left-turn phasing. This grouping is appropriate because the validation process emphasized the type of left-turn phasing. The intersection conditions of Alma School and Broadway, Stonex's (*2*) scenario, Warne's (*3*) scenario, and Scottsdale and Thomas all had an exclusive-permissive left-turn phase. The intersection conditions of University and Alma School and 44th Street and Thomas had a permissive left-turn phase. Finally, the intersection conditions of Broadway and Priest, and Dobson and Main, had an exclusive left-turn phase. The following section contains the results for one out of the eight intersection conditions. A summary of all statistical results is documented in a later section.

Results of Alma School and University (Permissive Left-Turn Case)

The number of critical 10-min periods selected were 11 for Alma School and University. The results obtained from running the TEXAS model 15 times for each 10-min period are presented in Table 6. Each average stopped delay value for the left-turn movement and for the through movement is the mean of the 15 runs for that time period.

Examining the results, both the eastbound approach and the westbound approach have similar trends. The mean stopped delay of the 11 time periods for both left turns obtained from the simulation are more than three times the mean stopped delay obtained from the field and presented in the observed column. The left-turn standard deviation for the simulation results are 61.5 and 76.2 sec, respectively, whereas the standard deviation for the field results are 15.3 and 13.3 sec, respectively. The through movements have the mean stopped delay for the simulation within 2 sec of the field results; the standard deviation for the mean stopped delay in the field is at least 4 times greater than in the model. The simulation results indicate that there is a higher variation in the delay between the different time periods for the left-turn movement than for the through movement on the same approach. This variation is not that extreme for the observed delay data.

A reason for the high variation in left-turn delay results in the model is because of the fact that left-turn movements take place when opposing through traffic is outside a conflict region. Therefore, the delay results vary depending on the combined effect of the opposing volume and the arrival pattern of the opposing traffic. This explanation is not applicable for the through movement; so the delay results are narrowly spread within a small range. For the left-turn movement, the delays from the model are high compared to the field delay because the model does not accurately represent driver behavior. The conflict region explanation is more conservative than gap acceptance where a left turn would be made when an appropriate gap occurs in the opposing traffic. In addition, the model makes left-turning vehicles stop at the stop bar and therefore only one vehicle can turn during the yellow phase per cycle. Field data indicated that two, and in some cases three, sneakers turned left during the yellow phase. Another

TABLE 6 AVERAGE STOPPED DELAY RESULTS FOR ALMA SCHOOL AND UNIVERSITY FOR PERMISSIVE LEFT TURNS

Eastbound Approach*			*Average Stopped Delay (sec)*							
Left Turn Volume	Opposing Volume	Time of Day	Eastbound Approach				Westbound Approach			
			Left Turn		Through		Left Turn		Through	
(VPH	(VPH)		Sim	Obs	Sim	Obs	Sim	Obs.	Sim	Obs
35	942	8:35 am	29.8	10.5	14.0	12.4	25.9	36.7	14.4	14.0
75	660	8:45 am	33.9	26.0	14.5	11.7	30.7	45.0	14.8	15.5
91	786	8:55 am	39.2	15.0	14.4	13.9	35.9	24.5	14.5	17.5
67	864	9:40 am	109.7	11.4	15.0	15.9	56.9	19.0	15.3	16.7
113	948	10:50 am	113.8	42.0	14.5	18.5	62.8	35.0	15.9	20.0
76	1278	11:30 am	180.1	24.5	16.1	19.2	179.1	57.7	15.5	13.2
93	996	1:50 pm	80.8	32.3	15.0	17.0	109.8	35.0	14.6	13.1
155	918	2:00 pm	31.8	14.5	14.2	13.3	50.7	19.2	14.3	11.5
127	1158	2:45 pm	90.3	33.1	14.5	16.9	90.4	28.0	13.9	11.9
114	1266	3:50 pm	147.2	48.2	17.0	35.8	222.6	-----	16.5	19.2
114	1245	4:20 pm	187.3	50.8	16.1	16.0	206.5	-----	16.5	23.3
Mean Delay			94.9	28.0	15.0	17.3	97.4	33.3	15.1	16.0
Std. Dev.			61.49	15.26	1.00	6.90	76.22	13.29	0.94	3.91

----- Not Available
*Time lapse Camera's Position

possible reason is that real-world driver behavior tends to be more aggressive than anticipated and depends on many factors.

The data points are graphically represented and are shown in Figures 1 and 2. Each figure has the simulated delay data plotted against the observed delay. On the same graph, a line that has a slope equal to unity is drawn that represents the plotting of the observed delay data against itself. The scatter plot indicates how much the simulated delay data deviates from the observed delay data. Evidence that the simulation model performs well would be found if the simulated data points fall on or closely around the unit slope line. Figure 1 shows the results of the eastbound left-turn delays with all the data points falling above the unit slope line. Figure 2 shows the results of the eastbound through delays in which the data points fall above and below the line and are grouped in a small range.

Figures 1 and 2 suggest, but do not prove, that the simulation model works well in some cases, and performs poorly in other cases in predicting average stopped delay. To prove or disprove that the model works well, statistical tests were used. The null hypothesis was that the difference between the observed average stopped delay (from field data) and the simulated average stopped delay (TEXAS model) equals zero. The paired-data *t*-test was adopted for the statistical analysis and the results are presented in Table 7. The mean of the differences and the standard deviation are used to obtain the calculated test statistic *t*. The results of the statistical analysis indicate that for a level of significance of 0.05 the null hypothesis was rejected for both eastbound and westbound left turns. The null hypothesis was not rejected for the through movements and on conducting the β test, the probability of accepting the null hypothesis was greater than 0.75. Therefore it was decided to accept the hypothesis that the difference between the observed means and the simulated means equals zero for both the eastbound and westbound through movements at a confidence level of 0.05. This finding suggests that for this intersection, the model does a good job in simulating the through movements; however, simulating left-turn movements appears to be inconsistent with the field data.

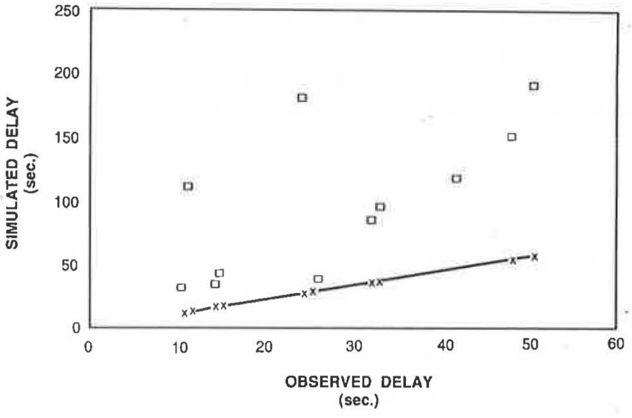


FIGURE 1 Alma School and University eastbound permissive left-turn results.

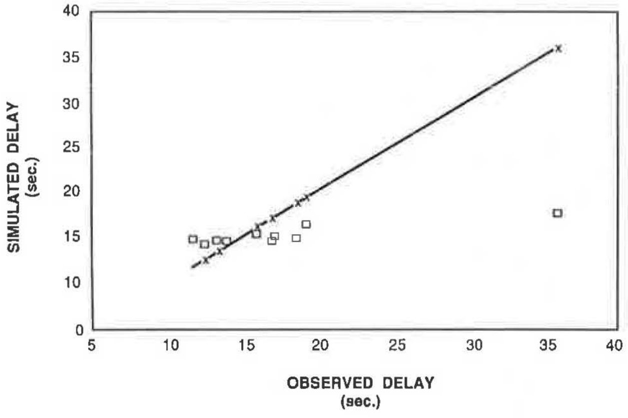


FIGURE 2 Alma School and University eastbound permissive through results.

TABLE 7 PAIRED-DATA *t*-TEST RESULTS FOR ALMA SCHOOL AND UNIVERSITY FOR PERMISSIVE LEFT TURNS

Time of Day	Difference Between Simulated and Observed Delay Results (sec)			
	Eastbound Approach		Westbound Approach	
	Left Turn	Through	Left Turn	Through
8:35 am	19.31	1.62	-10.81	0.36
8:45 am	7.94	2.85	-14.32	-0.67
8:55 am	24.19	0.50	11.37	-3.00
9:40 am	98.29	-0.87	37.94	-1.33
10:50 am	71.77	-4.01	27.76	-4.14
11:30 am	155.62	-3.09	121.41	2.28
1:50 pm	48.52	-1.94	74.77	1.46
2:00 pm	17.31	0.90	31.54	2.79
2:45 pm	57.20	-2.39	62.36	2.03
3:50 pm	99.01	-18.77	----	-2.63
4:20 pm	136.52	0.12	----	-6.75
Mean Difference	66.88	-2.28	38.00	-0.87
Standard Deviation	52.67	6.14	35.74	3.18
t calculated	4.212	-1.231	2.755	-0.911
Conclusion	reject H_0	cannot reject H_0	reject H_0	cannot reject H_0
Probability to Accept $H_0(1-\beta)$	———	0.83	———	0.89

Note: $n = 11, t_{(0.025,10)} = 2.228$

Results of the Permissive Left-Turn Combined Data

The statistical analyses of the second permissive left-turn phasing intersection (Thomas and 44th Street) indicated similar results to those of Alma School and University intersection. In both cases it was concluded that the TEXAS model overestimates delay for left-turn movements. This finding was previously shown in Figures 1 and 2 where points were scattered above the unit slope line. It was decided to combine the data of both intersections and attempt to fit a linear regression line for those scattered points. If a straight line parallel to the unit slope line is successfully fit, then this line and the new line would measure the average deviation between the field data and the TEXAS model results. Linear regression was conducted to fit a line to the scattered data points. The slope coefficient was statistically tested with the null hypothesis stating that the slope coefficient equals 1. If the null hypothesis is accepted, then the regression line through the data points is parallel to the unit slope line, but it is shifted upward at a certain value equal to the value by which the model overestimates delay. Results of the regression analyses are presented in Table 8. The statistical test conducted to find out if the slope coefficient equals 1 were all rejected except for the left-turn delay data for the near-side approach. A β test was conducted; the probability of accepting the null hypothesis that the slope coefficient equals 1 was 0.93. Therefore it is safe to say that the left-turn delays for the near-side approach obtained from the TEXAS model follow the same trend as the observed left-turn delay results for the same approach but with an increased delay of 66.8 secs.

Results of the Exclusive-Permissive and Exclusive Left-Turn Cases

The statistical analysis procedure applied to the two permissive intersection cases was also applied to the exclusive-permissive and exclusive left-turn cases. For exclusive-permissive left-turn phasing, four intersection cases were analyzed. In all four cases, it was observed that the model overestimates delay for left-turn traffic.

As for exclusive left-turn phasing, two intersections met the criteria, and the statistical analyses indicated that the model overestimates delay for left-turn traffic. Comparing the differences between the simulated and observed field data across all three types of left-turn phasing, it was concluded that the differences were the greatest for permissive phasing, smaller for exclusive-permissive phasing, and the least for exclusive phasing. Naguib (7) provided detailed information concerning the results.

A summary of all the statistical results for the eight intersection cases and the combined data is presented in Table 9. The significant conclusion reported for the eight intersection cases means that the simulated mean stopped delays produced by the TEXAS model are significantly different than the mean stopped delays observed in the field at a level of confidence of 95 percent.

TABLE 8 REGRESSION ANALYSIS RESULTS FOR PERMISSIVE LEFT-TURN PHASING

Parameters	Near Side Approach		Far Side Approach	
	Left Turn	Through	Left Turn	Through
Slope Value	1.132	0.231	1.814	0.227
Std. Error of Slope	0.266	0.075	0.330	0.066
Intercept Value	66.82	11.72	17.52	11.37
R Squared	0.489	0.331	0.640	0.380
t calculated	0.495	-10.226	2.467	-11.659
Conclusion	cannot reject H_o	reject H_o	reject H_o	reject H_o
Probability to Accept Ho (1-β)	0.93	———	———	———

Notes : H_o: $b_1 = 1$ (slope is unity), H_a: $b_1 = 1$ (slope is not unity)
$n = 21$, $t_{(0.025,19)} = 2.093$

TABLE 9 SUMMARY OF STATISTICAL RESULTS

Approach		Near Side		Far Side	
		Left Turn	Through	Left Turn	Through
Permissive	Alma School and University	S	N.S.*	S	N.S.*
	Thomas and 44th Street	S	S	S	N.S.
Exclusive/ Permissive	Alma School and Broadway	S	N.S.*	N.S.*	N.S.*
	Stonex's Scenario	S	S	S	N.S.
	Warne's Scenario	S	N.S.*	S	N.S.*
	Scottsdale and Thomas	S	N.S.*	N.S.	N.S.*
Exclusive	Broadway and Priest	S	N.S.*	S	N.S.*
	Dobson and Main	S	S	S	S
Combined Data	Permissive	N.S.*	S	S	S
	Exclusive/ Permissive	N.S.	S	N.S.	S
	Exclusive	N.S.*	S	N.S.*	S

Notes: S -denotes a significant conclusion from the null hypothesis.
N.S. -denotes a non-significant conclusion from the null hypothesis.
* -denotes it also passed the β-test criteria.

The combined data includes the three left-turn phasing treatments. The permissive left-turn treatment combines two intersection results, the exclusive-permissive treatment combines four intersection results, and the exclusive treatment combines two intersection results. The significant conclusion reported for the combined data means that the regression line fitted to the delay data produced by TEXAS has a slope significantly different from unity at a level of confidence of 95 percent.

Table 9 indicates that all the near-side left-turn delays rejected the null hypothesis that they do not differ from the observed delay, while all the through delays for the combined data were not parallel to the unit slope line. This finding may appear negative. The nonsignificance conclusion means that the left-turn delay estimated by the TEXAS model follows the same trend as the observed delay results. It is a positive finding because one can determine the deviation between the two lines and use it for calibrating the different variables of the model.

CONCLUSIONS

The main goal of this study was to test the TEXAS model, as provided to the users without changing its default parameters, against field data.

The limited number of trial runs indicated that the percentage of left-turning vehicles in the median lane and the percentage of right-turning vehicles in the curb lane had no significant effect on the left-turn and through delay figures.

However, the alpha and lambda parameters of the car-following logic were found to have significant impacts on the average stopped delay, especially for the left-turn movement.

The headway distribution used in the model to generate vehicles was found to have a significant impact on the delay results. The user is advised to collect field data on headway distribution to fit it to one of the distributions provided by the TEXAS model. As for the warm-up period, it was concluded that a more detailed experiment is needed to identify the region of steady state for different left-turn phasing schemes.

The statistical results concluded that the TEXAS model is capable of simulating through movements fairly well. The ability of the model to graphically display intersection geometrics and to provide animation of traffic movements was certainly helpful in coding the data properly.

The results further indicated that delay figures estimated for left-turn traffic were higher than those observed in the field for all three types of left-turn signal phasing (permissive, exclusive-permissive, and exclusive). The difference between the model results and the field data was found to be the greatest for permissive phasing and the smallest for exclusive phasing. It is believed that the permissive left-turn logic needs some refinements. Closer examination of the data revealed that the TEXAS model does not simulate left-turn laggers. Furthermore, vehicles who are attempting a left turn during a permissive phase are queued behind the stop bar, and not permitted to enter the intersection until an acceptable gap in the opposing traffic is presented.

Although the simulated delay figures were significantly different than the field data figures, all three types of left-turn phasing indicated that the delay estimated by the TEXAS model followed the same trend as the observed delay results.

ACKNOWLEDGMENT

The authors would like to acknowledge the Arizona Department of Transportation for sponsoring this research project.

REFERENCES

1. J. S. Matthias and J. E. Upchurch. *Left Turn Signal Warrants for Arizona*. FHWA/AZ 85/192. Arizona Department of Transportation, Phoenix, June 1985.
2. A. Stonex. *Effects of Changing Permissive Left Turn Phasing To Exclusive/Permissive*. M.S. thesis, Arizona State University, Tempe, 1985.
3. T. R. Warne. *Effects of Changing Exclusive/Permissive Left Turn Phasing From Variable To Fixed Cycle Length*. M.S. thesis, Arizona State University, Tempe, 1988.
4. C. E. Lee, R. B. Machemehl, R. F. Inman, C. R. Copeland, and W. M. Sanders. *User-Friendly TEXAS Model—Guide to Data Entry*. Research Report 361-1F, Center for Transportation Research, University of Texas, Austin, Nov. 1985.
5. C. E. Lee, T. W. Rioux, and C. R. Copeland. *The TEXAS Model for Intersection Traffic—Development*. Research Report 184-1. Center for Transportation Research, University of Texas, Austin, Dec. 1977.
6. C. E. Lee, V. S. Savur, and G. E. Grayson. *Application of The TEXAS Model for Analysis of Intersection Capacity and Evaluation of Traffic Control Warrants*. Research Report 184-4F. Center for Transportation Research, University of Texas, Austin, July 1978.
7. F. Naguib. Validation of the TEXAS Simulation Model Using Field Data. M.S. thesis, Department of Civil Engineering, Arizona State University, Tempe, Aug. 1989.

Views expressed in this paper are those of the authors and not of the Arizona Department of Transportation. The authors are responsible for the accuracy of the results presented in this paper.

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

Simulation of Two- and Four-Way Stop Control

R. J. Salter and E. A. Ismail

The development of digital computer simulation models representing vehicle interactions at highway intersections controlled by two- and four-way stop control is described. Driver actions at the stop lines are made on the basis of gap acceptance, and the appropriate distribution of gap acceptance may be selected on the basis of site experience. Demand flows on the approach highways may also be selected according to traffic flow conditions that vary from a uniform flow to a peak-hour flow with pronounced variation between maximum and minimum flow levels. Output from the simulation models is in the form of queue lengths and average and total delays. Observations of lag and gap acceptance and vehicle headway distributions are used in the simulation models to evaluate the relative advantages of two- and four-way stop control in terms of average delay. For the traffic flow conditions simulated, values of total practical intersection capacity occurred when the total inflow was approximately 1,400 veh/hr for a two-way stop intersection with one-lane approaches and 1,650 veh/hr for a four-way stop intersection with one-lane approaches.

When at-grade stop-controlled intersections are used, the capacity of a highway network is frequently limited by the capacity of the intersecting highways. For this reason, the capacity of, and delays at, at-grade intersections are of considerable interest to highway engineers.

The development of improved guidelines for the use of stop sign control at highway intersections was addressed by Upchurch (*1*), who selected as principal criterion for the evaluation of different control strategies the total cost of operation. Three types of control were analyzed: yield control, two-way stop control, and four-way stop control. The traffic experimental and analysis simulation (TEXAS) intersection simulation model was used to generate data on intersection operation for the calculation of vehicle operating and delay costs. Results indicated that for the accident rates used, yield control was the most economical type of control. Moreover, in most cases, two-way stop control was considered to be economically preferable to four-way stop control. However, for selected conditions of low volume or where minor-street volume was approximately equal to major-street volume, four-way stop control was found to be preferable.

Lee and Savur (*2*) had previously investigated intersection capacity and level of service (LOS) by using the TEXAS simulation model. Queue delay was used as a performance indicator and was recommended as the best indicator of LOS for stop-controlled intersections.

DEVELOPED SIMULATION MODELS

It was decided to use a simulation approach to investigate the relative advantages of two- and four-way stop control. Regular time scanning was adopted in the simulation models that were developed, the selection of the appropriate scanning interval depending on the level of accuracy required and on the available computer resources. From these considerations, a scanning interval of 0.5 sec was adopted. The performance indicators used were

1. Maximum and average queue length,
2. Total queuing delay to all vehicles, and
3. Average queuing delay to all vehicles.

The main models developed in this study were

1. Model TWSSIM: This model was developed to depict vehicular interactions at crossroad intersections controlled by a two-way stop (major-minor) priority rule.
2. Model FWSSIM: This model was developed to simulate the performance of a crossroad intersection controlled by a four-way stop (off-side) priority rule.

The main simulation models, together with the other necessary submodels, were built into a general and modular computer program, SIMPHINT. The details of the programming aspects are discussed by Ismail (*3*). Figure 1 shows an outline layout of the system. In general, the system was divided into three basic components: the main simulation models, the vehicle-driver submodels, and the computation submodels. The main simulation models described the movements of the different streams within the intersection area under the given priority rules and geometric layouts. The vehicle-driver submodels provided the required input values for the main simulation models. These input values, such as the critical acceptable lags and gaps, the time headways between vehicles, and the assignment of turning movements, are generated using the appropriate predefined distributions. The computation submodels carry out the necessary calculations required to update the values and compute the results summary during the simulation run.

THE GAP ACCEPTANCE SUBMODEL

Both types of priority rules considered depended mainly on the concept of gap acceptance, which has been extensively studied in the United Kingdom for two-way stop intersections

Department of Civil Engineering, University of Bradford, Bradford BD7 1DP Great Britain.

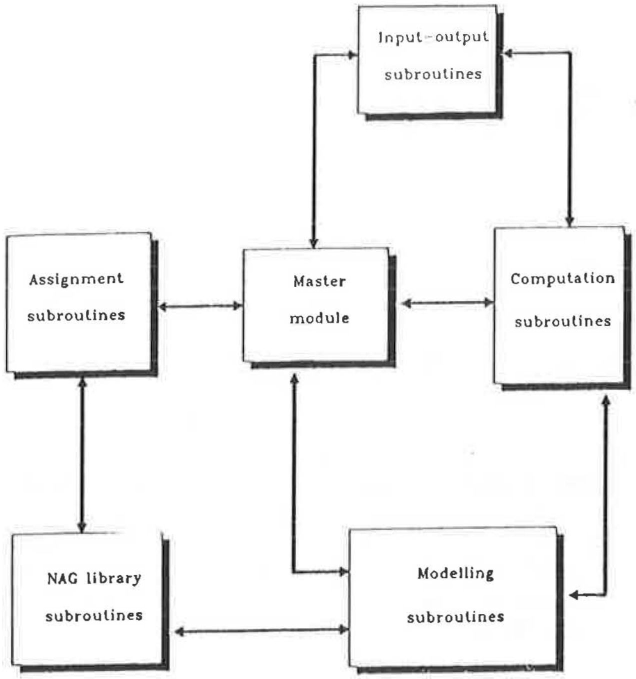


FIGURE 1 The general conceptual approach.

but not for four-way stop intersections. Vehicles arriving at the junction from controlled approaches are obliged to give way to vehicles on the controlling approaches. Four options for the acceptance distribution were included in the model:

1. A step function,
2. A cumulative normal distribution,
3. A cumulative log-normal distribution, and
4. A linear distribution.

