1225

TRANSPORTATION RESEARCH RECORD

Highway Capacity, Flow Measurement, and Theory

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL
WASHINGTON, D.C. 1989

Transportation Research Record 1225
Price: $21.50

mode
1 highway transportation

subject area
55 traffic flow, capacity, and measurements

TRB Publications Staff
Director of Publications: Nancy A. Ackerman
Senior Editor: Edythe T. Crump
Associate Editors: Naomi C. Kassabian
Ruth S. Pitt
Alison G. Tobias
Production Editor: Kieran P. O'Leary
Graphics Coordinator: Karen L. White
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.

Highway capacity, flow measurement, and theory.
p. cm.—(Transportation research record, ISSN 0361-1981 ; 1225)
ISBN 0-309-04952-0
1. Highway capacity. 2. Traffic flow. I. National Research Council (U.S.). Transportation Research Board. II. Series.
TE7.H5 no. 1225
[HE336.H48]
388 s—dc20
[388.3 142] 90-5819
CIP

Sponsorship of Transportation Research Record 1225

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

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

Committee on Application of Emerging Technology
Chairman: Earl C. Shirley, California Department of Transportation
Kenneth R. Agent, Roy E. Benner, William Bowlby, Wiley D. Cunagin, Donald R. Curphey, C. Page Fisher, Thomas A. Fuca, Richard C. Ingberg, Terry M. Mitchell, Kenneth S. Opiela, Justin J. Rennilson, Simon Slutsky, Walter A. Winter, David C. Wyant

GROUP 3—OPERATION, SAFETY, AND MAINTENANCE OF TRANSPORTATION FACILITIES
Chairman: James I. Taylor, University of Notre Dame

Facilities and Operations Section

Committee on Communications
Chairman: Frank J. Mammano, Federal Highway Administration, U.S. Department of Transportation
Secretary: John S. Ludwick, Jr., Mitre Corporation
John Auzins, E. Ryerson Case, Wiley D. Cunagin, Robert L. French, Charles J. Glass, Gerard J. Kerwin, Richard C. Lavigne, Robert W. Maher, Corwin D. Moore, Jr., Jim Morton, Umit Ozguner, Michael A. Perfater, John J. Renner, Richard E. Stark, S. J. Stephany, Charles F. Turner, Carroll F. White, Harold Williams, Philip Zove

Committee on Traffic Flow Theory and Characteristics
Chairman: Carroll J. Messer, Texas Transportation Institute
Secretary: Edmund A. Hodgkins, EAH and Associates
Siamak A. Ardekani, James A. Bonneson, E. Ryerson Case, John W. Erdman, Nathan H. Gartner, Fred L. Hall, Douglas W. Harwood, Richard L. Hollinger, Michael Kyte, Edward Lieberman, Feng-Bor Lin, David Mahalel, Hani S. Mahmassani, Patrick T. McCoy, Panos G. Michalopoulos, Harold J. Payne, A. Essam Radwan, Paul Ross, Nagui M. Rouphail, Steven R. Shapiro, Charles E. Wallace, James C. Williams, Sam Yagar

Richard A. Cunard, 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, 1988.

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

Transportation Research Record 1225

Contents

Foreword

The papers that compose this Record are related by their focus on highway capacity, traffic flow measurement, or traffic flow theory. However, the papers cover a wide range of problems reflecting the concerns of both the theoreticians and the practitioners.

The use of stopped delay as a measure for evaluating the quality of traffic flow is discussed in four papers. Teply and Teply and Evans discuss stopped delay and methods for measuring stopped delay to evaluate the performance of traffic signal systems. Lin discusses the reliability of the *Highway Capacity Manual* (HCM) procedure for estimating stopped delays at signalized intersections and proposes modifications to the HCM technique. Hagen and Courage compare the HCM delay computations with the corresponding computations performed by the Signal Operations Analysis Package and TRANSYT-7F, and they present an alternative delay model based on a deterministic queuing process.

The application of chaotic behavior as a means of understanding the discontinuous behavior and nonlinearity in the generalized car-following equation is examined by Disbro and Frame. Gilchrist and Hall examine the relationship between speed, flow, and occupancy in three-dimensional space and suggest that conventional traffic flow theory is insufficient to explain the data and that the catastrophe theory model of uninterrupted traffic flow provides better agreement. Banks also examines the speed, flow, and occupancy relationships in freeway traffic flow and provides more evidence and interpretations based upon data from San Diego.