Drivers are considered to be consistent in their decision to accept, or reject, the available gaps, i.e., the critical acceptable gap value assigned to the waiting driver remains unchanged until an available gap is accepted. In generating the critical acceptable values, the three different turning movements—left turn, straight ahead, and right turn (U.S. rules of the road)—are treated separately, because it is most likely that drivers' behavior in executing different turning movements is not unique. Different acceptance distributions may also be specified in the input data for the different intersection approaches. In assigning the critical acceptable values to each driver there are three possibilities, selection of which depends on the level of available data:

1. When detailed data are available to the traffic engineer, the waiting driver is first assigned a critical acceptable value from either a delayed gap or undelayed lag acceptance distribution, according to the vehicle's position in the queue. A delayed lag acceptance distribution is used for vehicles that arrive at the junction when a queue exists, whereas an undelayed lag acceptance distribution is used for vehicles arriving at the stop line when a queue does not exist. In the latter case, if the available lag is accepted, no further assignment is needed, whereas if that lag is rejected, then a critical acceptable gap is assigned for the waiting driver from the appropriate delayed gap acceptance distribution.
2. When less detailed data are available, no distinction is made between delayed and undelayed vehicles lags. Values of critical acceptable lags or gaps are assigned to waiting drivers regardless of whether they had joined the queue or arrived directly at the stop line. A critical acceptable gap is assigned to a waiting driver if the available lag is rejected.
3. A rather coarser approach in which each waiting driver is assigned a single value from a lag-and-gap acceptance distribution. This value is considered to represent the driver's critical acceptable value with which the available lag, and the subsequent available gaps, may be compared.

FLOW PATTERN SUBMODEL

The modeling of the flow pattern or profile throughout the simulation period is one of the factors that affects the realism of any simulation model. Generally, traffic flow varies considerably with time throughout the day and this variation is more pronounced during peak periods. Therefore, a demand flow profile that varies with time provides a more realistic and more accurate base for the calculation of queue length and delay at highway intersections. In the system developed, three options were considered to model the demand flow profile on the intersection approaches. In the first option, demand flow values, for each approach, can be input directly at specified time intervals. In the second option, it is only necessary to input three flow values to represent the peak period under consideration. These are prepeak, peak, and postpeak flow values. Three values on the time scale are also necessary; these are the time at which the flow starts to increase to the peak flow, the time of the peak flow, and the time at which the flow starts to decay to the postpeak flow. These flow and time values are used by the model to synthesize, from a normal distribution, the flow profile over the simulated period. The third option is to provide only the average demand flow value, for the whole simulation period, which will be used to synthesize the flow profile from a normal distribution using default values embedded in the model.

ASSUMPTIONS AND MODEL FRAMEWORKS

The general rules considered in the formulation of the simulation models in this study were as follows:

1. The headway distributions on the approaches to the intersection are represented by the displaced negative exponential distribution.
2. Drivers on the controlled streams, i.e., those who are waiting for an acceptable lag (or gap) in the controlling streams at two-way stop intersections, are assumed to be homogeneous and consistent in their responses to the available lags (or gaps) in those streams.
3. Vehicles arriving at the intersection in the controlled streams are put in a queue in the appropriate queuing lane according to the direction of movement and the layout of the approach.
4. No vehicle can enter the interaction zone while it is occupied by another vehicle, except for compatible movements.

5. Vehicles from the controlled stream can enter the intersection area if, and only if, an acceptable gap exists in all the controlling streams simultaneously.

6. The decision of direction of movement is assigned randomly to each vehicle arriving at the intersection on a given approach according to predefined turning proportions for that approach.

7. Drivers turning right (U.S. rules of the road) from the near-side controlling stream indicate their direction of movement at a suitable time prior to their arrival at the intersection, so that drivers on the controlled streams are able to consider the next available lag (or gap) on that controlling stream.

8. Parking is prohibited in the vicinity of the intersection area.

9. Overtaking is not allowed within the intersection area.

10. Interference between vehicles and pedestrians in the intersection area is not considered.

11. Queuing delay was only derived from the models.

12. All approaches to the intersection were considered to be subject to a 30-mph speed limit.

PARAMETRIC ANALYSES

The practical use of the simulation model can be demonstrated by performing a range of parametric analyses. This task may be accomplished by subjecting the validated model to hypothetical sets of input parameters, and by studying the effect of these parameters on the performance of the model. The analysis was mainly concentrated in two areas, type of intersection control and intersection geometry. In summary, more than 700 hr of vehicular interactions were simulated using the developed simulation program SIMPHINT, each simulation run representing 1 hr of real time.

EFFECT OF INTERSECTION CONTROL

Highway traffic engineers frequently face the problem of choosing the optimum type of control to be adopted at a particular intersection. At priority controlled intersections, the decision to use a certain priority rule or to change from one rule to another must be made according to guidelines on the basis of a specified criterion. The average delay per vehicle is widely accepted as such a criterion.

A series of simulation runs was made, and the required delay values were obtained from simulation model TWSSIM for an intersection with one-lane approaches under two-way stop control. These results are shown in Figure 2 over a range of input flow data. An average delay value of 60 sec, to all vehicles, was selected as a criterion to obtain the practical capacity of the intersection. Figure 2 shows that, at this delay value, the practical capacity of the intersection occurs when the total inflow is approximately 1,400 veh/hr over the range of minor-road flows shown.

Gap and lag acceptance is a critical phenomenon in the interaction between vehicles at priority highway intersections. During the validation of the simulation models, it was found that there were some differences between delayed and undelayed acceptance distributions. However, these differences were, in some cases, marginal. In order to investigate the effect of this factor, a series of simulation runs was made. In these runs, lag acceptance data were split into delayed and undelayed acceptance distributions. It was found that the effect was marginal without any significant difference in delay being observed, although it was expected that the difference could increase with an increase in total intersection flow. Nevertheless, in the light of more significantly different acceptance distributions, a different conclusion might be drawn.

The other form of intersection control that was of interest in this study was the four-way stop priority rule. In this priority rule, all the approaches to the intersection area are controlled by stop signs. Drivers approaching the intersection on a given approach are obliged to give way to those arriving from the left (U.S. rules of the road). In order to study the performance of highway intersections controlled by a four-way stop priority rule, the simulation model FWSSIM was used to obtain simulated delay values at hypothetical one-lane approach intersections controlled by a restricted four-way stop priority rule. Flow levels on the two intersecting roads were varied from 100 to 800 veh/hr in each direction, in increments of 50 veh/hr, to cover a wide range of traffic conditions. The turning proportions assumed were 15, 70, and 15 percent for left-turning, straight-ahead, and right-turning movements (U.S. rules of the road), respectively.

The variation of average delay with major-road flow, for different minor-road flow levels, was used to study the intersection performance. Figure 3 shows that for a range of flow values the average delay increases with an increase in major road flow. In order to obtain the practical capacity of the intersection under this priority rule, i.e., four-way stop, the same average delay criterion of 60 sec to all vehicles was selected. It was concluded that the practical capacity of the intersection under this type of control occurs when the total inflow is about 1,650 veh/hr.

COMPARISON BETWEEN TWO- AND FOUR-WAY STOP CONTROL

In order to obtain practical warrants that can be implemented by practising highway traffic engineers, four-way stop priority control was compared with two-way stop control under the same traffic and geometric conditions. Generally, at high flow levels, two-way stop control tends to cause higher average delays to those vehicles that are delayed on the minor road than four-way stop control, especially when the two intersecting roads have approximately equal traffic flows. Under these conditions, a four-way stop priority rule provides a better distribution of delays between the two intersecting roads. For this reason, when based on average delay, the practical capacity of the intersection (on the basis of observed U.K. gap acceptance) obtained from applying a four-way stop priority rule (1,650 veh/hr), as shown on Figure 3, was higher than that obtained from using a two-way stop priority rule (1,400 veh/hr), as shown on Figure 2, for the same traffic and geometric conditions.

A series of simulation runs was carried out with left-, straight-ahead and right-turning movements (U.S. rules of the road) in the proportions of 15, 70, and 15 percent, respectively, for two- and four-way stop control with single-lane approaches. For each simulation run, the flows on the four approaches

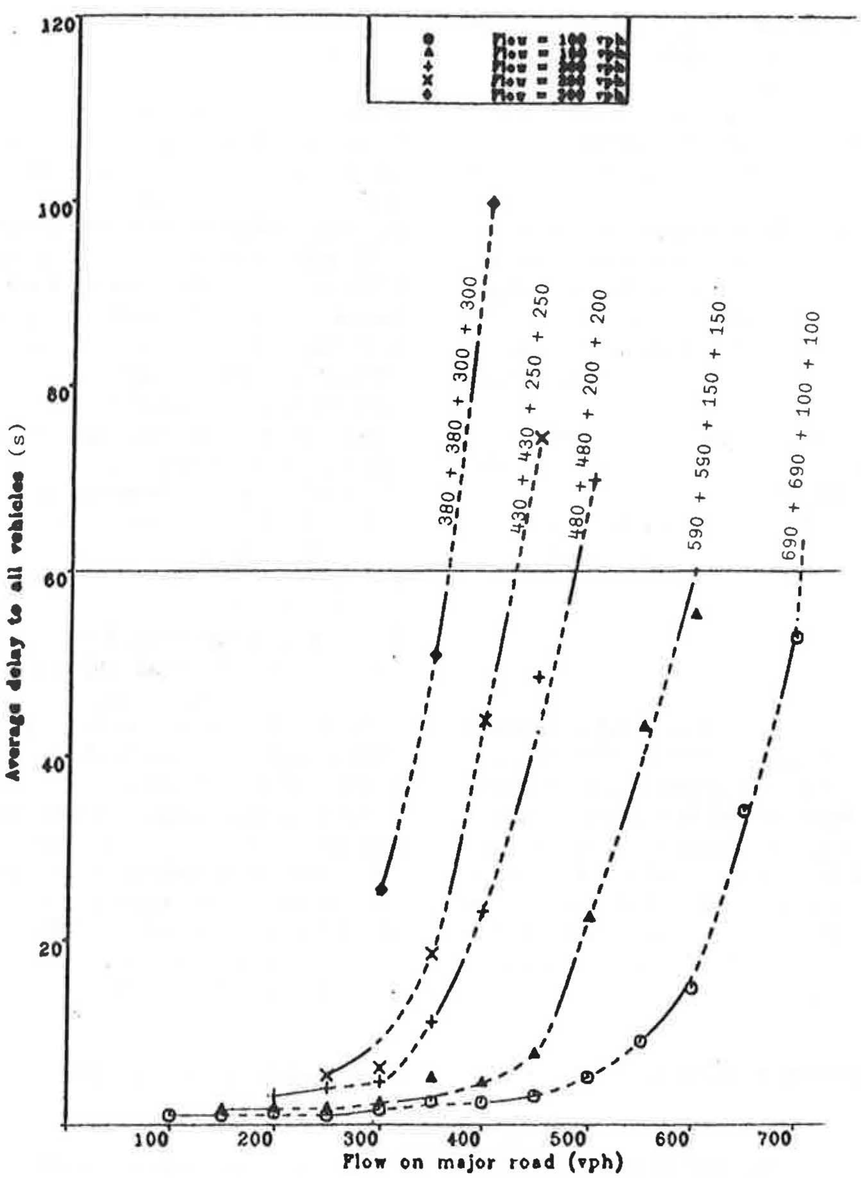


FIGURE 2 Flow-delay relationships at two-way stop priority highway intersections for one-lane approaches with flows for each approach (minor road flow, 100 to 300 vph).

were equal. Values of average delay obtained from the simulations are presented in Table 1.

These values indicate, for the flow conditions simulated, that at low flow levels there is no significant difference between the two control methods, whereas, at flow levels exceeding 250 veh/hr per approach, the use of a four-way stop priority rule yields lower delay values than the use of a two-way stop priority rule. This conclusion is justified in terms of both delay distributions and overall average intersection delay.

EFFECT OF INTERSECTION GEOMETRY

The geometrical layout of the intersection area has a considerable influence on intersection performance. Generally, the adoption of better design standards will be reflected in enhanced intersection performance. However, the strategy adopted in the design of a given intersection is a function not only of road users' convenience, but also of the space available, and the resources allocated for that intersection. The effect of the number of lanes on the intersection approaches is described under the two methods of control. The purpose of the analysis was to establish some guidelines regarding the adoption of a suitable layout at highway intersections. The number of queuing lanes provided on the controlled approaches to the intersection area is one of the factors that assist in providing improved performance. This improvement is achieved through the reduction of queuing delay on these approaches by increasing the service channels available, and hence increasing the opportunities for vehicle release.

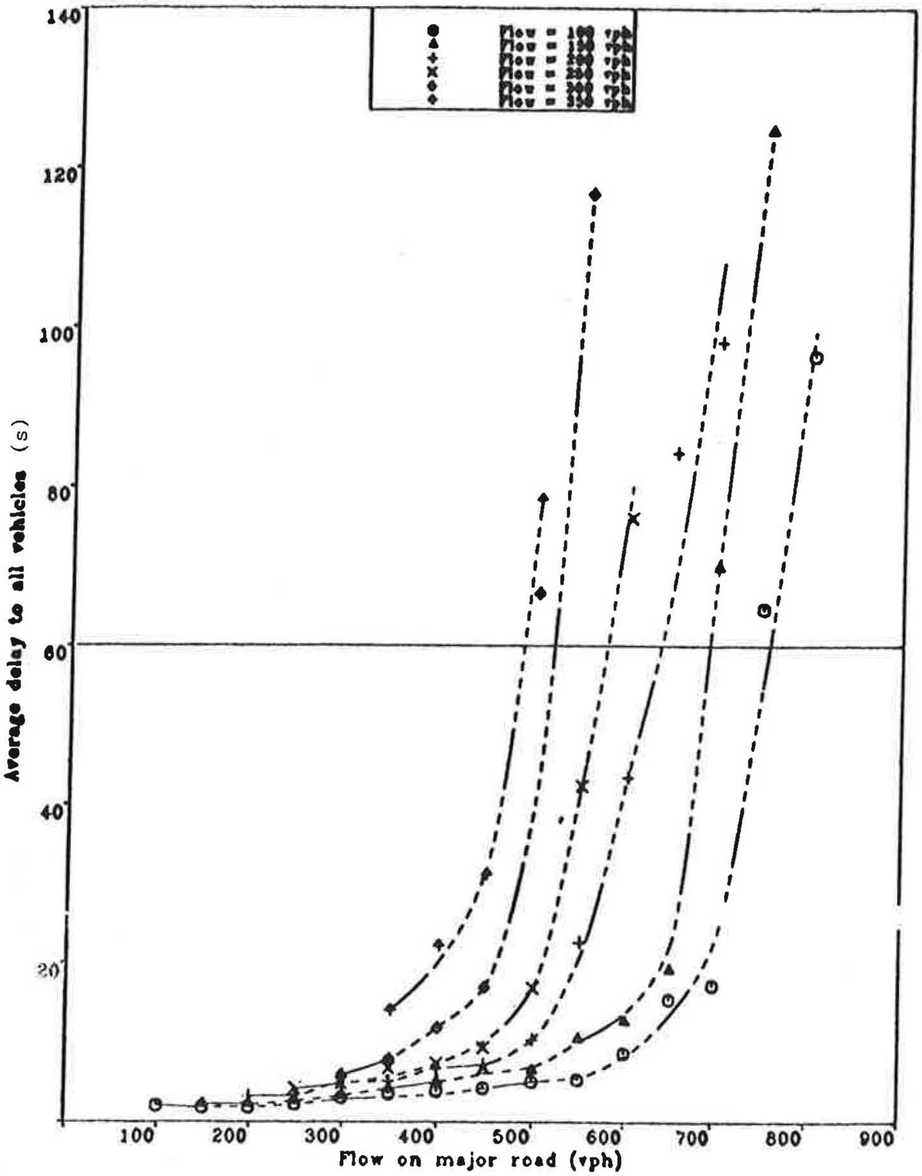


FIGURE 3 Flow-delay relationships at four-way stop priority highway intersections for one-lane approaches with flows for each approach (minor road flow, 100 to 350 vph).

In order to study the effect of this factor on highway intersections controlled by a two-way stop priority rule, the simulation model TWSSIM was used to generate a range of delay values for two different layouts. In the first, a one-lane major road intersects a one-lane minor road, whereas in the second, each of the two intersecting roads has a two-lane approach. In order to use these results to obtain guidelines for the provision of one or two approach lanes at intersections controlled by a two-way stop priority rule, certain criteria were defined. Using a 60-sec average delay as this criterion, average delay contours were derived as shown in Figure 4. This figure can be accessed with a combination of flow levels on the major and minor roads to select a suitable layout using delay criteria.

Similarly, the simulation model FWSSIM was used to study the effect of the number of approach lanes on the performance of intersections controlled by a four-way stop priority rule. Two cases were again considered in the analysis, one- and two-lane approaches. Using the same average delay criterion of 60 sec, guidelines were derived for the selection of the appropriate layout for intersections controlled by four-way stop intersections. The 60-sec average delay contours for the two layouts are shown in Figure 5 and can be used to choose the appropriate layout for a given combination of major- and minor-road flow levels.

CONCLUSIONS

The main conclusions that were drawn from the simulation study of two- and four-way stop control are as follows:

TABLE 1 COMPARISONS OF AVERAGE DELAY AT AN INTERSECTION OPERATING UNDER TWO- AND FOUR-WAY STOP CONTROLS

Flow on each approach (vph/approach)	Average delay to all vehicles (s)	
	two-way stop	four-way stop
100	1.1	2.0
150	1.6	2.2
200	3.1	3.2
250	5.3	4.2
300	25.5	5.9
350	*	14.0
400	*	30.3

* long queues on the minor road approaches

1. Field observations at the validation site indicated that the shifted negative exponential represented the appropriate headway distribution for both stop and nonstop conditions.

2. The analysis of lag-gap acceptance observations indicated that there was a marked difference between acceptance distributions for the different turning movements at the given intersection approach. Average acceptable lag values on the minor-road approaches of the validation site controlled by a two-way stop priority rule were found to be 5.5, 5.5, and 4.5 sec for left-turning, straight-ahead, and right-turning movements (U.S. rules of the road), respectively. On the other hand, average acceptable gap values for the three directions of movements on these approaches were 6.0, 4.85, and 4.75 sec, respectively.

3. A detailed analysis of lag-gap acceptance data was carried out where lags were split into two categories, delayed and undelayed, according to the queuing condition at the time of arrival of the vehicle. However, marginal differences were found between observed delayed and undelayed lag acceptance distributions for the different turning movements.

4. Theoretical distributions were fitted to the observed acceptance distributions in the form of normal, log-normal, and linear functions. The appropriate distribution was then se-

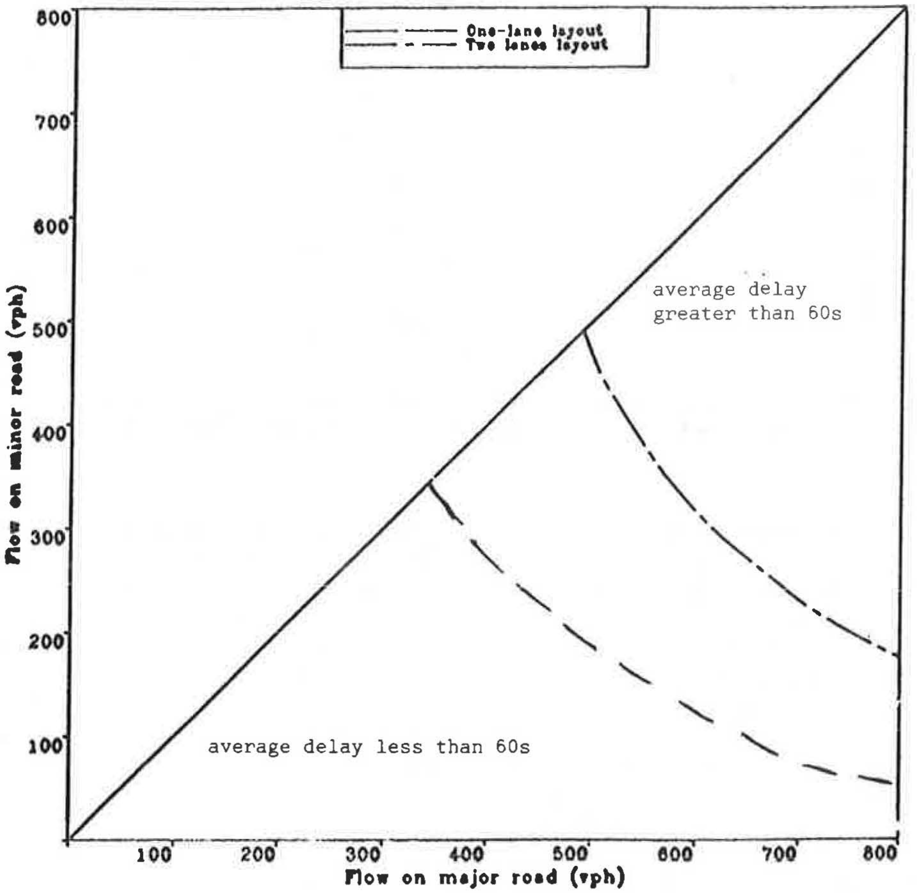


FIGURE 4 Average delay contours at two-way stop priority intersections with different geometric layouts for all one-way flows.

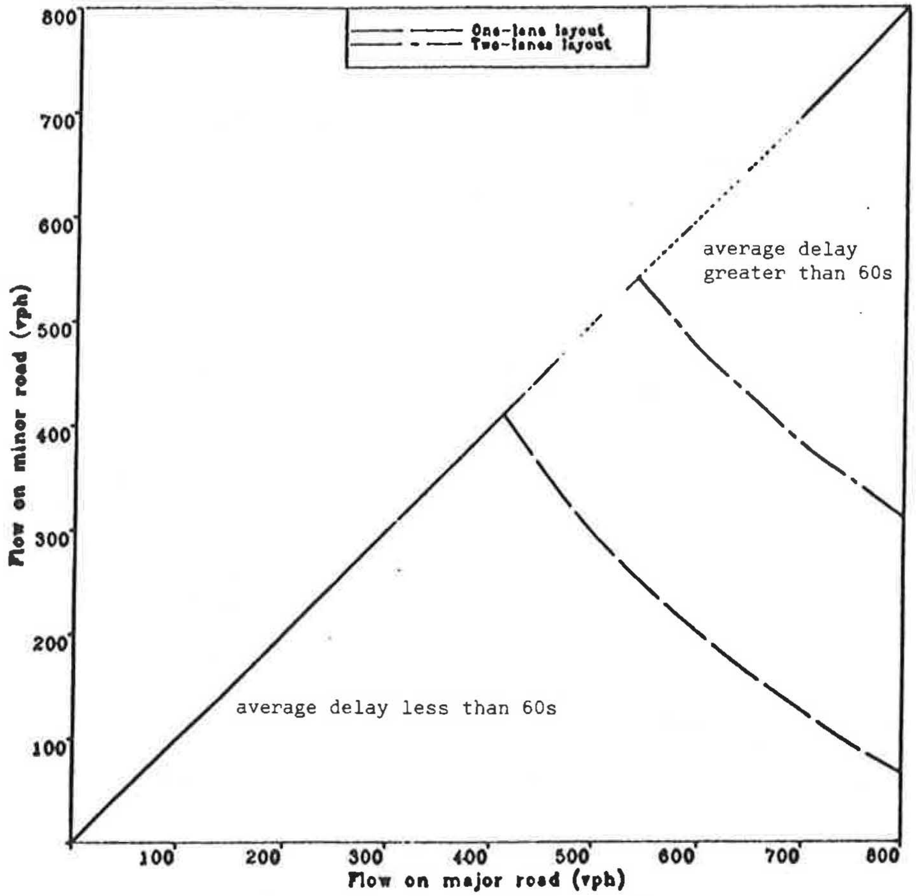


FIGURE 5 Average delay contours at four-way stop priority highway intersections with different geometric layouts for all one-way flows.

lected for each combination of intersection approach and lag-gap category and used in the validation of the developed simulation models. Simulated queue lengths and delay values obtained from the models were compared to the observed values, and good agreement was found.

5. At two-way stop priority highway intersections, the average delay per vehicle increased nonlinearly with the increase in flow levels. Using a selected delay criterion of 60 sec average delay to all vehicles, the practical capacity of the intersection with one-lane approaches occurred when the total inflow was approximately 1,400 veh/hr.

6. The effect on delay of using the detailed lag-gap acceptance categories, delayed lags, undelayed lags, and gaps, was found to be marginal. It is believed that this marginal effect is caused by the relatively small difference between the two lag categories observed at the validation site.

7. At four-way stop priority highway intersections, the average delay per vehicle was found to increase with the increase in flow levels in a nonlinear form. The practical capacity of one-lane approach intersection controlled by this priority rule was determined according to the same delay criterion of 60 sec average delay to all vehicles and occurred when the total inflow was approximately 1,650 veh/hr.

8. At equal flow levels and for the turning movements input on the major and minor roads, four-way stop control provided a more even distribution of delay than two-way stop control, and reduced queue lengths on the minor-road approaches. The overall intersection delay experienced under these conditions was found to be lower in the case of the four-way stop than the two-way stop for flows exceeding 250 veh/hr on each approach. Therefore, it is concluded that for equal flows and specified turning movements on each approach above this critical level, four-way stop rule provides a better control method than two-way stop control.

REFERENCES

1. J. E. Upchurch. Guidelines for Use of Sign Control at Intersections to Reduce Energy Consumption. *Journal of the Institute of Transportation Engineers*, Vol. 53, No. 1, Jan. 1983, pp. 22–24 and 31–34.
2. C. E. Lee and V. S. Savur. Analysis of Intersection Capacity and Level of Service by Simulation. In *Transportation Research Record* TRR 699, 1979, pp. 34–41.
3. E. A. Ismail. Highway Intersections with Alternative Priority Rules. Ph.D. dissertation, Bradford University, Bradford, Great Britain, 1989.

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

The Two-Capacity Phenomenon: Some Theoretical Issues

JAMES H. BANKS

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

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

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

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

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

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

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

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

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

OBSERVED PROCESS OF FLOW BREAKDOWN

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

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

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

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

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

THEORETICAL BACKGROUND

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

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

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

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

THE FLOW DECREASE

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

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

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

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

SPEED DIFFERENCES BETWEEN VEHICLES

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

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

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

$$t_B = t_A + \Delta x(\Lambda_2 - \Lambda_1) \qquad (1)$$

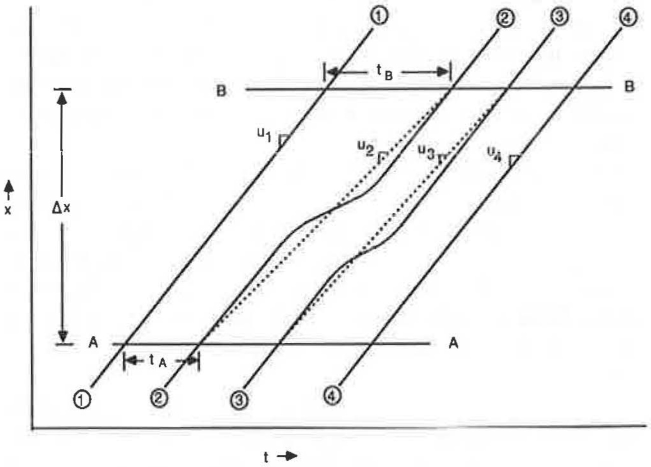


FIGURE 1 Hypothetical vehicle trajectories.

from which

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

or

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

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

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

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

or

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

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

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

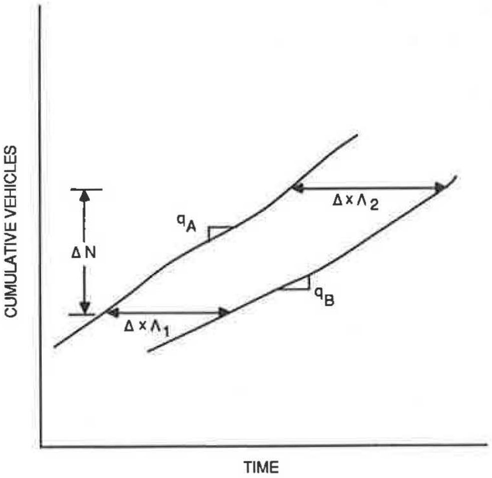


FIGURE 2 Hypothetical cumulative arrival curves.

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

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

INCREASE IN PASSAGE TIME

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

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

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

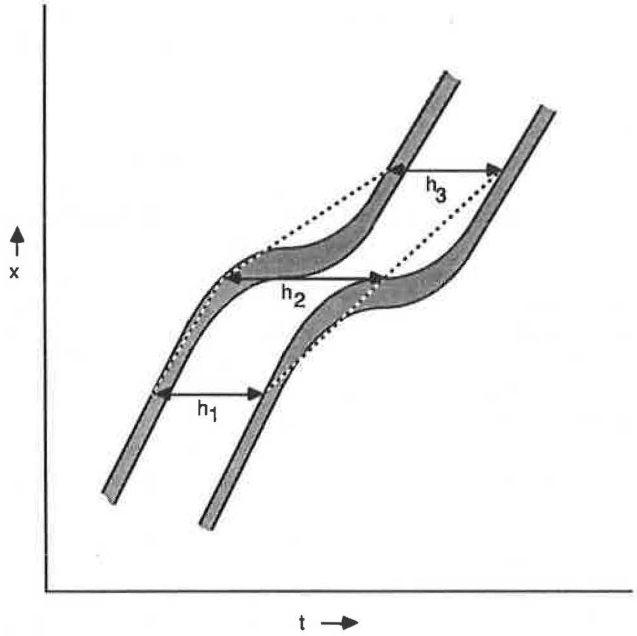


FIGURE 3 Effect of passage time on time headway.

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

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

The actual calculation of the average gap then is

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

where

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

But because

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

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

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

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

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

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

SHOCK WAVE STRUCTURE

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

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

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

where

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

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

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

SPEED INSTABILITY AND CAR-FOLLOWING

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

CONCLUSION

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

ACKNOWLEDGMENTS

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

REFERENCES

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

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

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

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

Empirical Analysis of Congested Traffic Flow Characteristics and Free Speed Affected by Geometric Factors on an Intercity Expressway

MASATO IWASAKI

In Japan, much lower traffic capacities occur at S-curves, sag sections, and the entrances of long tunnels on intercity expressways. The traffic capacities at such sections and points were estimated at about 60 to 70 percent that of the basic expressway segments. Oscillating characteristics of traffic flows under congested conditions on urban expressways are already known. However, the traffic flow characteristics on the urban expressways have yet to be analyzed. The traffic flow characteristics under congested conditions, especially the accordion-like actions, the maximum flow rate at bottlenecks, and the relationship between the speed characteristics and geometric factors are analyzed. The data used were collected from 106 detector stations set up in each lane on a 220-km section of the Tomei Expressway. Forty-eight stations were in the upstream direction and 58 stations were in the downstream direction. Three severe bottlenecks on the Tomei Expressway are described. Propagating speeds of the density boundaries between the free and congested-flow regions were estimated. The characteristics of the accordion-like actions caused during the congested conditions on the intercity expressway were almost the same as those of urban expressways. The free speeds in each lane at the detector stations were estimated in the $Q-V_s$ plane. Free speeds were affected by the geometric factors at the detector stations. Multivariate analysis was introduced to quantify the effects of the geometric factors on the free speeds.

In Japan, the first week of May usually has 4 or 5 days of holidays. Many people travel to various places using passenger cars. Therefore, heavy traffic is experienced on the expressways. As a result of the heavy traffic demand, severe congestion occurs at the bottlenecks on intercity expressways.

A much lower traffic capacity exists at S-curves, at sag sections, and at entrances of long tunnels on the intercity expressways. Some researchers have reported that the traffic capacities at these sections and points were from about 2,600 to 2,800 veh/hr per two lanes under congested flow conditions (*1*). These capacities were estimated to be from about 60 to 70 percent that of the capacity at the basic expressway segment under free flow conditions.

However, the effects of these geometric factors on traffic flows have yet to be analyzed. So the geometric factors described earlier affect the capacities and traffic phenomena in the congested flow appearing in the upstream road section at the bottlenecks. The speed characteristics of the traffic flow—for example, the free speed and critical speed represented on the $Q-K-V_s$ plane—may also be affected by the geometric factors.

Factors affecting traffic capacity were described in the American *Highway Capacity Manual* (*2,3*) and in the Japanese *Highway Capacity* (*4*) literature. In Japan, some factors affecting capacity on expressways were quantified. However, some of the geometric factors affecting capacity and running speed on expressways are still unstudied.

Oscillating characteristics of traffic flow under congested conditions on the urban expressways have already been clarified (*5–7*). The capacity analyses and empirical studies about $Q-K(\text{Occ})-V_s$ relationships on urban and intercity expressways have also been analyzed by many researchers and organizations (*3,8–10*). However, in this work few analyses concerning the geometric factors have been carried out, so the relationships between traffic characteristics and geometric factors still need to be quantified.

Koshi and others (*11–14*) have conducted many car-following experiments using test cars on intercity expressways. They have analyzed the car-following behavior at S-curves, sag sections, and at the entrances of long tunnels. The results of these experiments have provided important basic information about the factors that affect the capacity at these bottlenecks.

The traffic characteristics under congested conditions and the relationships between the speed characteristics and geometric factors should be important for the planning and design of expressways and expressway traffic surveillance and control systems.

Basic traffic flow characteristics near capacity and under actual congested condition, especially the capacities at bottlenecks and the accordion-like actions occurring upstream at the bottlenecks, are analyzed. Another objective is to quantify the effect of the geometric factors on free speeds at each observation station and lane on the intercity expressways using traffic detector data.

DATA ACQUISITION

The traffic data used in this study come from the Expressway Surveillance and Control System of the Tokyo First Bureau of Traffic Management of Japan Highway Public Corporation. The system comprised 106 mainline detector stations with

Department of Civil Engineering, Musashi Institute of Technology, 1-28-1 Tamazutsumi, Setagaya-ku, Tokyo 158, Japan.

double-induction-loop traffic detectors (5.5 m distance between loops) on each lane, both in the up- and downstream directions, on a 220-km section of the Tomei Expressway; off-ramp detector stations with ultrasonic detectors (count of volume) at each interchange; and closed-circuit television surveillance cameras in the long tunnels operated from a control center. The accuracy of the double-loop detector is about 95 to 97 percent in volume count. The data analyzed here come from only these double-induction-loop detectors set up at the main line of the expressway.

Forty-eight detector stations are in the upstream direction, and 58 stations are in the downstream direction. The 5-min volume (Q), volume of commercial vehicles, space mean speed (V_s), and time occupancy (Occ) from each detector station were recorded on magnetic tape. A traffic flow record consisting of 10 days during May 1989 was used for the analysis. Some detectors broke down and data from these detectors were eliminated from the analysis. The volume of commercial vehicles were not converted to passenger car units (pcu) in this stage of the analysis. Only percentages of commercial vehicles were calculated both in the up- and downstream directions at each detector station.

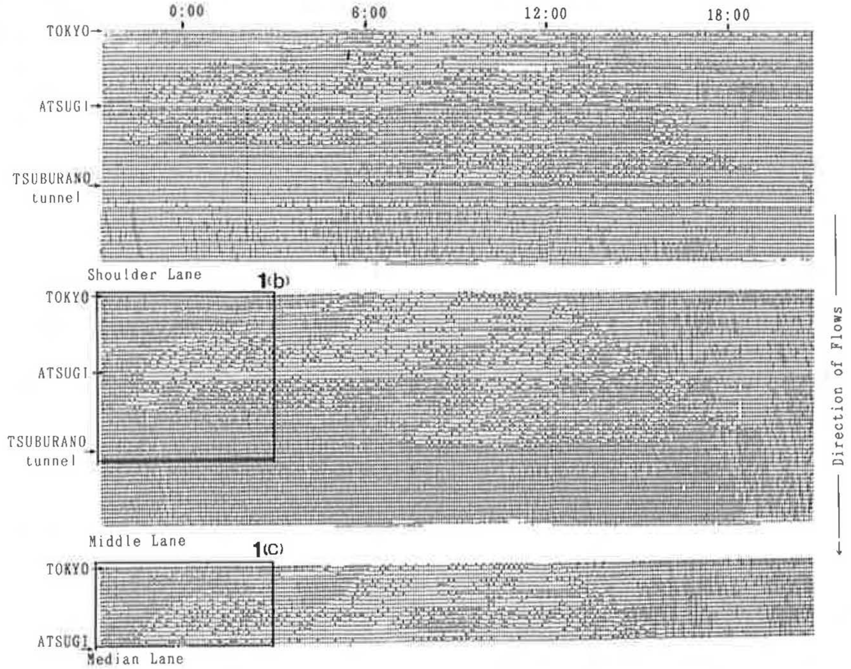
Geometric factor data were also collected with cooperation of the Japan Highway Public Corporation. The geometric factors collected were the gradients and radii at the detector stations, and data concerned with the S-curve, sag, and crest existing within 1.5 km up- and downstream at the detector stations. The 1.5-km distance was equivalent to the running distance when a vehicle runs for about 1 min at free speed.

CHARACTERISTICS OF TRAFFIC FLOW

The heavy holiday traffic occurring in the Tokyo metropolitan area created severe congested traffic conditions on the downward direction of the Tomei Expressway. More than a 70-km slow-moving vehicular line formed during the early morning hours on May 3 in the downward direction of the Tomei Expressway. The head of the slow-moving vehicular line was observed at the entrance of the Tsuburano Tunnel. The percentage of commercial vehicles was from about 20 to 55 percent through 10 days both in the up- and downstream directions of the Tomei Expressway.

Shock-Wave Characteristics near Capacity and Congested Region, and Traffic Flows under Congested Condition

Figure 1 shows congested traffic flow conditions for a 220-km section in the downstream direction on the Tomei Expressway. In this figure, the dots mean values that represent the product of Q and V_s. The traffic conditions are classified by means of the darkness of the dots on the time-space plane (*15*). The darker parts mean higher flow rate and higher speed, and the lighter parts mean lower flow rate and lower speed, indicating that severe congestion has occurred. In the downstream direction of the Tomei Expressway, it is specified that three major bottlenecks exist on the expressway. The first bottleneck is the Atsugi Interchange from where the three-



* Darker dots : Higher flow rate, higher speed
* Lighter dots : Lower flow rate, lower speed

FIGURE 1 Traffic congestion on each lane of the Tomei Expressway (May 1989, downstream direction).

lane roadway drops two lanes (about 35 km from the Tokyo entrance). The second one is the compound geometric section with an S-curve and a sag near Hatano (about 46 km from the Tokyo entrance). The last one is the entrance of Tsuburano Tunnel (about 1.7 km long) at about 68 km from the Tokyo entrance.

Traffic characteristics under congested conditions on urban expressways have been analyzed by the author (*7*) and others (*5,6*). It has been pointed out that traffic flow under congested conditions has some specific characteristics such as the following (*15*):

1. Traffic flow oscillates under congested conditions.
2. Oscillations in traffic flow are not uniform over the congested road section. There are small amplitudes immediately upstream from the bottlenecks. As they propagate upstream, the ripples grow larger, but the wave numbers become fewer in number.
3. The oscillations of the two neighboring lanes become more synchronous as they propagate in the upstream direction.
4. The propagating speeds of waves are almost 18 to 20 km/hr.
5. There are some upper speed limits of the order of the high-speed parts. The speed is thought to be about 45 km/hr on the urban expressways.
6. No specific period prevails in the oscillating characteristics of the traffic flow under congested traffic flow condition.

In Figures 1–3, the fluctuations of traffic flow in the congested regions on the three lanes are synchronized and propagate in the upstream direction at the speed of 17 to 24 km/hr, which is almost the same as the wave speed appearing on the urban expressways (*15*). The oscillating characteristics of congested flow condition, i.e., propagating speed, and periodicity of fluctuations of short-term traffic flow and average speed, have been analyzed by some researchers (*5,6,15*).

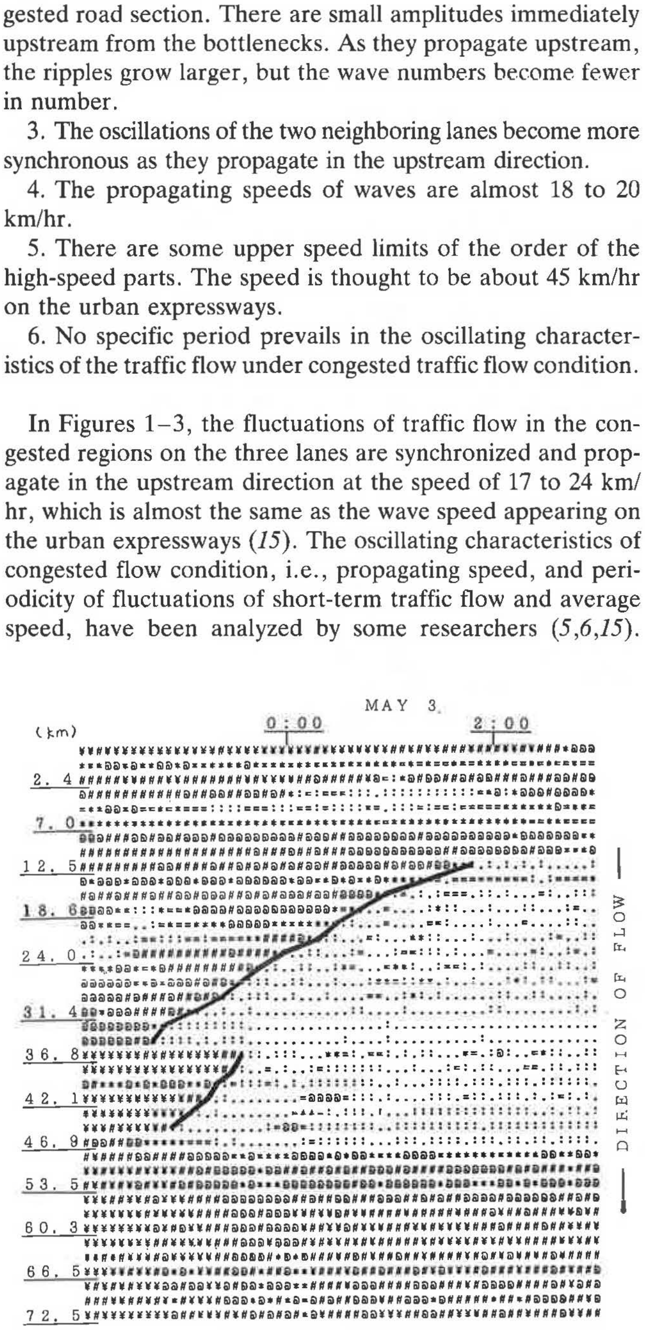


FIGURE 2 Shock wave propagation on middle lane.

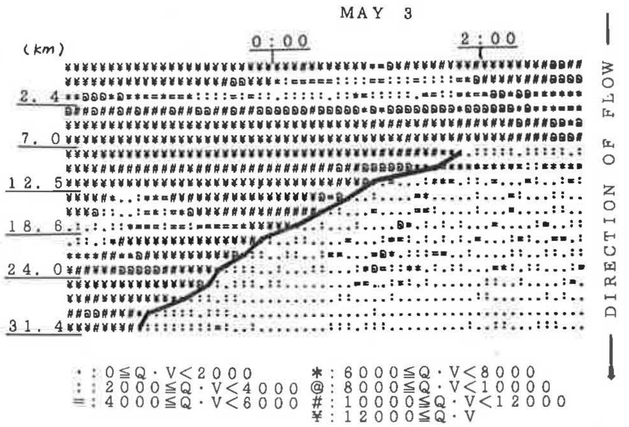


FIGURE 3 Shock wave propagation on median lane.

However, few researchers have dealt with the density boundary characteristics of free and congested regions using field data or observations. In Figures 2 and 3, the propagation of density boundaries between free and congested regions can discriminate. The discrimination threshold used for congested to free flow regions was not flow rate, but the QV value. The QV value, 6,000 veh-km/hr per 5-min interval (* in the figures), is introduced as the threshold. This value was decided on to represent the traffic condition that is 100 veh per 5 min with mean speed for 5 min of 60 km/hr. The average speed of these density boundaries can be estimated to be about 7 to 13 km/hr; it is lower than the propagating speed of traffic flow fluctuations under congested conditions.

The difference in the propagating speed, density boundaries of free and congested regions, and fluctuations of traffic flow in the congested regions may be explainable using the fundamentals of traffic shock wave motion (*15*). In this case, the density boundaries cause the transient conditions from the free to congested transition regions. At the traffic condition just before traffic congestion, the flow rate in the free flow region may be expected to be slightly higher than the bottleneck capacity (Q_{maxI}) (*1*). On the other hand, the flow rate immediately after changing traffic condition (from the free to congested regions) still stays at the higher level (q_{maxI}), but slightly lower than capacity (Q_{maxI}). Then, shock waves are formed by this change in condition of traffic flow (flow rate and density), and propagate with speed C_b shown in Figure 4. On the other hand, the propagating speed C_a of the traffic fluctuations in the congested region is caused only in the congested flow region, and the changes in the flow rates and densities may be larger than the changes caused at the boundary shown in Figure 4. Therefore, the shock wave speed (C_b) may be expected to be lower than propagating speeds (C_a) of the traffic fluctuations in the congested regions.

Volume-Occupancy Relationships

As described, three major bottlenecks exist in the downstream direction on the Tomei Expressway, and the entrance of Tsuburano Tunnel has been already pointed out to be one of the severest-capacity restrictions on the Tomei Expressway as a result of many analyses (*1*). Figure 1 also shows that the entrance

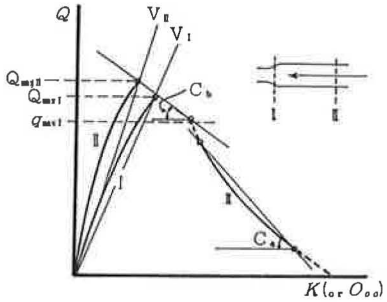


FIGURE 4 Conceptual diagram of shock wave speeds.

of Tsuburano Tunnel is the head of the traffic congestion on the downstream direction of the Tomei Expressway.

Figures 5–8 show the volume-occupancy relationships (*17*) on the shoulder lane immediately downstream of the Tsuburano Tunnel (1.0 km downstream from the exit), immediately upstream of the Tsuburano Tunnel (0.2 km upstream from the entrance), and two points upstream from the entrance, 1.3 and 3.3 km, respectively.