Signalized intersection operations are examined in papers by Rouphail and by Levinson. Rouphail presents analytical models for the estimation of progression adjustment factors to delays at signalized, coordinated intersection approaches. Levinson proposes a formula for estimating the capacities of shared left-turn lanes.

Tsay and Wang describe a new search method they have developed for TRANSYT-7F to reduce computational time and improve the TRANSYT-7F performance index.

Freeway weaving is examined in papers by Skabardonis et al. and Cassidy et al. Skabardonis examines the application of the INTRAS simulation model to freeway weaving areas. Cassidy examines six existing methods for the design and analysis of freeway weaving sections and compares the results with data collected at weaving locations in California.

Kyte and Marek present the results of a study of single-lane approaches to all-way stop controlled intersections and propose a methodology for analyzing their operational performance. Zion et al. look at three different models for determining the capacity and delay at four-way stop sign intersections and examine the validity of these models based on field data.

The two papers that conclude this Record examine the use of video image processing as a means for the automatic collection of traffic data. Taylor et al. present case studies of a freeway and a parking lot that utilized video-based equipment and automatic data-extraction techniques to determine vehicle speeds and headways, which were then compared to data collected by traditional methods. Michalopoulos et al. describe a video detection system that emulates loop detectors and can be used for automatic surveillance and real-time extraction of traffic data.

Whether the reader is a city traffic engineer trying to determine the capacity of an all-way stop intersection or a traffic flow theoretician pondering the vagaries of the traffic flow equations, the papers in this Record should be both interesting and informative.

Evaluation of the Quality of Signal Progression by Delay Distributions

S. TEPLY AND G. D. EVANS

The quality of service provided by signalized intersections can be expressed in a number of ways. The volume-to-capacity ratio, the probability of discharge, the load factor, the average delay, the number of stops, the length of queues, or composite measures such as level of service or cyclic flow profiles are frequently used. Among these criteria, the number of stops and the average delay have a special importance for coordinated systems because they are directly perceived by drivers. As a result, they greatly affect public acceptance of the system design. "Delay" is not a simple or exact term. It is defined as the difference between the actual travel time and the unimpeded travel time through a roadway section upstream of (and sometimes including) a signalized intersection. Driver expectations and behavior, speed distributions, and the arbitrary nature of the distance reference are some of the factors influencing the delay. As a result, its measured values are difficult to duplicate. Nevertheless, delay not only remains the most important signal performance criterion but can also be used as an indicator of the quality of signal coordination. The basic idea is simple: when the vehicular progression through a traffic signal is good, most vehicles will encounter no delay. On the other hand, if most platooned vehicles are stopped and delayed, the coordination of signals is suspect. This idea has been employed in a survey system developed at the University of Alberta. Portable field equipment and subsequent computer programs assist in the evaluation of signal coordination by using delays at individual intersections along a coordinated route. The survey method is based on a re-creation of a time-space diagram for a signalized intersection approach lane. The spatial reference extends from an arbitrary point upstream of signal queues to the stopline. The difference between the actual and unimpeded travel times determines individual vehicle delays. Individual vehicle information is summarized in the form of delay distributions and cyclic flow profiles. This paper describes the survey and employs practical examples to illustrate its use.

The quality of signal progression is one of the major determinants of delay magnitude. Naturally, this statement applies only to situations where traffic volumes are below capacity, since oversaturation is characterized by growing queues and, consequently, the stopping of all platoons.

TRB's *Highway Capacity Manual* (*1*) identifies five different arrival types for intersection approaches. They are characterized by flow patterns at the approach (platooning or random) and by their relationship to green or red signal intervals. Because it is difficult to quantify arrival type precisely, the manual recommends the use of an indicator in the form of a platoon ratio:

$$R_p = \frac{PVG}{PTG}$$

S. Teply, Department of Civil Engineering, University of Alberta, Edmonton, Alberta, Canada, T6G 2G7. G. D. Evans, University of Alberta, Edmonton, Alberta, Canada, T6G 2G7. Current affiliation: Transportation Department, City of Edmonton, 9803 102A Avenue, Edmonton, Alberta, Canada, T5J 3A3.

where *PVG* is the percentage of all vehicles arriving during the green interval and *PTG* is the green-to-cycle time ratio expressed as a percentage. Various arrival types can then be identified by their platoon ratios. For example, the most favorable arrival condition—in which most, if not all, vehicles arrive during the green interval—is characterized by a platoon ratio in excess of 1.5. On the other hand, a platoon ratio between 0 and 0.5 is typical of an approach on which most vehicles arrive during the red interval, with 0 indicating that no vehicles arrived during green. Arrival types are then linked to progression adjustment factors, which are used to adjust the value of delay (determined from a formula).