In Figure 3, which shows the volume-occupancy relationship immediately downstream of the exit of Tsuburano Tunnel, the appearance of no congested region contrasts in a striking way with the relationships at the detector stations

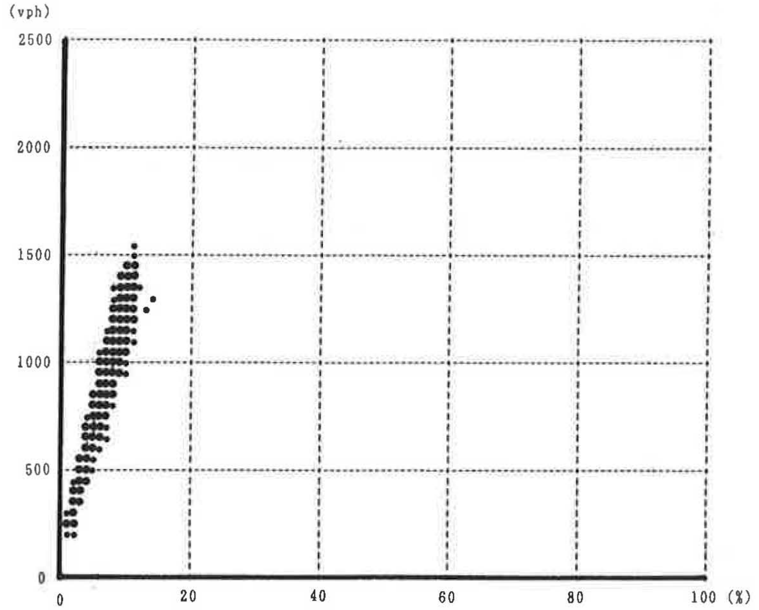


FIGURE 5 Flow rate (5-min) versus time of occupancy at 1.0 km downstream of the exit of Tsuburano Tunnel (shoulder lane).

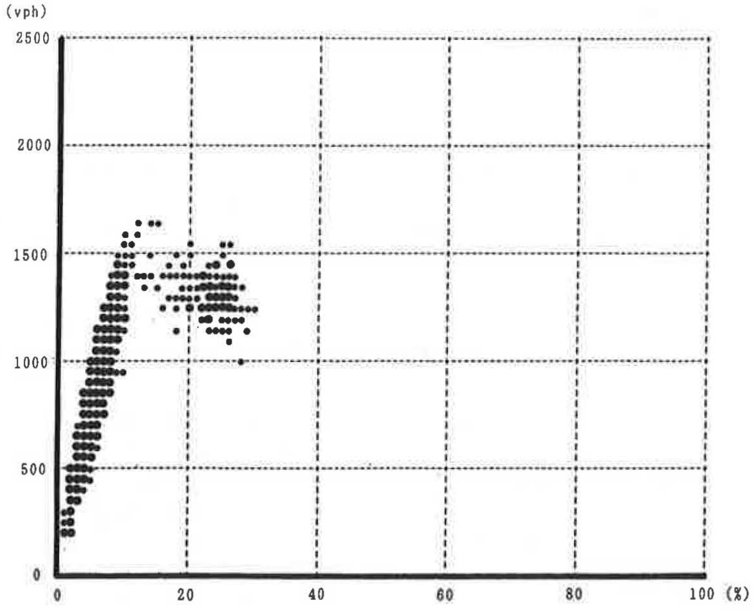


FIGURE 6 Flow rate (5-min) versus time of occupancy at 0.2 km upstream of the entrance of Tsuburano Tunnel (shoulder lane).

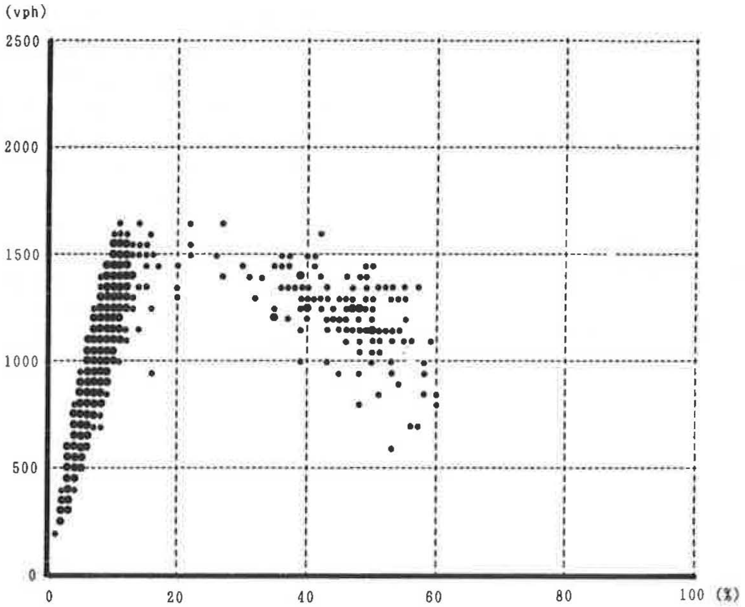


FIGURE 7 Flow rate (5-min) versus time of occupancy at 1.3 km upstream of the entrance of Tsuburano Tunnel (shoulder lane).

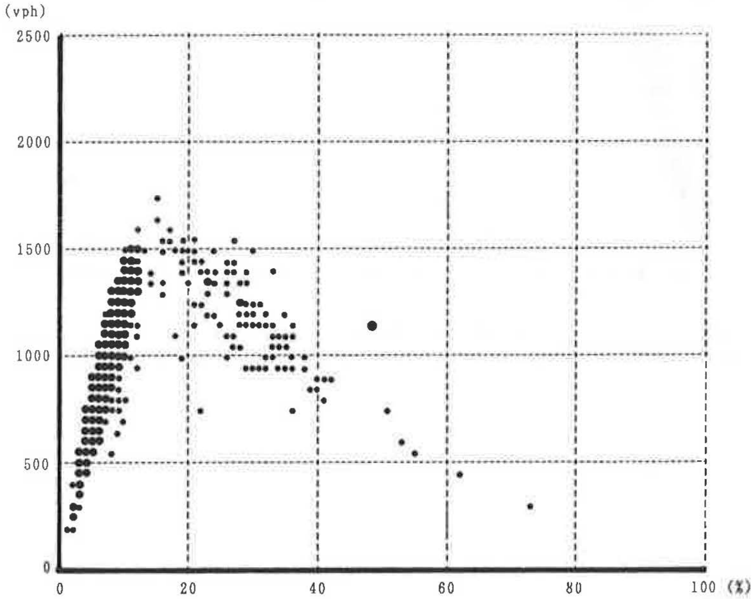


FIGURE 8 Flow rate (5-min) versus time of occupancy at 3.3 km upstream of the entrance of Tsuburano Tunnel (shoulder lane).

upstream of the entrance of the tunnel (Figures 6–8). During restriction of traffic capacity in the long tunnel, the traffic flow that passes through the bottleneck cannot exceed the maximum flow rate of the bottleneck. Therefore, the maximum flow rate appearing in Figure 3 reflects the capacity of the shoulder lane at the entrance of Tsuburano Tunnel.

In Figure 6, immediately upstream of the entrance of the tunnel, discontinuity and discrepancy between free and congested regions are apparent. The maximum flow rates in the free flow region are estimated to be about 1,650 veh/hr per lane on the shoulder lane. The maximum flow rates in the congested region are estimated to be 1,550 veh/hr per lane on the shoulder lane. From Figures 5 and 6, the capacity on the shoulder lane at the Tsuburano Tunnel is estimated to be about 1,550 veh/hr per lane.

Growth of amplitude of traffic fluctuations in congested regions is recognized from Figures 6–8. The tendency of characteristics of growing amplitude in traffic flow fluctuations is the same as on urban expressways, but the upper limit of speeds in the congested-flow fluctuation are significantly larger than that of urban expressways. This difference may be caused mainly by the differences in the design speed, roadway conditions, and regulatory speeds existing between urban and intercity expressways. The regulatory speeds of almost all urban expressways in Japan are 60 km/hr and 100 km/hr for intercity expressways.

ANALYSIS OF THE RELATIONSHIPS BETWEEN FREE SPEEDS AND GEOMETRIC FACTORS

Characteristics of Free Speeds

Free speeds at each lane of the detector stations are estimated using the $Q-V_s$ relationship. Figures 9 and 10 show examples of the $Q-V_s$ relationship in the free flow regions. In these figures, a speed of 60 km/hr is used as the boundary speed between the free and congested flow regions. All the Q and V_s data within the confines of the speed category greater than 60 km/hr are applied to the linear regression. The free speeds are estimated from the intercepts of the Y axis by linear regressions. The free speed is defined when traffic volume is zero. Therefore, the factors that affect the free speeds ought to be the geometric factors at the points and percentages of commercial vehicles.

Figures 11 and 12 show the distributions of the estimated free speed on each lane in the two- and three-lane roadway sections. Differences in mean free speeds clearly exist between the lanes. Figure 13 shows the free speeds at each

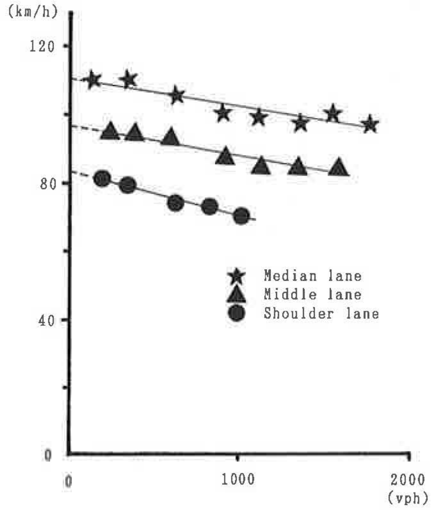


FIGURE 9 $Q-V_s$ **diagram in the free flow region (a three-lane segment).**

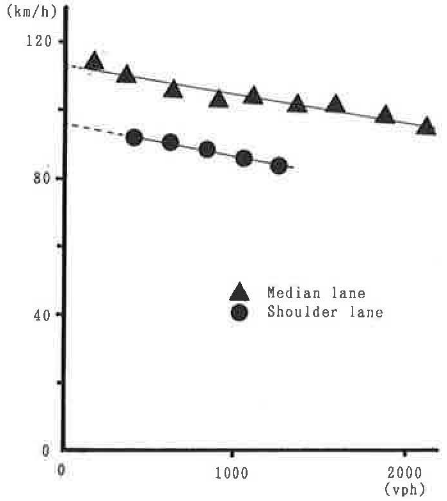


FIGURE 10 $Q-V_s$ **diagram in the free flow region (a two-lane segment).**

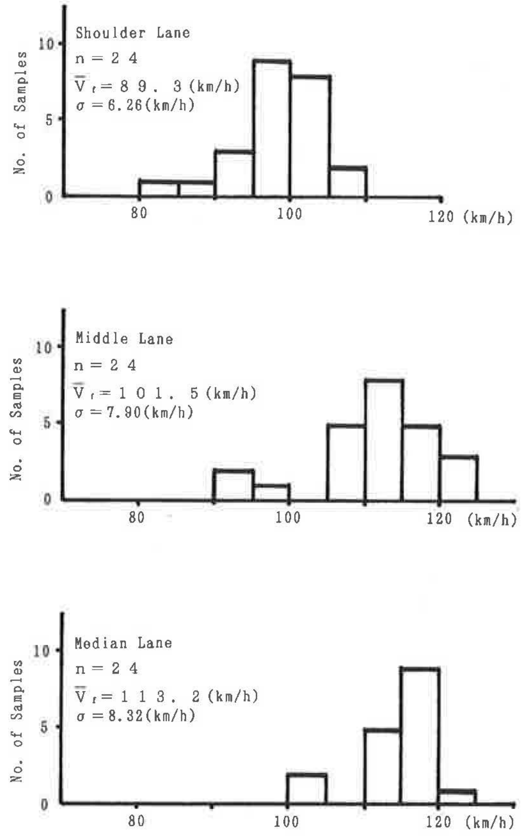


FIGURE 11 Distribution of free speeds in each lane for a three-lane segment in the downstream direction.

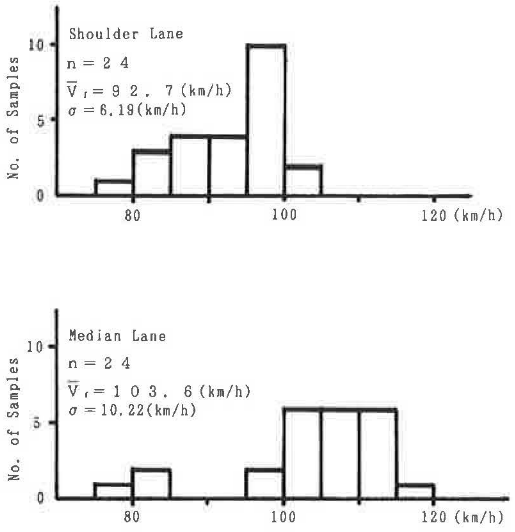


FIGURE 12 Distribution of free speeds in each lane for a two-lane segment in the downstream direction.

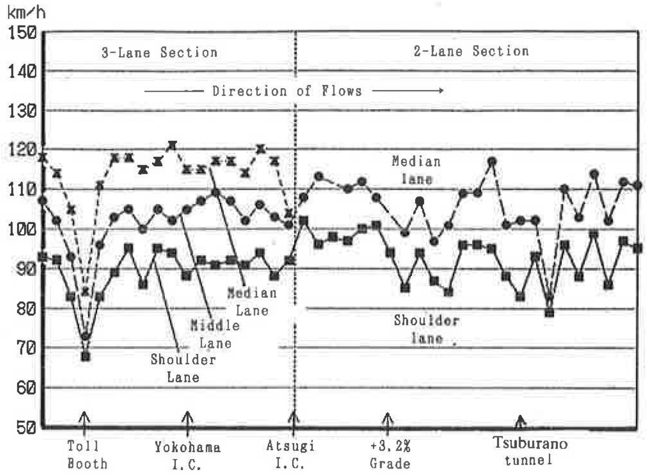


FIGURE 13 Fluctuations of free speed on each lane at detector stations in the downstream direction.

detector station. Significant differences also exist between three- and two-lane roadway sections. From these differences, it would be better to deal with the free speeds on the three- and two-lane roadway sections separately.

As described earlier, three major bottlenecks exist in the downstream direction on the Tomei Expressway. The free speeds at these bottlenecks are lower than the free speeds that appear on the detectors up- and downstream of the bottlenecks. In addition to these bottlenecks, there are some detector stations that have lower free speeds than detector stations on each side of it. Such drops in free speeds seem to depend on geometric factors (for example, upgrades and radii) at the stations.

Figures 14 and 15 show examples of the relationships between the free speeds and typical geometric factors (gradient and curvature radius at the detector stations). According to these figures, the free speeds are obviously affected by the gradients and radii. For example, the steeper the upgrades, the lower are the free speeds. Some of other geometric factors also have an influence on free speeds.

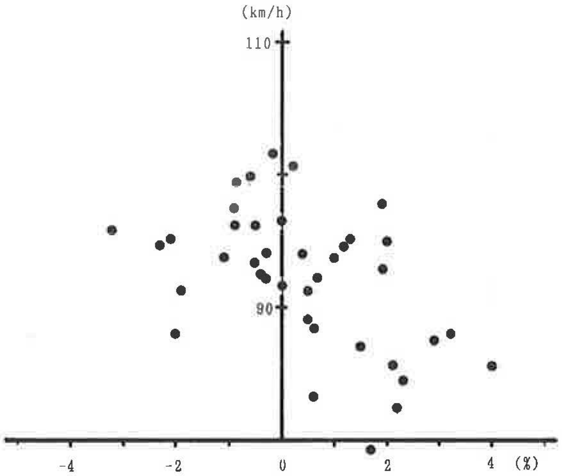


FIGURE 14 Free speed versus gradient in downstream direction on the Tomei Expressway (shoulder lane).

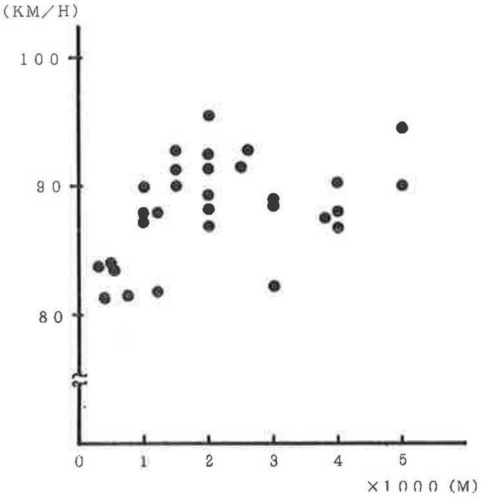


FIGURE 15 Free speed versus curvature radius in downstream direction on the Tomei Expressway (shoulder lane).

Analysis of the Relationships between Free Speed and Geometric Factors using Multivariate Analysis

In the preceding section, it was found that the free speeds are affected by some geometric factors. Next, a model is constructed to explain how the geometric factors and percentage of commercial vehicles affect the free speeds at each detector station on the Tomei Expressway.

The factors used are as follows:

1. The geometric factors at the detector stations.
   a. Radii, and
   b. Gradients.
2. Geometric factors that exist within 1.5 km in front of and behind the given detector stations.
   a. Sum of two radii that form the S-curve,
   b. Distance to the inflection point of the S-curve,
   c. Distance to the bottom of the sag,
   d. Sum of the absolute values of the gradients forming the sag,
   e. Distance to the summit of the crest, and
   f. Sum of the absolute values of the gradients forming the crest.
3. Percentage of commercial vehicles at the detector stations.

As mentioned earlier, the entrance of a long tunnel should be considered as a severe bottleneck to traffic. However, only one long tunnel, the Tsuburano Tunnel, exists in this study section. Therefore, the factors about long tunnels have been eliminated from this analysis.

Several of the geometric factors described do not exist within 1.5 km in front of and behind the detector stations. So, the theory of quantification I, a kind of multivariate analysis method developed by Hayashi (*18*), which corresponds to the multiple linear regression method, is introduced here. Table 1 presents one of the results of the relationship between the free speed and geometric factors. Each factor was categorized, and the degree of influence was calculated as the category weight presented in Table 1. The magnitudes of the range values presented in Table 1 correspond to the partial coefficients of correlation calculated by the multiple linear regression method. The results presented in Table 1 mean that the free speeds on the expressway are significantly influenced by the gradients, radii at the detector stations, and sum of the two radii that form the S-curve existing within 1.5 km downstream of the detector stations.

Figure 16 shows the relationship between the free speeds that were estimated by the model and free speeds estimated by the linear regression described previously. According to this result, the multiple coefficient of correlation was 0.847. Similar results were obtained from other data sets.

The percentages of commercial vehicles are not introduced as significant factors into the multivariate equation. The reasons may be that the classification of percentages are too large (two classes are introduced here, less than 30 percent, and greater than or equal to 30 percent) to be introduced as an influencing factor, and that the percentages of commercial vehicles do not vary between the detector stations.

TABLE 1 CATEGORY WEIGHTS AND RANGES

FACTORS	CATEGORY	NUMBER OF SAMPLE	CATEGORY WEIGHT	RANGE
GRADIENT (%)	i<-3	1	4.524	9.804
	-3≦i<0	8	3.105	
	i=0	1	0.774	
	0<i≦3	13	-1.912	
	i>3	1	-5.280	
RADIUS (m)	R≦1500	9	-3.731	6.867
	1500<R ≦3000	7	1.215	
	R>3000	8	3.135	
SUM OF RADII (m)	R ≦1500	5	-4.407	6.694
	1500<R ≦3000	6	2.287	
	R>3000	7	1.891	
	NON	6	-0.820	

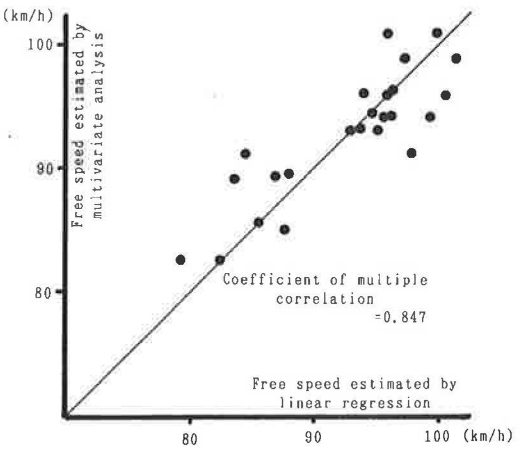


FIGURE 16 Free speed estimated by the model versus free speed estimated by linear regression method for a two-lane segment in the downstream direction (shoulder lane).

The steeper the upgrades (>3 percent) and the smaller the radii (<1,500 m) at the detector stations, the lower are the free speeds. In addition to these results, the two radii that form the S-curve, if less than 1,500 m, and existing within 1.5 km downstream of the stations, adversely affect the free speed at the stations (which is considered to be significant).

SUMMARY

More than 70 km of slow-moving traffic formed during the early morning hours on May 3, 1989, in the downstream direction on the Tomei Expressway. Three major bottlenecks exist in the downward direction on the Tomei Expressway. They are the Atsugi Interchange, at which point the three-lane roadway narrows to two lanes (about 35 km from the Tokyo entrance), the compound geometric section with an S-curve and a sag section near Hatano (about 46 km from Tokyo), and the entrance of the Tsuburano Tunnel (about 1.7 km long, and about 68 km from Tokyo).

The propagating speeds of the density boundary between the free and congested-flow regions were about 7 to 13 km/hr, and these speeds were lower than the propagating speeds of the traffic flow fluctuation in the congested region. The differences in the two propagating speeds can be explained using fundamentals of traffic shock wave theory. Accordion-like actions of the traffic flow were performed in the congested flow. The fluctuations of accordion-like actions of traffic flow on the two and three lanes were synchronized, and propagate upstream at speeds of about 17 to 24 km/hr. Growth in the amplitude of traffic fluctuations in the congested flow was also recognized. These characteristics of traffic flow in the congested flow were the same as those of the traffic flow on the urban expressways. There were some upper limit speeds in the fluctuations of the traffic flow in the congested flow region on the intercity expressways, the same as that of the urban expressways in nature. The maximum flow rate on the shoulder lane at the Tsuburano Tunnel will be estimated to be 1,550 veh/hr per lane.

The free speeds estimated in the $Q-V_s$ plane using the linear regression method have differences between the detector stations and between lanes. The differences seemed to be affected by the geometric factors at the detector stations. A multivariate analysis was applied to estimate the relationship between the free speed and geometric factors. According to the results, the free speeds were affected by the gradients, radii at the stations, and sum of radii existing within 1.5 km downstream at the stations.

ACKNOWLEDGMENTS

The support provided by the Japan Highway Public Corporation is gratefully acknowledged. The assistance of H. Mori, Mitsubishi Research Institute, in making the data available to the author was essential to this study. The author also wishes to express his appreciation for the efforts of Mr. Aoyama and Mr. Yoshizumi who assisted in the calculation and drawing up of some of the figures for this paper.