In the *Canadian Capacity Guide for Signalized Intersections* (*2*), only individual intersections are considered; consequently, no adjustments to the delay formula are used. Instead, the guide recommends that the whole route or network operation be examined after a local intersection is designed or analyzed. Such an examination could include cyclic flow profiles (*3*) or other tools to identify arrival and discharge patterns and relationships. For instance, the city of Edmonton evaluates signal control schemes prior to their implementation using a modified version of RCOORD (*4*) for route analysis and TRANSYT (*5*) for networks.

No simple field test exists for the "as-is" coordination of traffic signals. In the past, the evaluation and adjustment of implemented route or network progression schemes have relied on brief observations, floating car surveys, or driver complaints. All of these methods, however, have practical drawbacks. Observations that have not been quantified are unreliable and require review by an experienced engineer to be meaningful. Similarly, although complaints may be valid and may provide the initial impetus for an examination of the quality of coordination, they do not form an adequate basis for engineering analysis. Moreover, since even the best coordination schemes are a result of tradeoffs among conflicting objectives, some complaints can always be expected. Floating car surveys are time consuming, expensive, and previously have had statistically questionable results.

THE DELAY PRINCIPLE

The objective of this research was to develop a practical system for evaluating signal coordination at a series of intersections. It appeared that delay could be employed to achieve this objective.

Individual vehicles arriving during a signal cycle encounter different durations of delay. The three most important factors are the time when they arrive relative to the beginning of the

green interval, the rate of their arrival, and the rate of their discharge (see Figure 1).

In the situation illustrated in Figure 1A, which features low arrival rate and high discharge rate (saturation flow), the main factor determining the delay of individual vehicles is the time of arrival. Figure 1B is similar but shows a very low saturation flow, which causes delays to vehicles that could proceed unimpeded in Figure 1A. Figure 1C features a high arrival rate, which also forces delays on vehicles that arrive during the green interval. The latter two cases illustrate the need to supplement platoon ratios with additional information.

A survey system involving the interaction of arrival rate, discharge rate, and signal timing can be employed to determine individual delays. These delays, which depend on the progression of groups of vehicles through a signal cycle, can then be used as indicators of the quality of signal coordination.

In an ideal coordination case, all vehicles would arrive during the green interval and the flow rate of the platoon would not exceed the saturation flow. As a result, no vehicle would encounter a delay (see Figure 2A). The probability distribution of individual delays would be simple, with all vehicles falling into the minimum or zero delay class (see Figure 2B).

If, on the other hand, the arrival pattern remained unchanged but the signal operation were reversed (in other words, the red signal were in operation instead of the green signal), all vehicles would suffer from delays (see Figures 2C and 2D).

The distributions of delay shown in Figure 3 represent more realistic situations. Figure 3A shows what can be expected at signal approaches that provide excellent progression. In this example, most vehicles (the main platoon) pass through without delay. A secondary platoon of "on-turning" vehicles (those that turn into the coordinated route from side streets in a random fashion) experience a range of delay durations, as evidenced by the wide distribution. Figures 3B and 3C show a deterioration of the progression: fewer vehicles experience zero delays and more vehicles encounter increasing delays. Figure 3D represents an intersection approach in which the main platoon is apparently stopped during the red signal and the secondary flow proceeds unimpeded during green.

Since most of the delays in Figure 3 are experienced by vehicles that are stopped, the number of stops was considered to be a major concern.