REFERENCES

1. M. Koshi. Traffic Capacities at Motorway Bottlenecks. *Proc., Japanese Society of Civil Engineers*, No. 371/IV-5, 1986 (in Japanese), pp. 1–7.
2. *Special Report 87: Highway Capacity Manual*. HRB, National Research Council, Washington, D.C., 1965.
3. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
4. *Highway Capacity*, Japan Road Association. 1984 (in Japanese).
5. H. S. Mika et al. Dual Mode Behavior of Freeway Traffic. In *Highway Research Record 279*, HRB, National Research Council, Washington, D.C., 1969, pp. 1–13.
6. B. D. Hillegas et al. Investigation of Flow-Density Discontinuity and Dual Mode Behavior. In *Transportation Research Record 495*, TRB, National Research Council, Washington, D.C., 1974, pp. 53–63.
7. M. Iwasaki. Study on Car-Following Behaviors and Traffic Flow Characteristics on The Urban Expressways. Ph.D. dissertation, University of Tokyo, 1981 (in Japanese).
8. *Data Book for Highway Capacity Study*. Japan Society of Traffic Engineers, 1986 (in Japanese).
9. F. L. Hall et al. Empirical Analysis of Freeway Flow-Density Relationships. *Transportation Research A*, Vol. 20A, No. 3, 1986, pp. 197–210.
10. F. Navin and F. L. Hall. Understanding Traffic Flow at and near Capacity. *ITE Journal*, 1989, pp. 31–35.
11. T. Ooba et al. Study on Perception and Reaction Time in Car-Following Behavior. *Proc., 42nd Annual Meeting of Japanese Society of Civil Engineers*, 1987 (in Japanese), pp. 58–59.
12. H. Akahane et al. Development of a Measuring System of Spacing, Speed, and Acceleration of a Following Vehicle. *Proc., Infrastructure Planning*, No. 11, 1988 (in Japanese), pp. 63–70.
13. T. Oguchi et al. Running Behavior of a Test Car at Hikone Tunnel on Meishin Expressway. *Proc., 44th Annual Meeting of Japanese Society of Civil Engineers*, 1989 (in Japanese), pp. 156–157.
14. T. Oguchi et al. Comparison of Car-Following Behaviors Between Basic Segment and Tunnel on Expressway. *Proc., Infrastructure Planning*, No. 12, 1989 (in Japanese), pp. 75–80.
15. M. Koshi et al. Some Findings and an Overview on Vehicular Flow Characteristics. *Proc., 8th International Symposium on Transportation and Traffic Theory*, 1983, pp. 403–426.
16. *Special Report 165: Traffic Flow Theory*. TRB, National Research Council, Washington, D.C., 1975, pp. 111–165.
17. F. L. Hall. Roadway Occupancy as a Criterion for Freeway Management. Presented at the annual meeting of the Roads and Transportation Association of Canada, 1986.
18. T. Komazawa. Quantification Theory and Data Processing. Asakura Shoten, Tokyo, 1982 (in Japanese).

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

Freeway Control Using a Dynamic Traffic Flow Model and Vehicle Reidentification Techniques

REINHART D. KÜHNE

Freeway traffic flow is described in terms of control theory. The detecting elements of millimeter-wave radar sensors, which detect speed and occupancy time by a 61-GHz continuous-wave doppler radar, are used. The regulating unit consists of variable traffic signs for traffic-dependent speed limit and alternative route guidance. The control unit consists of a local computer and a control center. The control strategy is based on a continuum theory of traffic flow, which takes into account characteristics of the speed distribution for different traffic states. For incident detection and early warning criteria, the model yields the traffic density as a crucial stability parameter. For measuring the traffic density, a correlation technique is presented that for dense traffic uses the radar reflection signals as fingerprints for reidentification of vehicles.

To avoid congestion, freeway traffic control systems are designed that detect traffic flow and influence traffic by display of variable traffic signs for speed reduction or for alternative route guidance.

The acceptance of such systems and the improvement potential increases with a good adaptive control strategy. A proper control strategy needs data from accurate traffic detectors as well as an advanced modeling of traffic flow. For this aim, new millimeter-wave radar detectors are developed that measure vehicle speed by doppler frequency shift of the transmitted 61-GHz radar beam within an accuracy of ±2 km/hr. The detectors can easily be mounted in overhead position on a traffic sign bridge. No interruption of traffic is necessary as in the case of inductive loop installation. Advanced modeling of traffic flow is possible on the basis of a continuum description. This continuum description contains a relaxation to the static equilibrium speed-density relation and an anticipation of traffic conditions downstream.

In this way, the premises of a well-adapted traffic control system are fulfilled. The proper control unit to transform this traffic detector data and traffic model into a control strategy is a hierarchical control architecture. This architecture consists of local control units, which handle nearly autonomously the traffic data and threshold values derived from the traffic model.

It consists of a control center that coordinates the local displays of the variable traffic signs within a harmonizing strategy also considering external weather influences. Figure 1 shows the elements of the control circuit for a line control setup.

Besides the traffic detectors (with millimeter-wave radar detectors used advantageously as detecting elements), visibility detectors and rain detectors are installed. The local control unit is in the cabinet at the street border, and the control center is in a control room within the traffic agency. The changeable traffic signs form the regulating units in the sense of a control circuit and indicate not only traffic-dependent speed limits but also Keep in Lane, No Passing for Trucks, or Lane Blocked, because the displays have universal contents. A famous example of a freeway network used for alternative route guidance is the New York Integrated Motorist Information System (IMIS) corridor (*1*). Such network control systems need sensors for travel time detection and a forecasting strategy for traffic demand on main and alternative routes.

For this purpose, vehicle reidentification techniques, which classify the radar reflection patterns by pattern recognition methods, are developed. In this sense, each vehicle produces a fingerprint that can be compared at neighboring measurement sites with preceding fingerprints. If two fingerprints are in accordance, travel time and traffic density can easily be derived.

CONTINUUM DESCRIPTION OF TRAFFIC FLOW

A macroscopic description of freeway traffic uses the variables traffic flow q (in vehicles per hour), traffic density ρ (in vehicles per kilometer), and mean speed v (average speed over an interval such as 5 min). (Throughout this paper ρ is used as density symbol rather than k, which is used by other authors.) The relation

$$q = \rho v \tag{1}$$

and the empirical speed-density relation

$$v = V(\rho) \tag{2}$$

always hold. The latter can be transformed directly to the fundamental diagram $q = Q(\rho)$. (The capital letters V and Q denote functional relations, and the lowercase letters v and q refer to actual variables.)

For a dynamic description, the continuum limit of a general car-following behavior is considered. The speed v_n of the nth car in a line of cars reacts with a time lag τ on the change of

Daimler-Benz Aktiengesellschaft Research Center, Wilhelm-Runge Strasse 11, 7900 Ulm, Germany.

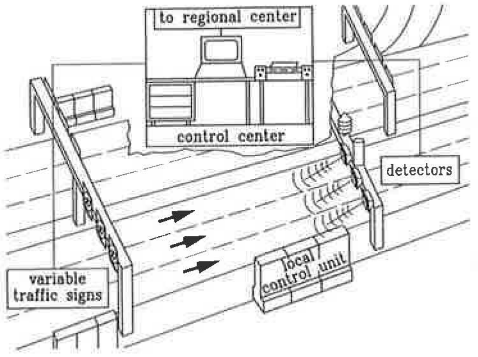


FIGURE 1 Elements of the control circuit.

headway h_n of the subject car relative to that of the preceding car:

$$v_n(t + \tau) = F(h_n) \tag{3}$$

In this relation, a general headway function $F(h_n)$ is used. Introducing differentiable functions $v(x,t)$ for speed and $\rho(x,t) = 1/h$ for density leads to the acceleration equation [compare with that of Kühne (*2*)] with a general speed-density relation instead of the general headway function $V(\rho) = F(1/\rho)$:

$$\frac{dv}{dt} = v_t + vv_x = \frac{1}{\tau}[V(\rho) - v] - c_0^2 \frac{\rho x}{\rho} + \nu_0 v_{xx} \tag{4}$$

In Equation 4, subscripts t and x denote partial derivatives with respect to time and space coordinates, respectively. The acceleration equation contains a relaxation to the empirically fitted equilibrium speed-density relation $V(\rho)$ and an anticipation of traffic conditions downstream. The first-order derivative (pressure term) leads to a deceleration (acceleration) when traffic density grows (decreases), and the second-order derivative leads to temporal speed change when the spatial density trend changes (viscosity term). The relaxation time τ, the "sound" velocity c_0, and the dynamic viscosity ν_0 (in the limit $\nu_0 \rightarrow 0$) are constants. The relaxation time τ denotes the response time of a vehicle collective for compensation of a speed difference. The "sound" velocity results from the spreading velocity of disturbances in the absence of relaxation and viscosity. Finally, the dynamic viscosity introduces a small shear layer to smear out sharp shocks. The viscosity term even in the limit $\nu_0 \rightarrow 0$ guarantees a continuous description of roll waves and other traffic patterns for unstable traffic flow.

Besides this, the equation of continuity holds,

$$\rho_t + q_x = 0$$

which states that a temporal change in density takes place only if a spatial change of net flow occurs.

One possibility is to use the model for spatial or temporal forecasting by integrating the equations given traffic volume and mean speed time series as boundary conditions. In this way, time series of traffic volume or mean speed for cross sections downstream or at later times can be predicted (*3*).

The traffic flow model also is a powerful tool for explaining general properties of traffic flow such as stability regimes, stop-start waves, and critical fluctuations. Stability analysis of the homogeneous solution $[\rho_0, V(\rho_0)]$ yields stability when the traffic parameter is negative. When

$$a \equiv -1 - \frac{\rho_0}{c_0}\frac{dV(\rho_0)}{d\rho} < 0$$

traffic waves decay to the static speed-density relation $V(\rho)$. When

$$a \equiv -1 - \frac{\rho_0}{c_0}\frac{dV(\rho_0)}{d\rho} > 0$$

the static flow $\rho = \rho_0$ and $v = V(\rho_0)$ are no longer stable. The traffic parameter reflects the small-signal performance and not only the speed-density relation itself. The quantity $a(\rho_0)$ contains the operating point density ρ_0, the slope of the speed-density relation, and the "sound" velocity c_0. These are the crucial dynamic parameters of traffic flow besides the relaxation time τ.

The critical density ρ_c at which the change from stable traffic flow to unstable traffic with jams and stop-start waves is given by

$$-1 - \frac{\rho_c}{c_0}\frac{dV(\rho_c)}{d\rho} = 0$$

To estimate the slope of the speed-density relation $v = V(\rho_c)$, a proper fit of the measurement data with restriction to data from noncongested traffic flow is necessary. The fit procedure described by Kühne (*4*) yields, for West German autobahns under normal conditions,

$$\rho_c = 25 \text{ veh/km per lane}$$

The speed-density relation and its critical density depend on street curvature, light, and weather conditions.

TRAFFIC PATTERNS FOR UNSTABLE TRAFFIC FLOW

Far beyond the critical density at which traffic flow nearly breaks down completely, stop-start waves occur. These stop-start waves can be derived from the traffic flow model when the equations are transformed to the collective coordinate

$$z = x - v_g t$$

as the only independent variable. This transformation means that only running profiles with stable shapes are considered. The group velocity v_g of the profile motion will be determined by proper boundary conditions. Integration of the continuity equation yields

$$\rho(v - v_g) = Q_0$$

This equation means density and speed in a frame running with group velocity v_g must always serve as supplements. Wave solutions with a profile moving along the highway are only possible if the density at one site increases in the same proportion as the mean speed with respect to the group velocity v_g decreases, and vice versa. The constant Q_0 has the meaning of an external given flow.

The remaining acceleration equation

$$\nu v_{zz} + F'(v)v_z + H(v) = 0$$

has the form of a nonlinear wave equation with an amplitude-dependent damping term and an unharmonic force (*5,6*). In the limit of vanishing viscosity, $\nu \rightarrow 0$, sawtooth oscillations occur.

Figure 2 shows the stop-start wave solutions together with measurements from West German autobahns. The oscillations are strongly asymmetric. Decreases of speed occur much more quickly than increases of speed. This asymmetry was pointed out early by Gazis et al. (*7*). It is a consequence of nonlinearities caused by convection and indicates again the power of the traffic flow model. Even such details of traffic behavior as the difference between deceleration and acceleration behavior coincide with observation. It is therefore not necessary to discriminate in the relaxation time between relaxation from high speed levels downward or from low speed levels upward.

In the vanishing viscosity limit, a simple relation between the amplitude $A = v_{\text{max}} - v_{\text{min}}$ of the stop-start waves and the oscillation time T of these waves can be derived (*2*).

$$T = \frac{1}{|v_g|a} \oint dz \approx \frac{2\tau}{|v_g|a} \sim A$$

The amplitude A describes the difference between the maximum speed v_{max} and the minimum speed v_{min} during an oscillation. The oscillation time T defines the repeating time of the start-stop waves.

The approximate proportionality between amplitude and oscillation time is significant for nonlinear oscillations. In contrast for harmonic oscillations, amplitude and oscillation time are independent and are only fixed by geometrical dimensions. In Figure 3, four measurements of German autobahn stop-start waves are shown in which each measurement had a duration of several hours with continuous stop-start wave conditions. The theoretical predictions are well fulfilled. This again is an example for the capacity of the traffic flow model under consideration.

Stop-start wave propagation is not the only possible solution in the congested traffic regime. The premise for such an oscillatory solution is the existence of a limit cycle for the nonlinear wave equation.

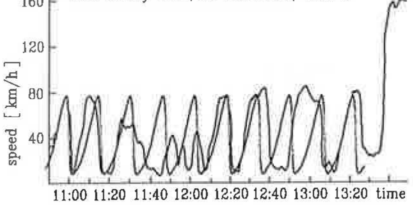


FIGURE 2 Sawtooth oscillations of mean speed for stop-start traffic together with measurements from Autobahn A5 near Karlsruhe, West Germany (*6*).

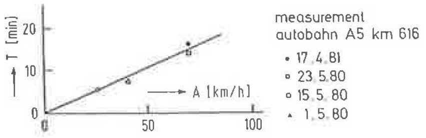


FIGURE 3 Oscillation time *T* and stop-start wave amplitude *A* from four measurements of stop-start traffic each of several hours duration (*8*).

A limit cycle occurs if an unstable-focus fixed point has a neighboring saddle point. Analyzing the fix points of the nonlinear wave equation by examination of the vicinity of the zeros of the force term $H(v)$ indicates that the only fixed point that can form an unstable focus is the fixed point corresponding to the operating point $\rho = \rho_0$, $v = V(\rho_0)$. When this operating point is also a saddle point, only shock front spreading occurs and no oscillatory solution is possible. The shock front jumps between the two equilibrium solutions corresponding to the zeros of $H(v)$, one corresponding to the operating point and the other lying in the creeping regime or in the nearly free condensed-traffic regime. The propagation (group) velocity v_g of such shock waves is given by

$$v_g = V(\rho_0) - c_0$$

in which $V(\rho_0)$ is the equilibrium speed of the operating point from the speed-density relation and c_0 = 65 km/hr, the "sound" velocity of disturbances spreading in the absence of relaxation and viscosity terms (*3,8,9*). As a typical value for dense traffic operation, $V(\rho_0)$ = 50 km/hr leads to a negative group velocity

$$v_g \approx -15 \text{ km/hr}$$

which means backward spreading of traffic stops. The propagation velocity differs clearly from the "sound" velocity c_0. The "sound" velocity would be the disturbance propagation velocity in the absence of the relaxation term and viscosity. An artificial separation of anticipation and relaxation is not possible, because the two effects are coupled.

Besides the dynamics, which are described by convection, anticipation, and continuity equations, the form of the speed-density relation $v = V(\rho)$ and the position of the operating point determine which of the possible solutions in the congested traffic flow regime will occur. The speed-density relation is derived from local measurements of speed and traffic volume $v = V(\rho)$. Such measurements are shown in Figure 4 from West German autobahns and from a field survey of parkways in the New York area (*10*). The latter data are early measurements reported by Roess et al. (*10*) under truly ideal conditions—no trucks or buses, lane width of 3.60 m, and adequate lateral clearance.

The speed-volume relation can be transformed to a volume-density relation (fundamental diagram) using the relation $q = \rho v$. A fundamental diagram produced from Figure 4 is shown in Figure 5 together with different traffic state regimes in the unstable traffic regime derived from the traffic flow model. In contrast to the level of service subdivision, where the unstable traffic regime is not further subdivided, the traffic

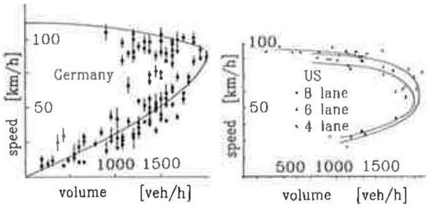


FIGURE 4 Speed-volume relation for West German (*left*) and U.S. (*right*) freeways from local measurements (*9,10*).

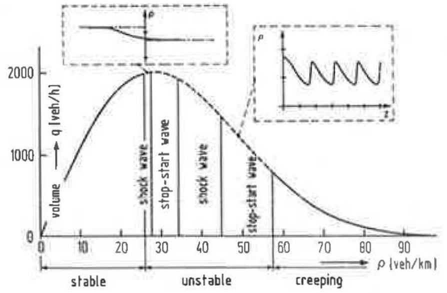


FIGURE 5 Fundamental diagram corresponding to speed-flow characteristics of Figure 4. The unstable traffic regime is subdivided into several regimes of different traffic behavior on the basis of the traffic flow model (*11*).

flow model yields in that specific case various different traffic behaviors such as regular stop-start waves or spreading of single shock waves. Because there are many different neighboring traffic states, under nearly the same congestion conditions on a given day, regular stop-start waves occur, and on another day sticky traffic with irregular perturbation spreading occurs.

Shock wave spreading is described within the kinematic wave theory (*11–13*). The propagation velocity of these shock waves can simply be calculated from the corresponding volume and density jumps. The only promise for such shock wave solutions is the existence of two fixed points between which the jump occurs. The proper statement of Figure 5 is the coexistence of shock waves and stop-start waves in the unstable regime. To derive the oscillating solutions, drivers' reactions not only to the amount of density or speed variations but also to the tendency of these variations must be taken into account. A large deviation justifies the introduction of the viscosity term together with the general feedback mechanisms, which establish the nonlinearities. This driver behavior leads to a change of topology as the control parameters [for instance, the position of the operating point ρ_0, $V(\rho_0)$] are changed.

As a particularly interesting example of stop-start waves, Figure 6 shows a control parameter configuration in which the stop-start wave solution surrounds a stable fixed point and an unstable limit cycle. This subcritical bifurcation leads to bistability: a stop-start wave solution and a homogeneous solution (divided by an unstable limit cycle) exist at the same time. The change from one to the other is connected with hysteresis effects. For presentation, a v_z – v-phase portrait was chosen. The amplitude depends on the external given flow for the complete subcritical case, which is also indicated in the same figure.

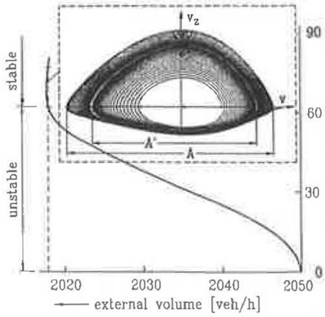


FIGURE 6 Bistability of stop-start wave solution, homogenous solution as v_z – v–phase portrait, and subcritical amplitude dependence with respect to the external given flow Q_0 (c_0 = 60 km/hr, ρ_0 = 34 veh/km) (*4,5*).

TRAFFIC FLOW MODEL AND CONTROL STRATEGY

The basis for the traffic control strategy is operating level of service (LOS) derived from idealized speed flow characteristics and field measurements as reported in Figure 4. From such measurements, LOS standards are deduced that define regimes of free, nearly free, or unstable traffic flow. The classification is shown in Figure 7. The draft is transformed from the 1985 *Highway Capacity Manual* (HCM) (*14*) to West Germany autobahn conditions; additionally a polygon approximation is shown that is used in the control center of the traffic area of Bavaria North (*15*).

The traffic flow model allows the derivation of a dynamic traffic state classification and appropriate early warning criteria. Because the traffic flow model describes traffic flow in terms of fluid motion, one expects that (as in hydrodynamics in which the change from steady flow to turbulent flow is connected with critical fluctuations and strongly irregular motions in the turbulent regime) the change from steady traffic to traffic with jams and stop-start waves is indicated by large fluctuations. The fluid analogy points out the reason for the strong spreading of the measurement values in the unstable regime; in fluid motion the convection nonlinearity serves as a feedback circuit. Beyond a certain feedback strength, the fluid motion is characterized by turbulence irregularities. Large eddies are diminished because the nonlinearity describes saturation; the following limited motions are increased because the nonlinearity amplifies more than proportionally. Also, in traffic flow, besides the ever-present random influences (e.g., bumps, irregularities in street guidance, and fluctuations in driver attention), the nonlinearity stochastics caused by feedback by convection motion produce critical fluctuations. The influence of the omnipresent fluctuations was investigated by Kühne (*2,4,16*). The irregularities and dynamic fundamental diagram have also been interpreted in the sense of chaos theory (*17–20*).

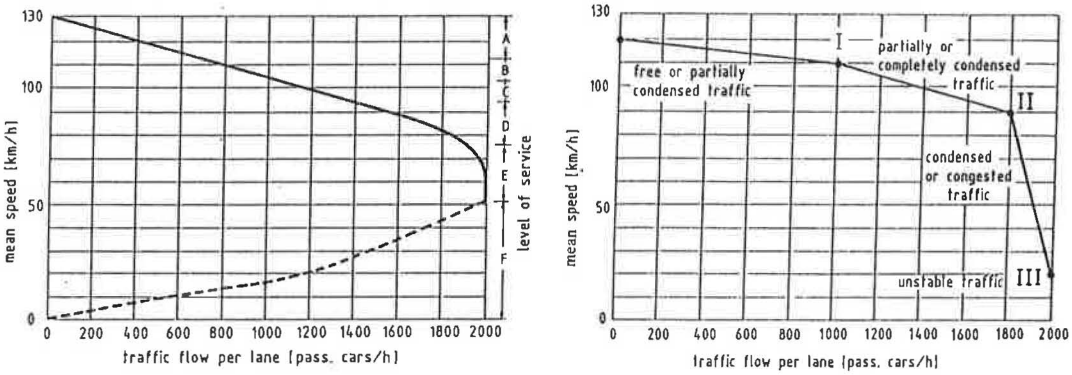


FIGURE 7 Highway capacity and LOS—transformation from HCM to West German conditions (*left*) together with a simplified polygon approximation used in the traffic control area Bavaria North (*right*) (*15*).

The essential result of all these investigations is that the speed distribution of traffic flow broadens before traffic breaks down. The standard deviation of the speed distribution as a measure of the broadness of the speed variety is therefore an early warning criterion of impending congestion and turbulent flow.

In Figure 8, the speed distribution is shown including the standard deviation for traffic with several traffic breakdowns from German Autobahn A5 near Karlsruhe. In each case, the standard deviation crosses the threshold of 18 km/hr before the mean speed sags. This provides the basis for an early warning strategy by detecting the broadening of the speed distribution. On the basis of the derived early warning strategy, a control logic was derived that is shown in Figure 9. It will be implemented in the control center of the Bavaria North control area in West Germany.

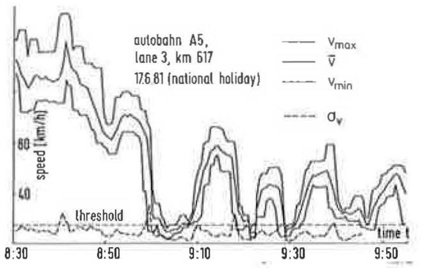


FIGURE 8 Speed distribution, standard deviation, and early warning principle (*6*).

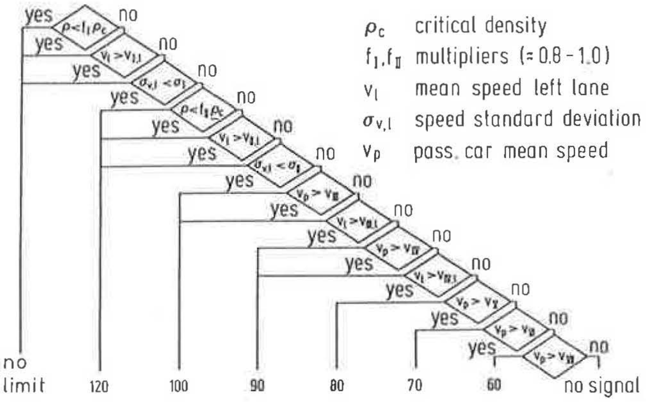


FIGURE 9 Control strategy derived from early warning principles by evaluating speed distribution.

The control logic presented in the flow chart uses a combination of several threshold values. The decision whether a specific speed limit is switched is made on the basis of the actual traffic density. As the traffic flow model indicates, this traffic density determines the appearing traffic patterns. The traffic density is a variable that is related to traffic conditions of a complete street segment in contrast to local variables such as traffic volume or mean speed, which refer only to a local point. As threshold value for density comparison, a first value can be derived from the traffic flow model condition $a(\rho_c) = 0$ with a proper speed-density fit stemming from free traffic flow conditions. In practical cases, a lane with low proportion of trucks (and therefore uniform vehicle composition) can be taken as a measurement reference. As an additional parameter for threshold comparison, the mean speed is taken. Because the calculated and measured traffic densities for heavy traffic exhibit strong fluctuations, an additional criterion is needed for a safer decision. As a third decision variable, the standard deviation of the speed distribution is used. As a speed distribution characteristic, the standard deviation measures the erratic character of the traffic flow. The more worriment in heavy traffic, the more danger of traffic breakdown exists. Readings taken on the broadness of the speed distribution yield an early warning. This criterion therefore is used on rather high speed levels.

The flow chart is designed for six different speed limits and two neutral states: no limit, 120 km/hr, 100 km/hr, 90 km/hr, 80 km/hr, 70 km/hr, 60 km/hr, and traffic jam. Translated to U.S. conditions, these values would lead to speed limits of 60, 55, 50, 45, 40, and 35 mph. The basis for the decision chart is static and stretched under practical aspects. Using a complex objective function for a dynamic decision logic makes no sense because there is no traffic diversion by a speed-influencing system.

The control strategy based on main-line control by speed limitation, temporal prohibition of passing trucks, lane keeping, or general warning homogenizes traffic flow, suppresses critical fluctuations and stabilizes traffic in a situation where without provisions traffic would have broken down. The success of such measurements is shown in Figure 10 in which the standard deviations of 2-min speed distributions are plotted against the mean speed with and without the working speed-

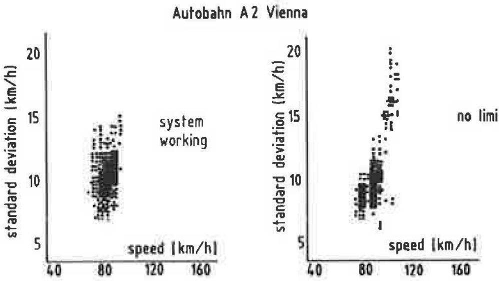


FIGURE 10 Standard deviation of speed distribution and mean speed with and without working line control system (*21*). The standard deviation is clearly reduced, impressively demonstrating the homogenizing effect.

influencing system. The data are taken from the Südautobahn A2 near Vienna, Austria (*21*). By application of the control strategy, the fluctuations are significantly suppressed.

Details of the speed distribution have proven to be an excellent tool for both assessment of traffic control provisions and early warning. To guarantee that the speed distribution has been sufficiently updated and the regarded ensemble is sufficiently stationary, the Sturges rule of thumb (*4*) is applied for speed class width estimation. In order to divide the speed distribution into a proper number of classes, this rule of thumb states that for the class width,

$$\Delta v = \frac{v_{\max} - v_{\min}}{\mathrm{ld}(2q_m t)}$$

where

ld = logarithmus dualis, and
q_m = average volume.

which means a ratio of speed range $v_{\max} - v_{\min}$ and average information content of the measurement events. For a 2-min interval and a maximum traffic flow of q_m = 2,000 veh/hr, the class width is (*4*)

$$\Delta v = 10 \text{ km/hr}$$

The speed distribution in this case is built up by speed classes each of width 10 km/hr. A finer subdivision would make no sense because of the strong fluctuations and a rougher subdivision would blur unnecessary details. The speed detection must therefore be done within an accuracy of less than 5 km/hr. This accuracy is easily obtained by millimeter-wave Doppler radar detectors described in the previous section. Meanwhile, the stated class width subdivision is standard for a modern traffic control center (*15*).

VEHICLE REIDENTIFICATION FOR TRAFFIC DATA ACQUISITION

The traffic flow model and the derived control strategy use traffic density as the crucial parameter. If stable or unstable traffic flow discrimination is performed, traffic in the unstable regime will be characterized by shock wave spreading or stop-start wave propagation, even if the speed distribution is used as an early warning criterion when the critical density is approached. In any case, the position of the operating point on the speed-density relation (and so the density itself) is the crucial traffic parameter. Traffic variables like traffic volume or mean speed are local variables. They are measured at a given location during a distinct time interval. Because of the discrete character of the traffic measurements, especially for short measurement intervals, large fluctuations occur. Likewise, large fluctuations occur in unstable traffic flow with shock front spreading and stop-start wave formation. The quotient $\rho = q/v$ has the dimensions of traffic density, but in the case of strong fluctuations it refers to the density only in the vicinity of the measurement site. In this case, there is a large difference between the calculated density and the overall density or lane occupancy of vehicles along the main-line facility. An extrapolation of $\rho = q/v$ for the complete main-line segment is not allowed. The correct determination of density takes into account the nonlocal character of traffic density. Traffic density, in contrast to traffic volume and mean speed, is a segment-related variable. It is by definition (strongly speaking) only measurable by aerial photo evaluation.

In practice, correlation techniques are applied for density determination. If two or more neighboring measurement sites are correlated, traffic variables with respect to street segments can be deduced. The most sophisticated method is the vehicle reidentification method by inductive loop pattern analysis (*22*). The impedance change of an inductive loop buried in the road surface when a vehicle passes the loop of several windings of wire is analyzed by pattern recognition methods (*23*) and used as a fingerprint. At the next measurement site, this finger print is compared with the passing finger prints. In the case of coincidence, the number of passed vehicles yields the traffic density (= number of passed vehicles within the subject main-line segment) and the waiting time until coincidence occurs gives the travel time. To avoid failures by artificial coincidence, thresholds for reidentification are regarded as well as reidentification windows for vehicle platoon reidentification. In this way, density and travel time as mean values for traffic engineering purposes are derived. Meanwhile, the traffic density and travel time detection based on inductive loops are commercially available (*24*).

Inductive loops have numerous disadvantages. They must be buried in the street pavement. Therefore on the occasions of installation, maintenance, and repair the traffic flow has to be interrupted. Under heavy truck traffic, pavement and inductive loops are damaged. Inductive loops need a metal-free environment which does not occur on bridges, steel-reinforced concrete pavement, or in tunnels. Millimeter-wave radar detectors mounted in overhead position are oblivious to these disadvantages.

After an extensive test phase, it is now possible to apply the vehicle reidentification technique also with millimeter-wave radar as traffic detectors (*25,26*). The finger print radar reflection signature is scattered back from the vehicle to the radar transmitter-receiver unit. Figure 11 shows the experimental set-up for radar signature collection. Some 1,000 signatures are collected in several measurement cycles to obtain a proper data base. Parallel to radar signature recording, the traffic scenes were stored by video to obtain the vehicle class and to reidentify vehicles using optical impressions. To sim-

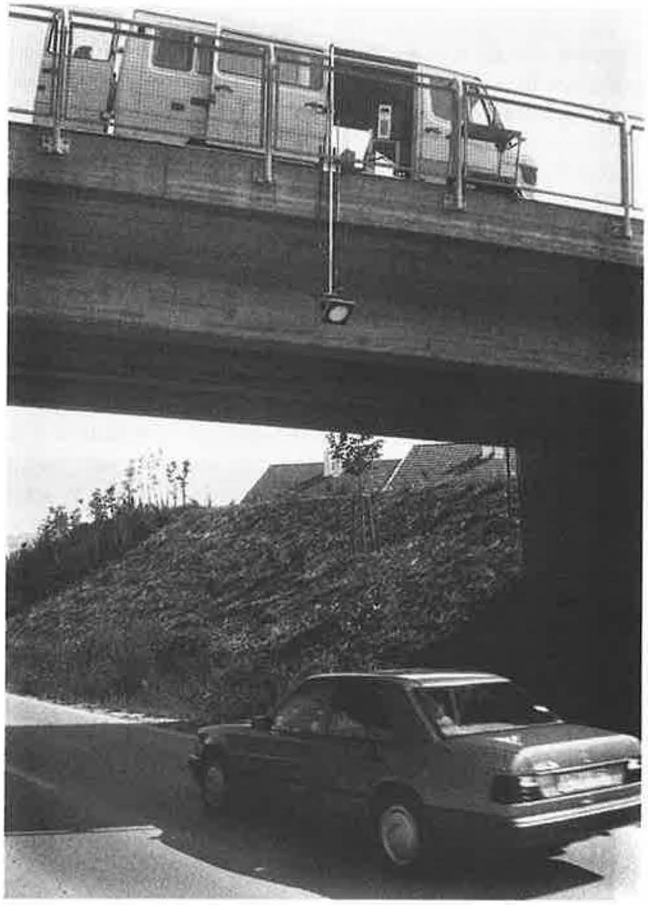


FIGURE 11 Measurement test setup for radar signature recording with parallel video recording. Small size and easy ambulant installation of the millimeter-wave radar detector are clearly recognizable.

ulate the various detector conditions, a group of test vehicles was detected more than 600 times under defined conditions by the same detector set-up.

The radar signature processing is based on envelope detection with a sampling rate of 32 kHz at a dynamic range of 60 dB. Figure 12 shows the power spectra of the same vehicle from six different measurements. The spectra vary because of different orientation of the vehicles with respect to the radar sensors. The Doppler frequency, which is determined by means of fast Fourier transformation (FFT), serves for normalizing the spectra as well as the power content. As in the simple case of only local traffic data determination (restricted to speed, signal duration, and length), the maximum spectral amplitude of the averaged power spectrum is obtained from the Doppler frequency (*27*).

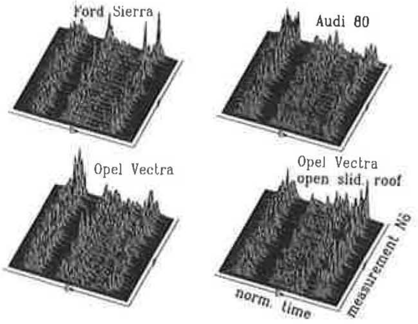


FIGURE 12 Power spectra of millimeter-wave radar reflection patterns from different measurements. The resemblance of the different patterns is recognizable, as well as the difficulty of similarity quantification.

For pattern extraction, front and tail ends of the vehicle reflection patterns are determined by proper thresholds. These thresholds are periodically adapted to react on the noise level fluctuations. Short-time amplitude breakdowns caused by interferences or radar beam deflections while the vehicle passes may be taken into consideration by controlling the time of breakdown. Beyond time control, the end threshold has been reduced to one-half of the start threshold, to avoid having to specify as an end what is actually an intermediate amplitude breakdown. After begin and end determination, the signal contour is smoothed and then coarse grained by a restriction to 100 scanning points. Mean value subtraction and covariance matrix formation finally lead to an eigenvector description of the radar signature, which can be restricted to seven eigenfunctions. This number is sufficient for accurate pattern extraction and reidentification and reduces the data amount for correlation data exchange to a data value of 1,200 bits/sec. This data rate on an additional direct data link is standard for traffic data transmission. Higher data rates press the limits of existing data transmission hardware.

In a correlation unit, the extracted finger print is compared with the finger prints of the neighboring measuring site similar to vehicle reidentification techniques using an inductive loop pattern. The correlation unit detects traffic density and travel time with respect to the preceding street section like the local traffic variables speed and volume. The complete set of data relevant to local and street sections is sent by a control module to the control center. Figure 13 shows the conception of local control stations with data acquisition relevant to street sec-

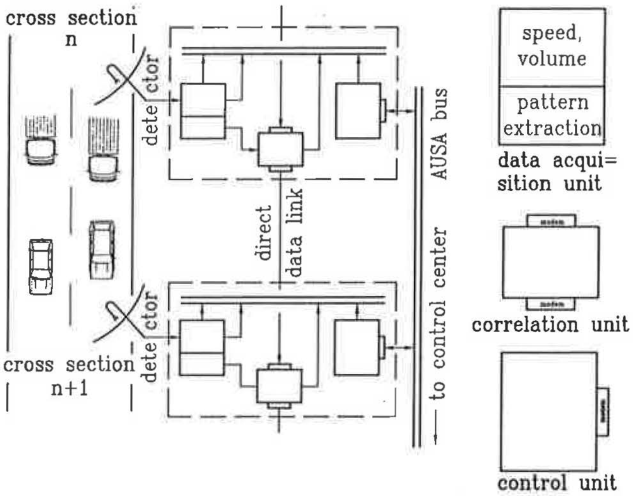


FIGURE 13 Local control stations with acquisition of traffic data relevant to street sections. Local traffic data are speed and volume; nonlocal data as determined by correlation techniques are travel time and traffic density.

tions. The data acquisition unit is the normal radar signal processing unit for local traffic data (*27*) enlarged by a pattern extraction processor. Both are installed on one European-sized board including the ECB-bus interface, the analyzing filter, and the A/D converter.

The functional architecture of the local control centers comes up to the modular set up required by the transportation authorities (*27*). With this modular concept, it is possible to use units from several suppliers, which is of great advantage in traffic control systems.

DESIGN OF A FREEWAY CONTROL SYSTEM

Traffic flow model and vehicle reidentification techniques enable the design of freeway control systems in which emphasis is put on new technologies in traffic engineering. As components of such freeway control systems, millimeter-wave radar detectors as speed and volume detectors have sufficient accuracy. These detectors can be mounted advantageously in overhead positions. For collective driver information and speed influencing, variable traffic signs are used. These signs are built up with a matrix technique; optical fibers with a pin-up lens for magnifying the light spot end in a matrix such that regular traffic signs are displayed. The fibers are collected in a bundle, which is illuminated by a halogen lamp via a coupler.

Traffic detector and variable traffic signs as well as meterological sensors are connected over a local data bus with the control unit in a local controller forming the local control station. The local control station is nearly autonomous and is responsible for the control of a street segment of 0.5 to 5 km in length. For mainline control, several such local control stations are put together. To avoid implausible jumps in the speed displays at the local control sections, the local control stations are connected by modems over a special data bus (AUSA-bus). At the end of this chain, which can take a length up to 50 km or more, is the control center where a Sun workstation is implemented. The complete control concept is shown in Figure 14. It is typical for a freeway control system in Western Europe (*28*).

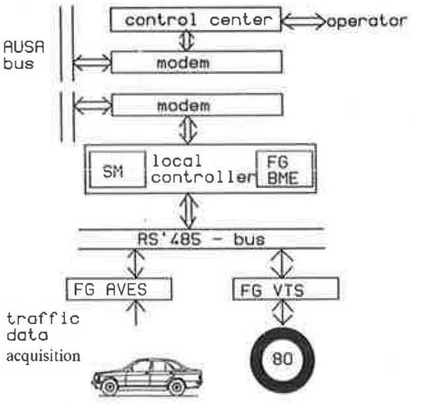


FIGURE 14 Traffic computer concept for freeway control systems. The local control stations contain a local bus connection between control unit, traffic detector (AVES), variable traffic signs (VTS).

CONCLUSION

Freeway control systems with line control and alternative route guidance in most cases use a speed-volume relation derived from local traffic data measurements as the basis of the control strategy. Traffic flow with a great variety of observable traffic patterns can be described by the dynamic traffic flow model, which uses traffic density as the crucial traffic parameter. Traffic density determines whether stable traffic flow occurs—stability conditions can be derived from the traffic flow model—or unstable traffic flow occurs with stop-start wave propagation or shock wave spreading. The density decides if nonlinearities caused by convection lead to irregularities that—together with the ever-present fluctuations—form critical fluctuations as an early warning criterion. From all these model ideas, a control system can be designed that controls traffic flow in the sense of a control circuit and that uses traffic density as the crucial control parameter. Turning away from only local traffic data affords new concepts for detection of traffic variables such as traffic density and travel time. For these, a correlation measurement concept is derived from vehicle reidentification techniques (previously used with inductive loop pattern recognition), which uses the radar reflection signature as a fingerprint from each vehicle. A satisfying concept is presented for a freeway control system based on the proposed ideas. This concept is partially realized in West German traffic control areas and can stimulate further traffic control systems.

REFERENCES

1. L. D. Powers. The Federal Highway Administration Advanced Technology Traffic Operation Research and Development Program. *Proc., 1st International Conference on Application of Advanced Technology in Transportation Engineering*, San Diego, Calif., 1989, pp. 250–264.
2. R. Kühne. Freeway Control and Incident Detection Using a Stochastic Continuum Theory of Traffic Flow. *Proc., 1st International Conference on Application of Advanced Technology in Transportation Engineering*, San Diego, Calif., 1989, pp. 273–279.
3. M. Cremer. *Traffic Flow on Freeways*. Springer, Berlin, 1979, pp. 85–114.
4. R. Kühne. Freeway Speed Distribution and Acceleration Noise—Calculations from a Stochastic Continuum Theory and Comparison with Measurements. *Transportation and Traffic Theory*, N. H. Gartner and N. H. M. Wilson, eds., Elsevier, New York, 1987, pp. 119–138.
5. B. Sick. *Dynamical Effects of Nonlinear Wave Equations with Applications to Traffic Flow Theory*. Diplom thesis, Department of Mathematical Physics, Ulm University, West Germany, 1989.
6. W. Leutzbach. *Mean Speed Time Series from Measurements on West German Autobahns*. Institute of Traffic Science, University of Karlsruhe, Karlsruhe, West Germany, 1981.
7. D. C. Gazis, R. Herman, and R. W. Rothey. Non-Linear Follow-the-Leader Models of Traffic Flow. *Operation Research*, Vol. 9, 1961, pp. 545–567.
8. H. Zackor, R. Kühne, and W. Balz. *Investigations on Traffic Flow in the Regime of Maximum Capacity and Instabilities*. Research Report 524, Federal Transportation Ministry, Bonn, West Germany, 1988.
9. R. Wiedemann. How Certain is a Fundamental Diagram. In *25 Years of Institut of Traffic Science*. Report 36, Institut of Traffic Science, University of Karlsruhe, Karlsruhe, West Germany, 1987, pp. 74–96.
10. R. P. Roess, W. R. McShane, and L. J. Pignataro. Freeway Level

of Science: A Revised Approach. In *Transportation Research Record 699*, TRB, National Research Council, Washington, D.C., 1979, pp. 7–16.

11. R. Kühne. Gunn Oscillator, Freeway Traffic Flow and Shallow Water Waves—a Comparison. In *Verhandlungen der Deutschen Physikalischen Gesellschaft, Spring Meeting, Solid State Physics—Dynamics and Statistical Mechanics*, Karlsruhe, West Germany, 1988, pp. 53.
12. M. J. Lighthill and G. B. Whitham. On Kinematic Waves. *Proc., Royal Society*, Vol. A229, 1955, pp. 281–345.
13. G. B. Whitham. *Linear and Nonlinear Waves*. Wiley, New York, 1974, pp. 68–95.
14. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
15. *Traffic Engineering Concept for Control Logic*. Special Report, Autobahn Direktion Bayern Nord, Nürnberg, West Germany, 1990, pp. 1.17–2.47.
16. R. Kühne. Microscopic Distance Strategies and Macroscopic Traffic Flow Model. *Proc., Control Computers Communication in Transportation*, Paris, 1989, pp. 383–389.
17. M. Rödiger. Chaotic Solutions of Nonlinear Wave Equations with Applications in Traffic Flow Theory. In *Verhandlungen der Deutschen Physikalischen Gesellschaft, Spring Meeting, Solid State Physics—Dynamics and Statistical Mechanics*, Regensburg, West Germany, 1990, pp. 689.
18. R. Kühne and M. Rödiger. Chaotic Behaviour of Traffic Flow. *Transportation Science*, 1991.
19. R. Kühne. Chaotic Behaviour of Shallow Water Waves. In *Verhandlungen der Deutschen Physikalischen Gesellschaft, Spring Meeting, Solid State Physics—Dynamics and Statistical Mechanics*, Regensburg, West Germany, 1990, pp. 690.
20. M. Schlett. *Continuum Model of Traffic Flow on Freeways With Unstable Flow*. Diplom thesis, Institut of Traffic Science, University of Karlsruhe, Karlsruhe, West Germany, 1990.
21. G. Steierwald, K. Stöcker, W. Fusseis, and V. Stottmeister. *Situation Dependent Traffic Influencing System, Südautobahn 12*. Research Report, Ministry of Building and Technology, Vienna, Austria, 1985.
22. P. Böhnke and E. Pfannerstill. A System for the Automatic Surveillance of Traffic Situations. *Proc., 1st Symposium Land Vehicle Navigation*, Münster, West Germany, 1984, pp. 4.1–4.17.
23. E. Pfannerstill. Automatic Monitoring of Traffic Conditions by Reidentification of Vehicles. *Proc., IEE Conference on Road Traffic Monitoring*, Report 299, 1989, pp. 172–175.
24. *MAVE-S-Innovative Traffic Flow Analysis*. AVE Verkehrs- und Informationstechnik GmbH, Aachen, West Germany, 1990.
25. R. Kühne, *Automatic Vehicle Classification by Means of Microwave Techniques*. Technical Report 12.060/82, AEG-Telefunken, Ulm, West Germany, 1982.
26. W. Stammler, N. Korn, and J. Schacherl. Radar Signal Processing for Traffic Flow Control. *Proc., EUSIPCCO 88 Grenoble*, 1988, pp. 143–147.
27. *Technical Terms of Delivery for Local Traffic Control Stations*. Special Report, German Federal Transportation Ministry, Bonn, West Germany, 1990.
28. R. Kühne, M. Rödiger, and H. Burhenne. Traffic Engineer Workstation—Traffic Data Evaluation in a Control Center on a SUN Workstation. *Proc., 10th Workshop on Real Time Processing*, Ulm University, West Germany, 1989, pp. 191–202.