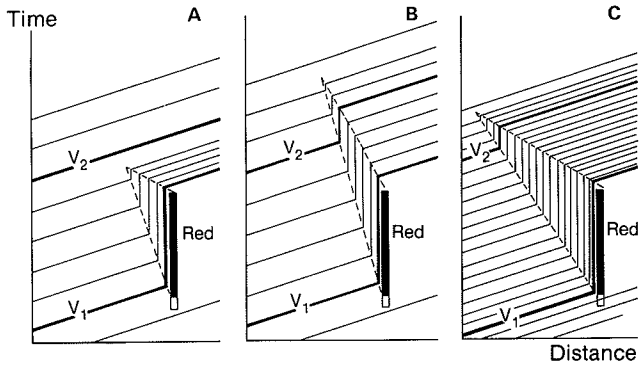
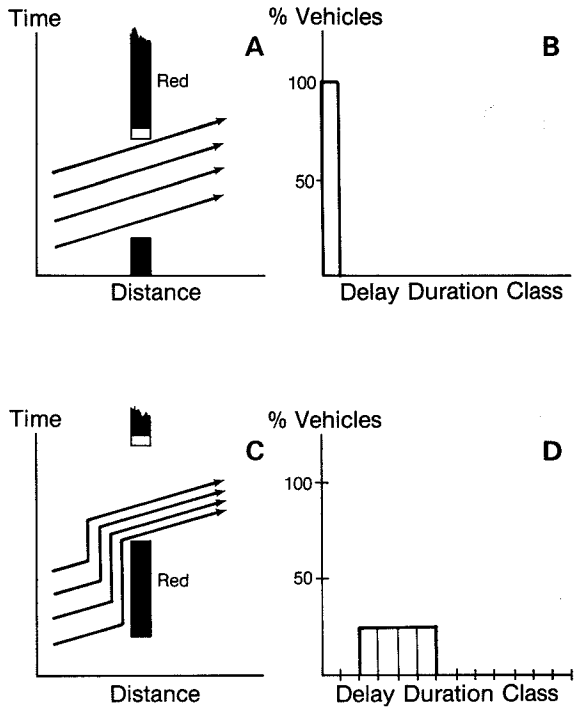


FIGURE 1 Examples of the influence of arrival rate and saturation flow on delay: A, low arrival rate/high saturation flow = short average delay; B, low arrival rate/low saturation flow = long average delay; C, high arrival rate/high saturation flow = long average delay.



NOTE: In Example B, all vehicles fall into the "no-delay" class. Example D has no vehicles in the "no-delay" class since all vehicles suffer some delay.

FIGURE 2 Examples of the influence of arrival pattern on delay: A, arrivals during green interval only; B, the corresponding delay distribution; C, arrivals during red interval; D, corresponding delay distribution.

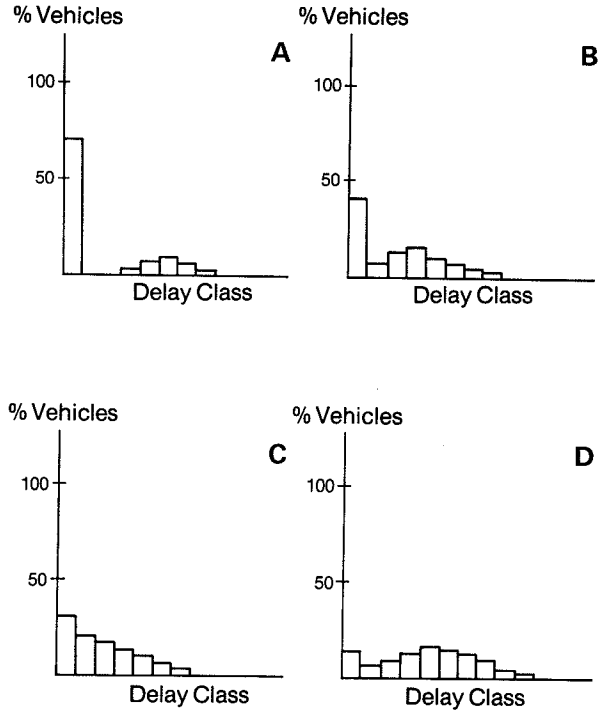


FIGURE 3 Examples of delay probability distributions for four signal progression conditions: A, excellent; B, good; C, fair; D, poor.

As a result of these considerations, delay distributions on individual intersection approaches or approach lanes were the main tool used in this research to assess the quality of signal coordination.

It should be noted that this form of investigation deals with the problem of coordination strictly on an individual intersection basis. As a result, the quality of the whole route or network can be judged only by an examination of all approaches. A single approach with a poor delay distribution may not signify bad design. Good coordination designs frequently sacrifice some approaches to achieve a global benefit of short delays or a minimum number of stops.