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

Simulation of Crowd Behavior and Movement: Fundamental Relations and Application

Saad A. H. AlGadhi and Hani S. Mahmassani

A model of crowd behavior and movement in the Jamarat system, which is a critical bottleneck during the Hajj, the Muslims' annual pilgrimage to Makkah, Saudi Arabia, is described. The model consists of a set of partial differential equations that are solved numerically following a discretization of time and space. Mathematical relations are defined for three fundamental processes: (a) radial movement, (b) lateral movement, and (c) stoning process. These relations are developed and calibrated using actual measurements taken at the site. As a result, a bidirectional speed-concentration model is developed to describe radial movement, revealing that the impedance to movement from facility users going in the same direction is about twice that from those going in the opposite direction. The overall model is applied to the evaluation of possible design and control strategies aimed at improving the efficiency and throughput of the system.

The efficient and safe operation of heavily crowded facilities used in special circumstances by unfamiliar users is a major concern to those responsible for the design and control of such public places and events. Crowd disasters in which people are seriously injured, sometimes fatally, can occur, and have occurred at sports events, religious gatherings, and rock music concerts (*1*).

A remarkable example of pedestrian overcrowding is that of the Hajj, or Muslims' pilgrimage to Makkah, Saudi Arabia. The Hajj consists of a set of prescribed rites at specific hours and days in assigned locations in and around the city of Makkah. Each year, over two million pilgrims converge on Makkah to perform this religious duty at the same time, resulting in a crowded event of extraordinary magnitude. Every hour, there are hundreds of thousands of pilgrims walking between the various rites of the Hajj (*2*). Bottlenecks can turn into disasters, and one of the most difficult of such bottlenecks is the relatively small Jamarat system. One of the pillars of the Hajj is the throwing of (seven) small pebbles at each of the three stone monuments located at a place called Jamarat. In performing this ritual, many pilgrims suffer considerable hardships because of congestion and overcrowding in the tight Jamarat area, because of space and time constraints. For example, in the 1983 Hajj season, 64 fatalities were reported in this area because of overcrowding (*3*).

Little scientific work has been directed at the characterization of crowd behavior and movement, leaving considerable gaps in the knowledge base and methodological capabilities pertinent to this problem. The limited body of the available studies related to this subject can be classified, for convenience, into two broad categories: (a) those that deal with steady state pedestrian flows under normal conditions, and (b) those that address crowding conditions, which correspond to the higher end of the concentration spectrum, and generally involve multidirectional flow patterns with strong interactions.

The first class of studies deals primarily with streamlined and orderly pedestrian flows encountered on a daily basis in high traffic areas such as passageways, stairways, plazas, and sidewalks (*4–11*). It appears that the theory of pedestrian traffic flow for orderly movement under normal conditions has been sufficiently developed to reflect the actual behavior of foot traffic flows in such contexts, and to provide an adequate basis for design and control purposes (*11*).

The second category addresses the behavior of large crowds under nonroutine circumstances, often involving extremes of human emotion, such as excitement, fear, religious fervor, anger, or exaltation. Movement is therefore far less orderly, multidirectional interactions are strong, and dynamic effects are of greater importance than in the previous case. Some studies have qualitatively discussed the sociopsychological aspects of crowd behavior, rather than its physical aspects (*12,13*). Another group of studies has sought to develop models of individual behavior in a crowd, recognizing the constructed as well as the behavioral environments (*14–21*).

These studies do not, individually or collectively, provide a sufficient technical basis for management of crowd movements at special events. For example, no speed-concentration models exist for crowd movement with multidirectional flow patterns with strong interactions. Similarly, limited research has been directed at crowd behavior and movement during the Hajj (*2,22–27*). Few studies have attempted to characterize pilgrims' behavior in the Jamarat system (*3,28–30*). One exception is a study by the Hajj Research Center (*3*) that successfully characterized, in qualitative terms, the problems encountered in the Jamarat area. However, no models appear to have been developed to characterize pilgrims' behavior in that context.

JAMARAT SYSTEM CONFIGURATION

The Jamarat system is composed of three stone monuments symbolizing the devil, which are supposed to be stoned (i.e., pebbles thrown at each of them) by pilgrims in a given sequence, starting with the first, Small, second, Middle, and

S. A. H. AlGadhi, Department of Civil Engineering, King Saud University, P.O. Box 800, Riyadh 11421, Saudi Arabia. H. S. Mahmassani, Department of Civil Engineering, University of Texas, Austin, Tex. 78712.

then the third, Big Jamarah, as shown in Figure 1. On the first day of stoning, only the Big Jamarah is stoned by pilgrims over a period extending from sunrise to sunset. However, on each of the other days of stoning all three jamarat are stoned in sequence, over a period extending from noon until sunset. Each monument (called a Jamarah in the Arabic language) is enclosed by a ring of about 16 m in diameter, referred hereafter as the Jamarah ring, to collect pebbles. Pilgrims squeeze through the crowd to get their pebbles into the ring, and, preferably, to hit the monument. At the same time, others who have finished stoning squeeze and push their way out. On each of the three days of stoning, pilgrims arrive at about the same time, creating extremely crowded conditions. Peak crowding occurs at the same time every year (*3*). To meet the demands of the enormous mass of pilgrims, an upper level was added to the Jamarat system in 1975, so that there are now two levels from which the stoning of the Jamarat can be performed (Figure 1).

THE JAMARAT MODEL: AN OVERVIEW

Previously, the authors (*31*) developed a general crowd behavior and movement model, which consists of a set of simultaneous partial differential equations describing the principal processes that govern the dynamics of the system. The system of governing differential equations can be solved numerically by discretizing in time and space. Such solution yields a profile of the system's evolution and allows the computation of various performance measures and figures of merit. The application of this general model to the Jamarat system of the Hajj is described here, focusing on the processes taking place on the upper level of the Jamarat bridge around an individual Jamarah ring. Pilgrims' behavior and movement there are characterized by a set of fundamental relations among the principal underlying variables. In the following subsections, an overview of the principal elements of the model is given. The details of the individual model elements are reported elsewhere (*31*).

Space Discretization Scheme

Because Pilgrims' major movements around the Jamarah ring are mainly in the radial direction toward and away from the ring, the physical space around the ring is divided into a number of concentric rings, NRING, each of a width of 1 m. Each ring is divided into a number of cells, NCELL, by taking equally spaced radials from the center of the Jamarah (stone monument) to the outer perimeter of the external ring (i = NRING + 1), as shown in Figure 2. Each cell is identified by its (i,j) index ($i = 2, 3, \ldots,$ NRING + 1; $j = 1, 2, \ldots,$ NCELL; $i = 1$ identifies the Jamarah ring itself). The areas of cells in the same ring are equal, whereas the cell area increases with Ring Index i.

Facility Users' Classification and Movements

A fundamental concept in the model is the classification of the pilgrims in the Jamarat system into two major classes: Class 1 users, who have the intention of performing stoning (or Rajm), and Class 2 users, who intend to leave the Jamarah after having completed stoning. Accordingly, $K(i,j)$, the total concentration of users (in persons/m^2) in Cell (i,j), is the sum of the respective concentrations of the two classes; i.e.,

$$K(i,j) = K^1(i,j) + K^2(i,j) \qquad (1)$$

For each class of facility users, two types of movement can be identified: in the radial direction (forward and backward), and in the circumferential direction (clockwise and counterclockwise). However, it is assumed that those who have yet to perform the Rajm (Class 1) do not go backward and those who have completed it (Class 2) no longer move towards the

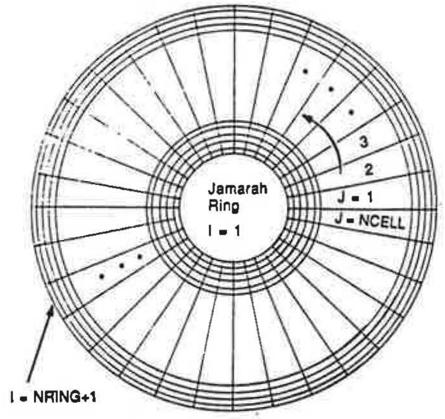


FIGURE 2 Space discretization around the Jamarah ring.

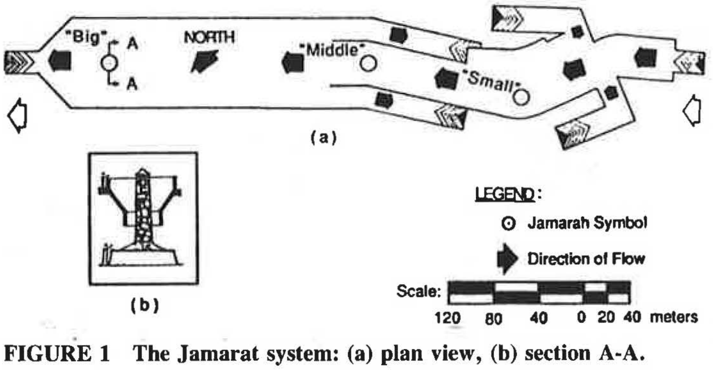


FIGURE 1 The Jamarat system: (a) plan view, (b) section A-A.

ring. Correspondingly, the following speeds of movement are defined:

U_{ij}^{1F} = Speed of Class 1 users, in Cell (i,j), moving towards the Jamarah ring (i.e., forward radial direction), in meters per hour.

U_{ij}^{2B} = Speed of Class 2 users, in Cell (i,j), moving away from the Jamarah ring (i.e., backward radial direction), in meters per hour.

$U_{ij}^{m\mathrm{LR}}$ = Speed of Class m users (m = 1, 2), in Cell (i,j), moving in the clockwise lateral direction, in meters per hour.

$U_{ij}^{m\mathrm{LL}}$ = Speed of Class m users (m = 1, 2), in Cell (i,j), moving in the counterclockwise lateral direction, in meters per hour.

Conservation of Mass Law

For a given Cell (i,j), the general balance states that the rate of change of facility users in the cell equals the net creation rate of users there plus the net rate at which users flow into the cell across its boundaries. The continuity equations used in the numerical discrete time simulation approach, as applied to the Jamarat system, take the following difference forms (Figure 3):

For Class 1 users:

$$\begin{aligned} A(i,j) * K^1(i,j)_{t+\Delta t} = {} & A(i,j) * K^1(i,j)_t + [Q^{1\mathrm{F}}(i+1,j) \\ & - Q^{1\mathrm{F}}(i,j) - \mathrm{SW}(i,j) \\ & + Q^{1\mathrm{LR}}(i,j+1) - Q^{1\mathrm{LL}}(i,j) \\ & + Q^{1\mathrm{LL}}(i,j-1) - Q^{1\mathrm{LR}}(i,j)]_t \end{aligned} \qquad (2)$$

for i = 2, 3, . . . , NRING + 1;

j = 1, 2, . . . , NCELL

For Class 2 users:

$$\begin{aligned} A(i,j) * K^2(i,j)_{t+\Delta t} = {} & A(i,j) * K^2(i,j)_t + [Q^{2\mathrm{B}}(i-1,j) \\ & - Q^{2\mathrm{B}}(i,j) + \mathrm{SW}(i,j) \\ & + Q^{2\mathrm{LR}}(i,j+1) - Q^{2\mathrm{LL}}(i,j) \\ & + Q^{2\mathrm{LL}}(i,j-1) - Q^{2\mathrm{LR}}(i,j)]_t \end{aligned} \qquad (3)$$

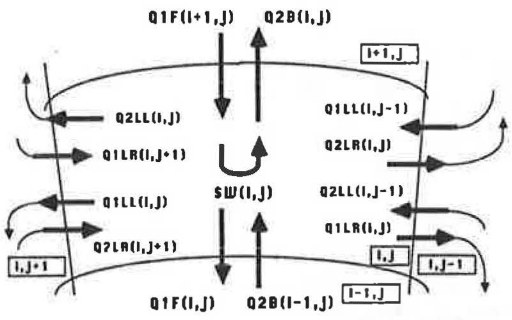


FIGURE 3 Concentration fluxes operating simultaneously in Cell (i,j).

for i = 2, 3, . . . , NRING + 1;

j = 1, 2, . . . , NCELL

where

$K^m(i,j)$ = Concentration of Class m users (m = 1, 2) in Cell (i,j), in persons/m^2.

$A(i,j)$ = area of Cell (i,j), in m^2.

$Q^{1\mathrm{F}}(i,j)$ = net flux of Class 1 users, in Δt, from Cell (i,j) in the radial forward direction, in persons.

$Q^{2\mathrm{B}}(i,j)$ = net flux of Class 2 users, in Δt, from Cell (i,j) in the radial backward direction, in persons.

$Q^{m\mathrm{LR}}(i,j)$ = net flux of Class m users, in Δt, from Cell (i,j) in the lateral clockwise direction, in persons.

$Q^{m\mathrm{L1}}(i,j)$ = net flux of Class m users, in Δt, from Cell (i,j) in the lateral counterclockwise direction, in persons.

SW(i,j) = number of Class 1 users switching into Class 2 users, in Cell (i,j) in Δt, after finishing stoning. This quantity acts as a sink term for Class 1 and a source term for Class 2 users.

Equation 1 indicates that three principal processes are identified as governing crowd behavior and movement at the Jamarat: (a) movement in the radial direction towards the Jamarah ring by pilgrims seeking to perform stoning (Class 1), and away from it by those leaving the system after completing the stoning (Class 2); (b) movement in the circumferential (lateral) direction around the ring of both classes of pilgrims in their effort to escape crowd pressure by moving to a less-congested space around the ring; and (c) the mechanism by which Class 1 pilgrims switch to the other class after completing stoning. Each of these principal processes is discussed hereafter, with the corresponding underlying variables identified, and relations among them established.

Radial Movement of Pilgrims

The radial movements of pilgrims at the Jamarah are the dominant ones relative to the lateral movements. Their flow rates were modeled by the mathematical product of the prevailing radial speed and the corresponding concentration. For example, the flux $Q^{1\mathrm{F}}(i+1,j)$ of Class 1 pilgrims, in a given time step Δt from Cell $(i+1,j)$ to Cell (i,j), is given by

$$\begin{aligned} Q^{1\mathrm{F}}(i+1,j) = {} & U^{1\mathrm{F}}(i+1,j) * K^{1\mathrm{R}}(i+1,j) \\ & * \mathrm{L1}(i+1,j) * \Delta t, \end{aligned} \qquad (4)$$

where

$K^{1\mathrm{R}}(i+1,j)$ = concentration of Class 1 users who are moving radially in cell $(i+1,j)$(persons/m^2), and

L1$(i+1,j)$ = length of the interior edge of Cell $(i+1,j)$, in meters.

The average radial speed of a given class of users in Cell (i,j) is a function not only of the concentration of users of the same class in the cell, but the concentration of both classes of users. The following calibrated relation is a central component of the Jamarat model:

$$U^{\text{in}} = 1{,}907 * [1 - (K^{\text{in}} + 0.38 * K^{\text{opp}})/K_{\text{jam}}] \qquad (5)$$

where U^{in} is the average speed, in meters per hour, of users of a given class, moving with a stream of pedestrians of the same class, with concentration K^{in} (persons/m^2), and against another stream of pilgrims from the other class with a corresponding concentration level K^{opp} (persons/m^2). The jam concentration K_{jam} was estimated to be 5.68 persons/m^2. (In the simulation runs, it is assumed that movement continues at a shuffling speed of 200 m/hr at jam concentration.) The parameters of this relation were estimated using field data, in the form of time lapse photographs, which were taken for this purpose at the Jamarat system. The details of the observational work are reported elsewhere (*31*). The calibrated relation revealed some interesting features of the process. For instance, the impedance, on the speed of a given class of facility users, caused by people moving in the same direction is more than twice that caused by those moving in the opposite direction. Another important aspect of this relation is that the marginal impedance of each of the two types of concentrations involved, on the speed of movement, is independent of the level of the other stream's concentration. That is, there is no interaction term between the two types of concentration in the estimated speed-concentration model.

Lateral Movement of Pilgrims

By analogy with the treatment of Gazis et al. (*32*) of interlane concentration oscillations in multilane highway traffic, it is assumed that the exchange of pedestrians between two neighboring cells (in the same ring) is proportional to the difference of their respective concentrations, and only when this difference exceeds a certain threshold, that is:

$$Q_{ij}^{mL} = \begin{cases} \alpha[\Delta K(i,j)]^{\beta}, & \text{if } \Delta K(i,j) \geq \Delta K^{m}_{\text{min}}, \\ 0 \text{ otherwise} \end{cases} \qquad (6)$$
$$m = 1, 2$$

where

Q_{ij}^{mL} = net lateral flux of Class m users (m = 1, 2) in Δt, from Cell (i,j) to the neighboring cell, in persons;
$\Delta K(i,j)$ = concentration differential between the two cells;
$\Delta K^{m}_{\text{min}}$ = Class m (m = 1, 2) concentration differential threshold; and
α, β = parameters.

Analysis of limited relevant data has suggested that the threshold level $\Delta K^{m}_{\text{min}}$ is about 1.0 persons/m^2 for each class of users, and yielded parameter estimates of α = 850 pilgrims/hr and β = 0.005 for purposes of the present illustration of the behavior of the Jamarat model (*31*).

Switching Mechanism (Stoning Process)

The number SW(i,j) of Class 1 users switching to Class 2 in Cell (i,j), in a given Δt depends on two principal factors: (a) the fraction FR(i,j) of Class 1 users in the cell who perform stoning from that cell, and (b) the time TRAJM(i,j) it takes an individual to perform Rajm. The relation between SW(i,j), FR(i,j), and TRAJM(i,j) can be expressed as follows:

$$\text{SW}(i,j) = [A(i,j) * K^{1}(i,j) * \text{FR}(i,j) * \Delta t]/[\text{TRAJM}(i,j)] \qquad (7)$$

The fraction FR(i,j) of Class 1 users switching to Class 2 depends on two main factors: the distance from the rim of the Jamarah ring, in meters, and the prevailing pilgrim concentration level in Cell (i,j), in persons/m^2, that is,

$$\text{FR}(i,j) = \begin{cases} 1, \text{ if } r_i \leq 1 \\ 0, \text{ if } r_i > R_{\text{max}} \text{ or } K(i,j) < K_{\text{FR}} \\ 1.005\ r_i^{-0.793}\ [K(i,j)]^{2.341}, \text{ otherwise} \end{cases} \qquad (8)$$

where R_{max} is the maximum throwing distance, estimated to be 11 m from the perimeter of the Jamarah ring; and K_{FR} is the concentration threshold, estimated to be 3 persons/m^2, below which no switching takes place (*31*). Equation 6 indicates that the fraction is initially equal to unity at the rim of the Jamarah ring, where all Class 1 users have to switch regardless of the prevailing concentration level. Beyond the first ring, the fraction is expressed as a function that decreases with distance from the rim until it equals zero beyond the maximum throwing distance from the Jamarah ring, R_{max}. Furthermore, the fraction switching increases with the prevailing concentration beyond a threshold concentration, K_{FR}, below which pilgrims would prefer to proceed further to perform stoning from a closer ring. The values of the parameters reported in Equation 7 are based on measurements taken on site (*31*).

Regarding the other principal factor, TRAJM, the observational evidence did not support the hypothesis that it is a function of either the distance from the ring or the prevailing concentration level; it was estimated to have an average value of 27 sec (*12*).

Choice of Solution Time Step and Space Discretization Scheme

Finally, the choice of time step, Δt, and space discretization scheme (cell width in the radial direction, Δr), when using numerical methods to solve partial differential equations, has important implications for the properties of the solution. Of course, it is generally desired that both Δr and Δt be small enough to approximate the assumed underlying smooth function of $K(i,j)$ with respect to r and t. In addition, the ratio $\Delta r/\Delta t$ should be greater than the mean free speed U_{max}, to ensure that no pilgrim transfer is possible from one cell to another nonadjacent cell in any time step of the numerical solution. The step sizes used in the solution of this model are Δr = 1 m, and Δt = 10^{-4} hr, for a ratio of 10 km/hr ($>>$ U_{max}).

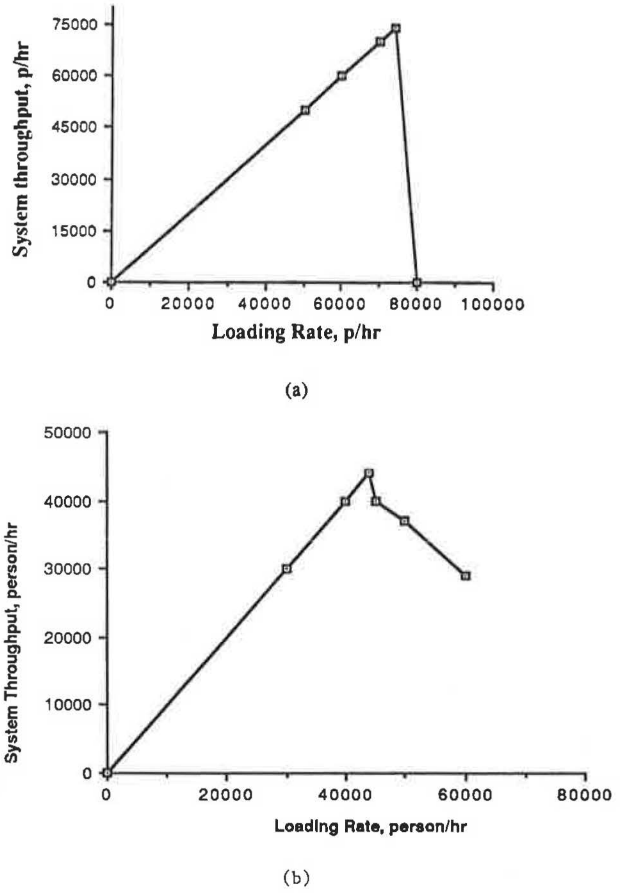
MODEL APPLICATION

In this section, the results of several numerical simulations are presented to illustrate the Jamarat system model and its application to the investigation of the system's properties under several possible crowd control and management schemes. The system throughput is examined first, followed by an investigation of the effect of several design and operational strategies on system performance.

Loading Strategies and Throughput Analysis

The Jamarat system was simulated under different spatial and temporal pilgrim loading patterns. Two different spatial loading patterns are considered here, as shown in Figure 4. In the first pattern, pilgrims approach the Jamarah from all surrounding directions; in the second pattern, they load from the ring's half side that first faces the pilgrims approaching from the main entrance (i.e., from the southeast, Figure 1). For a given spatial pattern, a variety of temporal loading patterns can be simulated. In the present exploration of maximum system throughput, the system was loaded continuously at a constant rate (for a given spatial pattern) until a steady state was attained. The system's maximum throughput was obtained by performing a series of simulations under different values of the loading rate and observing the behavior of the system.

Figures 5a and 5b show the maximum throughput as a function of the loading rate when loading from all directions and from one side only, respectively. For the first spatial loading pattern, and using the parameter values and relations described earlier, the maximum system throughput achieved is about 74,000 pilgrims/hr. Loading at rates higher than the optimum caused the system to break down. Similarly, the maximum throughput of the system achieved for the case of semiuniform loading (from one side only) was found to be about 44,000 pilgrims/hr. However, loading at higher rates caused the throughput to decline accordingly to lower but stable values.