SURVEY METHOD

The survey system used in this research is shown in Figure 4. At an intersection approach lane, a reference distance is identified that extends beyond the reach of usual queues. This distance is defined by the stopline (B) and a prominent point upstream (A).

The survey consists of recording the following three time series:

1. The passage of vehicles over point A;
2. The discharge of vehicles across stopline B; and
3. The beginning and end of the green interval.

Information is gathered manually on a classification counter that has been modified with additional keypads. The data is subsequently transferred to a microcomputer. Manual data entry by two surveyors has proven to be a fast and sufficiently accurate mode of operation. All vehicles are entered at the times when they pass the reference points. With the exception of the first and last vehicle in the survey, no matching of arrivals and discharges is required. An occasional miss of green interval entries is not critical, even in the event of traffic-actuated phases, since the data manipulation program can detect these omissions.

The survey consists of a set of individual intersection approaches along a route under investigation. A single "typical" lane (which may be difficult to identify) is generally sufficiently representative of arrival conditions. Since the data collection time required for each approach is relatively short, it is not necessary to survey all approaches in one direction simultaneously. It is usually possible to collect data for two intersection approaches within an hour, provided traffic conditions remain reasonably similar. Naturally, cycle times, green intervals, and offsets in fixed time coordinated systems must not be changed during the survey period. Under these conditions, the surveys along one route can employ the same time of day periods for several days.

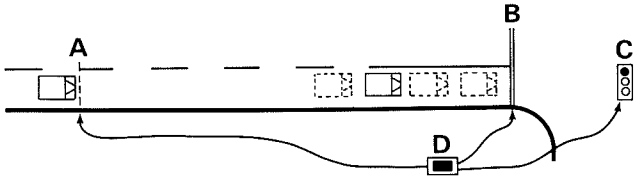


FIGURE 4 A typical layout of the survey system: A, arrival detection point; B, discharge detection point; C, traffic control signal; D, data acquisition device.

SOFTWARE SYSTEM

Data stored on location include a site and survey identification, a vehicle arrival and discharge time series (with a possible classification of vehicles), and a signal interval series. Stored data from all surveys are transferred to a microcomputer. The transfer software provides for both a visual and an automatic inspection on the screen. Illegal data items are stored in a separate file for future reference in the debugging process.

The data are then analyzed on the microcomputer. The main software system performs the following tasks:

- Checking and testing;
- Adjustments to the matching of arrivals and departures;
- Determination of unimpeded travel time;
- Calculation of the delays of individual vehicles;
- Sorting and classification of delays;
- Calculation of aggregate delay statistics, checking, and testing; and
- Sorting and classification of arrivals and departures.

Checking and Testing

The purpose of this software is to match vehicles arriving at the upstream reference point with vehicles discharged at the stopline. Typical errors, such as an arrival/departure imbalance or unusual speed values, are identified and communicated to the user. Negative or extremely high speed values, for example, would indicate a missing vehicle upstream; an unusually low speed may indicate a missing discharge record.

Adjustments to Matching

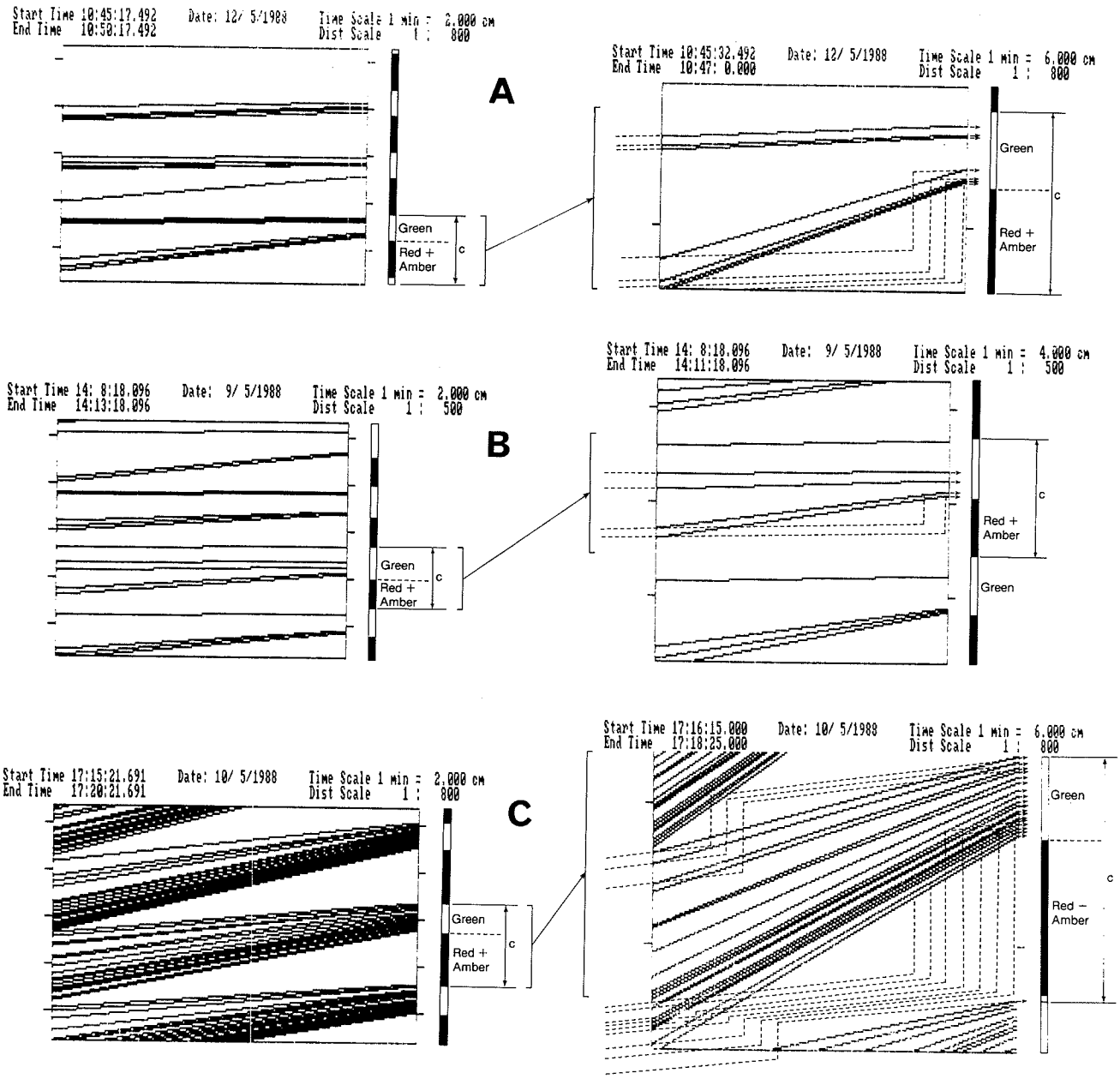
In most cases, some adjustments are necessary to match the arrival/discharge pairs. This task is performed by visually inspecting the simplified time-space diagrams produced by the computer software (see Figure 5). The screen can be scrolled up or down and the scales adjusted to "zoom" on problem areas. Irregularities in the pattern are readily apparent, and erroneous vehicles are deleted. These infrequent corrections do not appear to introduce significant errors to the final delay distribution.

Determination of Unimpeded Travel Time

The time-space display makes it possible to identify those vehicles that proceed through the survey section without delay. The average speed of these vehicles is calculated and can be considered by the user as the unimpeded travel time.

Calculation of Delays

The total travel time of individual vehicles is calculated as the difference between the matched arrival and discharge times. Individual delays are then determined by subtracting the unimpeded travel time from the total travel time. Negative values are possible for vehicles travelling faster than the speed



NOTE: The graphs on the right are "zoom" enlargements of those on the left. They are supplemented by dotted lines showing the usual form of the trajectories.

FIGURE 5 Examples of simplified vehicle trajectories in time-space diagrams: A, low-volume approach; B, medium-volume approach; C, oversaturation.

input by the user. Such vehicles are assumed to have zero delays.