(a)

(b)

FIGURE 5 Maximum throughput of the Jamarah when loading uniformly from (a) all directions around the ring, and (b) one side only.

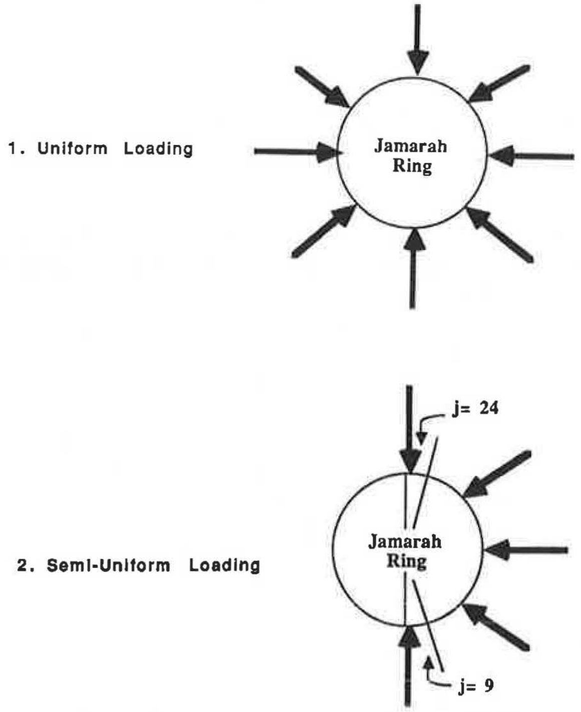


FIGURE 4 Spatial loading patterns of Jamarah ring.

Overall system behavior in both cases appears reasonable. In the first case, the system gets locked up beyond a maximum loading rate that it can absorb, because masses of each class of users eventually block the others' progress. In the second case, one-half of the Jamarah is used for loading whereas the other half functions as an evacuation outlet; the decline in throughput at rates higher than optimum is caused by the higher concentration levels, especially of Class 1 pilgrims, which results in more impedance to those who have already finished the stoning (Class 2) and are therefore seeking to leave the system.

It is interesting to point out that a field study conducted independently at the Jamarat measured the maximum throughput of the system at 69,000 pilgrims/hr (only at the top level of the Jamarat bridge) for a case that approximates that of loading from all directions (during the first day of stoning when pilgrims are required to stone only the Big Jamarah, which can be approached from any direction), and at 43,000 pilgrims/hr for loading only from one-half of the Jamarah (during the second and third days of stoning) (*28*). The findings are in remarkably close agreement with these measurements even though the measurement procedure by the Hajj Research Center (*28*) is not directly comparable to the manner in which throughput is calculated in the simulation.

An extensive investigation of the dynamic properties of the system is reported elsewhere (*31*). An illustration of the dynamic behavior of the system is given here. For this purpose, pilgrims are injected from one side only at a constant rate (of 70,000 pilgrims/hr, well in excess of the optimum level of 44,000 pilgrims/hr) over a given period (0.3 hr in this case, which is sufficient to illustrate concentration build-up), after which the inflow is shut off until the system clears. The model yields the values of the concentrations of the two classes of users in every cell at every time step, as well as a variety of associated summary measures. Figure 6 shows the dynamic concentration profile, for each class of facility users, prevailing at the rim of the Jamarah ring (i.e., for Ring $i = 2$), after the inflow has been shut off. Note that loading was taking place from one side only, i.e., for Sectors $j = 9, 10, \ldots, 24$. Figure 7 shows the corresponding dynamic concentration profile for a selected sector ($j = 13$). These figures correspond to an extremely congested situation, because the loading rate was well in excess of the system capacity. Class 2 users have been able to push their way out radially, away from the rim, which is dominated by Class 1 users, and into Rings 7 through 13.

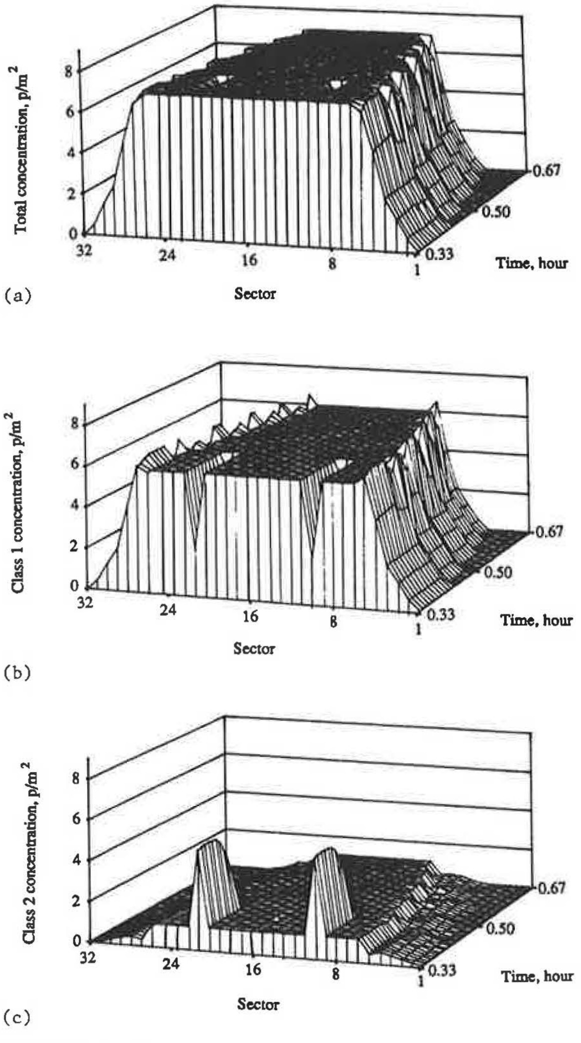


FIGURE 6 Concentration profile dynamics, at the rim of the Jamarah (Ring $i = 2$), with 70,000 pilgrims/hr loading rate from one side only, with loading shut off at time = 0.3 hr, for (a) both classes, (b) Class 1, and (c) Class 2 facility users.

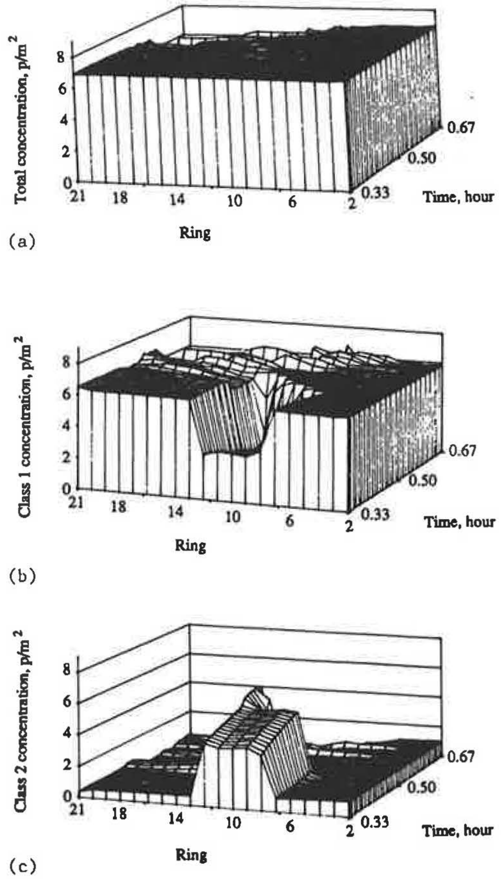


FIGURE 7 Concentration profile dynamics, at Sector $j = 13$, with 70,000 pilgrims/hr loading rate one side only, with loading shut off at time = 0.3 hr, for (a) both classes, (b) Class 1, and (c) Class 2 facility users.

Design and Control Strategies

So far, only a constant-rate temporal loading pattern was considered to examine the maximum throughput of the system. Another loading pattern was considered by pulsing pilgrims into the system: the system is loaded continuously for a given time, TPULSE, then shut off for an equal time before resuming loading again, and so on. Different runs were performed using different values of TPULSE ranging from 5 to 15 min. A similar strategy was successfully used for improving vehicular traffic flow in tunnels (*33*). However, for the case considered here, no improvement in system performance was observed, probably because of the fact that the case is not dealing with a continuous movement of individual users, because each pilgrim is supposed to completely stop to perform stoning.

In order to attain the maximum throughput of the Jamarat system under the current situation, the analysis presented earlier suggests that the input rate should be maintained at about the optimum level. However, system stability considerations suggest that a safety buffer should be allowed for both types of spatial loading patterns considered. The actual rate at which pilgrims would be admitted to the system should be below the theoretical optimum by the corresponding buffer. Furthermore, on the basis of the analysis presented earlier, the all-directions spatial loading pattern should have a larger buffer than the one-side-only loading pattern because of impending system breakdown when the Jamarah is loaded from all directions at rates higher than the optimum.

In practice, pilgrims' flow into the system could be controlled by scheduling their arrival times through their official guides, who are responsible for providing them with accommodations, transportation, and logistical support. However, this may not work too well in part because of the fact that those guides do not have much control over their pilgrims' local movement within the Mina area, where the Jamarat is located. Therefore, use of some kind of direct control at the facility itself may be considered. One possibility would be to use flexible physical barriers at the entrances of the Jamarat bridge that could funnel the flow of pilgrims into the system so as to result in the established loading rate (e.g., using an inverted cone-shaped barrier). Extreme caution should be taken to avoid the use of fixed hard barriers for this purpose, because it might result in a dangerous pedestrian crush; one possibility is to use horses to form the boundaries of the funneling apparatus, with personnel riding on the horses to direct the crowd.

On the other hand, to improve the maximum throughput of the system, one may consider changing the design of the physical system or affecting the behavior of the crowd in one way or another to improve the efficiency of the system.

From the design standpoint, one could change the size of the Jamarah ring itself. The circumference of a circle increases linearly with the radius, so in order to allow more pilgrims to simultaneously perform stoning at the Jamarah, one could increase the radius of the ring. The simulation model was used to explore the maximum throughput of a system when the ring has a 12-m radius, instead of the present 8 m. The maximum throughput was found to be about 116,000 pilgrims/hr when loading from all directions and 65,000 pilgrims/hr when loading from one side only. Thus, one way to increase the service rate of the system is to increase the size of the Jamarah ring itself. However, before considering the implementation of such a strategy, the effect of enlarging the size of the ring on the surrounding physical geometry, and movement within it, should be carefully evaluated. In addition, the height of the stoning monument should be increased accordingly to keep functioning as an orientation mark for the crowd.

Another possible physical design strategy is to change the shape of the Jamarah ring into an elliptical one. Using an ellipse instead of a circle with the same area would provide a larger circumference, thereby allowing more pilgrims to simultaneously perform stoning. For each spatial loading pattern, two simulation runs were performed using an ellipse with major and minor radii of 1.33 and 0.75 times the radius of the circular ring with equivalent area, respectively. It was found that the throughput would only improve slightly (3 to 5 percent) at the expense of increasing the prevailing average concentration level. However, such shape is expected to encourage more lateral movement of pilgrims and thus more uniform distribution of pilgrims around the ring, provided that the major axis of the ellipse is aligned along the centerline of the Jamarat bridge. One could also enlarge the stone monument, located at the center of the elliptical ring, to provide a wider target at the long sides of the ellipse, therefore encouraging further lateral spread of the pilgrims. These two strategies, which involve changes to the physical system, would only be applicable if they are in accordance with the religious rules governing the stoning process. It is the understanding of the authors that such changes are permissible, but opinions of the appropriate religious scholars should be obtained before implementing such strategies.

Even if the current physical layout of the system is kept unchanged, pilgrims approaching the Jamarah could be encouraged to perform stoning sooner than is currently done (i.e., having a larger fraction of Class 1 users switch from a given distance from the ring). This can be reflected in the model by a different value of the parameter γ in the relation: $\mathrm{FR}[r_i, K(i,j)] = 1.005\ (r_i)^{\gamma}[K(i,j)]^{2.341}$ (Equation 8). In the current model, $\gamma = -0.793$, but using values of -0.5, -0.3, and -0.1 instead yield maximum throughputs of 87,000, 98,000, and 110,000 pilgrims/hr, respectively, when loading from all directions; and 51,000, 59,000, and 65,000 pilgrims/hr, respectively, for the one-side only loading pattern. In practice, this could be implemented through educating pilgrims to stone sooner, and possibly encouraging them to do so on the spot through a public address system. Also, one could elevate the position of the throwers by slightly inclining the surface towards the Jamarah ring. One might also consider encouraging pilgrims to stone faster (i.e., lower TRAJM). Theoretically, such strategy could result in a dramatic increase in the system throughput. However, it would be extremely difficult to implement in practice because pilgrims who have made it through the crowd all the way to their stoning position would like to spend enough time to perform stoning successfully. This reluctance could be further appreciated when one realizes that the stoning time (average of 27 sec) is already small when compared to the time it takes a pilgrim to push through the crowd, which was observed to range from 4 to 10 min.

Finally, it was shown earlier that in the extreme case the throughput increased from 44,000 to 74,000 pilgrims/hr when the distribution of pilgrims changed from being semiuniform to being uniform around the Jamarah ring. This process suggests that encouraging more lateral movement of pilgrims would result in better overall system performance. Several simulation runs were performed by varying each of the parameters of the lateral movement relation (Equation 4). In these runs, values of 1,700, 1.0, and 0.1 persons/m^2 for the parameters α, β, and $\Delta K^{m}_{\min}$, respectively, were used instead of the previously reported values of 850, 0.005, and 1.0 persons/m^2, respectively. The change in the values of each of these three parameters resulted in higher system throughput, lower average concentration on the side where loading is taking place, and higher concentration on the other side of the ring. In addition, these changes resulted in better use of the space around the Jamarah and faster evacuation of the system after the input flow of pilgrims was shut off. In practice, this could be implemented through providing the pilgrims with real-time video information (e.g., through a giant elevated high-resolution screen indicating the concentration profiles around the ring) regarding the availability of vacant or less-congested positions at the other side of the Jamarah where no loading is taking place.

CONCLUSION

The model presented provides a useful framework for the simulation of crowd behavior and movement, and for the analysis of design features and control schemes for the effective management and operation of the Jamarat system. The need for such a modeling capability and for the development and application of sound scientific principles for crowd management problems was recently heightened by the tragic events in which over 1,400 people perished in the latest pilgrimage season in July 1990 (*34*).

The model consists of three fundamental processes: radial movement, lateral movement, and stoning process characterization (switching mechanism). Mathematical representations were developed for each of these processes, and the parameters of the principal relations were calibrated using actual observations. However, these observations were somewhat limited, and may not have covered a sufficiently wide spectrum of conditions that might be encountered in the system. These limitations affect the three processes of interest to varying degrees. For instance, there is little confidence at the present time in the parameter values of the lateral movement model. More extensive data collection is necessary for a more definitive model of lateral movement, especially because the limited data available seemed to suggest the existence of possible directionality reflecting users' preferred movement directions. On the other hand, the speed-concentration model developed here for bidirectional crowd movement may be robust. In particular, the conclusion that the impedance of facility users going in their own direction is about twice that of those going in the opposite direction appears to be rather.well supported by the available data. In addition, it corresponds to the authors' own experience in other heavily crowded conditions. Nevertheless, it is not clear that such a conclusion can be extrapolated to all situations regardless of the relative mix of the two user classes. Relations and parameter values for the third process, namely the stoning process, are also based on a sufficient number and range of observations to justify a reasonable degree of confidence in the results.

While motivated and presented for the Jamarat system, the modeling approach developed here is general and should essentially be applicable to a variety of heavy crowding situations. Of course, the application of any particular situation would require customization in terms of defining a spatial discretization scheme and identifying the principal classes of facility users. Of the relations developed for the Jamarat system, the bidirectional speed-concentration model is likely to be the most directly applicable in other situations, though some adjustment in parameter values would likely be necessary to reflect different cultural norms. It is hoped that additional effort will be directed at crowd phenomena in a variety of settings in order to build a body of scientific knowledge and engineering procedures to enhance the quality of the human experience in such settings and improve the safety and efficiency of the associated operations.

REFERENCES

1. J. J. Fruin. Crowd Disasters—A System Evaluation of Causes and Countermeasures. *Proc., Experts' Workshop on Crowd Ingress to Places of Assembly*, F. T. Ventre et al. (eds.), National Bureau of Standards, Gaithersburg, Md, 1981.
2. *Masarat Al-Mushat fi Al Mashair Al-Muqaddasah (Pedestrian Footways in the Holy Rites)*. Hajj Research Center, Umm Al Qura University, Makkah, Saudi Arabia, 1982.
3. *Al-Jamarat: Dirasat Alwada' Alrahin Wa Was'ail Tatweeruh (Jamarat: Study of Current Conditions and Means of Improvement)*. Hajj Research Center, Umm Al Qura University, Makkah, Saudi Arabia, 1984.
4. B. Hankin and R. Wright. Passenger Flow in Subways. *Operational Research Quarterly*, Vol. 9, No. 2, 1958, pp. 81–88.
5. D. Oeding. Verkehrsbelastung und Dimensionierung von Gehwegen und anderen Anlagen des Fussgangerverkehrs (Traffic Loads and Dimensions of Walkways and other Pedestrian Circulation Facilities). *Strassenbau und Strassenverkehrstechnik*, Vol. 22, Bonn, West Germany, 1963.
6. V. M. Predtechnskii. Dvizhenie Lyndkikh Potokovi Operadeknie Razmerove Kommunikatsionnykh Pomeshch Enii. Glov v Uchenbnike, Arkhitekura Grahdanskikh i Promyshlennykh Zdanii (Pedestrian Traffic Flow and the Determination of Passage Dimensions in Buildings. Chapter: Architecture of Public and Industrial Buildings), Stroiizdat, Moscow, 1966.
7. S. J. Older. Movement of Pedestrians on Footways in Shopping Streets. *Traffic Engineering and Control*, Vol. 10, No. 4, 1968, pp. 160–163.
8. F. Navin and R. Wheeler. Pedestrian Flow Characteristics. *Traffic Engineering*, Vol. 39, No. 9, 1969, pp. 30–33, 36.
9. J. J. Fruin. *Pedestrian Planning and Design*. Metropolitan Association of Urban Designers and Environmental Planners, New York, 1971.
10. A. Ceder. An Algorithm to Assign Pedestrian Groups Dispersing at Public Gatherings Based on Pedestrian-Traffic Modelling. *Applied Mathematical Modelling*, Vol. 3, April, 1979, pp. 116–124.
11. V. M. Predtechnskii and A. I. Milinski. *Planning for Foot Traffic Flow in Buildings*. (Translated from Russian, Stroiizdat Publishers, Moscow, 1969.) Amerind Publishing Co., New Delhi, India, 1978.
12. O. E. Klapp. *Currents of Unrest*. Holt, Rhinehart, and Winston, 1972.
13. J. L. Freedman. *Crowding and Behavior*. W. H. Freeman and Company, San Francisco, Calif., 1975.
14. M. Wolf. The Behavior of Pedestrians on 42nd Street, New York City. City University of New York, 1970.
15. J. Wolpert and D. Zillmann. The Sequential Expansion of a decision Model in a Spatial Context. *Environment and Planning*, Vol. 1, 1969, pp. 91–104.
16. I. B. Stilitz. The Role of Static Pedestrian Groups in Crowded Spaces. *Ergonomics*, Vol. 12, No. 6, 1969, pp. 821–839.
17. I. B. Stilitz. Pedestrian Congestion, RIBA Publication 61–72. *Architectural Psychology*, D. V. Canter (ed.), London, 1970.
18. A. E. Baer. *A Simulation Model of Multi-Directional Pedestrian Movement Within Physically Bounded Environments*. Report 47, Institute of Physical Planning, Carnegie-Mellon University, Pittsburgh, Pa., 1974.
19. W. Boles. Planning Pedestrian Environment: A Computer Simulation Model. *Proc., 4th National Seminar on Planning, Design, and Implementation of Bicycle and Pedestrian Facilities*, 1975.
20. K. Hirai and K. Tarui. A Simulation of the Behavior of a Crowd in Panic. *Proc., 1975 International Conference on Cybernetics and Society*, San Francisco, Calif., 1975, pp. 409–411.
21. F. H. Harlow and D. L. Sandoval. Human Collective Dynamics: The Mathematical Modeling of Mobs. Report LA-10765-MS. *Mathematical Modeling of Biological Ensembles*, Los Alamos National Laboratory, Los Alamos, N. Mex., 1986, pp. 31–49.
22. P. Endean. *Pedestrian Movement In and Out of the Holy Areas of Arafat, Muzdalifah, and Mina during the 1395 (1975) Hajj*. Hajj Research Center, Umm Al Qura University, Makkah, Saudi Arabia, 1977.
23. P. Endean. *The Pedestrian Movement During the Nafrah (Departure from Arafat to Muzdalifah)*. Hajj Research Center, Umm Al Qura University, Makkah, Saudi Arabia, 1977.
24. I. J. Gibson. *An Analysis of Tawaf, Sa'ee, and Umra for the 1396 (1976) Hajj*. Hajj Research Center, Umm Al Qura University, Makkah, Saudi Arabia, 1977.
25. E. Haug, B. Gawenal, and M. A. B. Rasch. *Optimization of the*

Tawaf Capacity of the Mataf Area. Islamic Agency for Technology, 1987.

26. *Dirasat Harakat Al-Hujjaj fi Al-Masa'a (Study of Pilgrims' Movement During Performing Sa'ee)*. Hajj Research Center, Umm Al Qura University, Makkah, Saudi Arabia, 1983.
27. *Dirasa'at Al-Masjid Al-Haram: Dirasat Al-Harakah fi Al-Mataf (The Sacred Mosque Studies: A Study of the Movement in the Tawaf Area)*. Hajj Research Center, Umm Al Qura University, Makkah, Saudi Arabia, 1987.
28. *Counting Pilgrims on the Jamarat Bridge During 1984 Pilgrimage*. Hajj Research Center, Umm Al Qura University, Makkah, Saudi Arabia, 1985.
29. A. H. Al-Rabeh, and S. Z. Selim. An Analysis of Congestion in the Jamarat Area. *Proc., 9th National Computer Conference*, Riyadh, Saudi Arabia, 1986.
30. S. A. H. AlGadhi, and H. S. Mahmassani. Modelling Crowd Behavior and Movement: Application to Makkah Pilgrimage. *Proc., 11th International Symposium on Transportation and Traffic Theory*, M. Koshi (ed.), Yokohama, Japan, July, 1990, pp. 59–78.
31. S. A. H. AlGadhi. *Characterization of Crowd Behavior and Movement*. Ph.D. dissertation, Department of Civil Engineering, The University of Texas, Austin, 1990.
32. D. C. Gazis, R. Herman, and G. H. Weiss. Density Oscillations Between Lanes of a Multilane Highway. *Operation Research*, Vol. 10, 1962, pp. 658–667.
33. L. C. Edie, and R. S. Foote. Effect of Shock Waves on Tunnel Traffic Flow. *Proc. HRB*, 1960, p. 39.
34. The Washington Post, July 3, 1990.

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