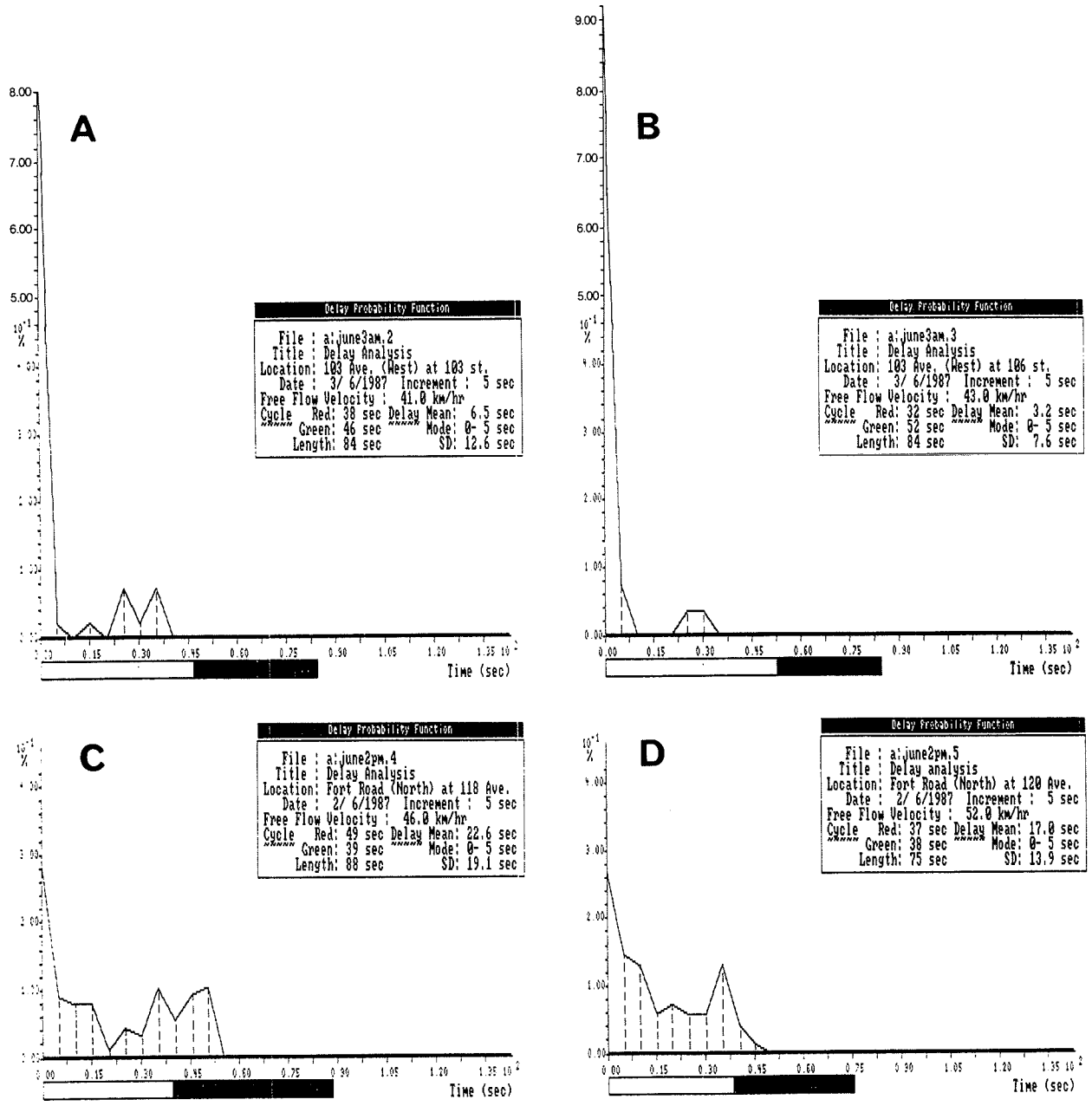
Sorting and Classification of Delays

The user can choose the delay class interval to be used in the determination of delay frequency and delay probability distributions. Intervals of 5 sec are practical in most cases.

Figure 6 shows sample probability distribution diagrams (frequency distribution histograms can also be printed). The first two distribution diagrams illustrate an almost ideal progression at two intersections of a one-way street that has coordinated signals. The second two examples illustrate a very poor signal progression at two consecutive intersections of a major arterial roadway. Only about one-fourth of the vehicles are free-flowing; most experience delays, some of which are considerable. The duration of both major intervals (green and red plus amber) is printed for information, although it is not related to the distribution. The intersections used in these examples operate below capacity since only a few delays exceed the duration of the red interval.

Calculation of Aggregate Delay Statistics

Block inserts in the probability distribution diagrams allow the user to label the survey (by day, location, file/approach,



NOTE: The white and black bars under the diagrams show the length of the green and red intervals as a delay magnitude reference but have no direct relevance to the graphs.

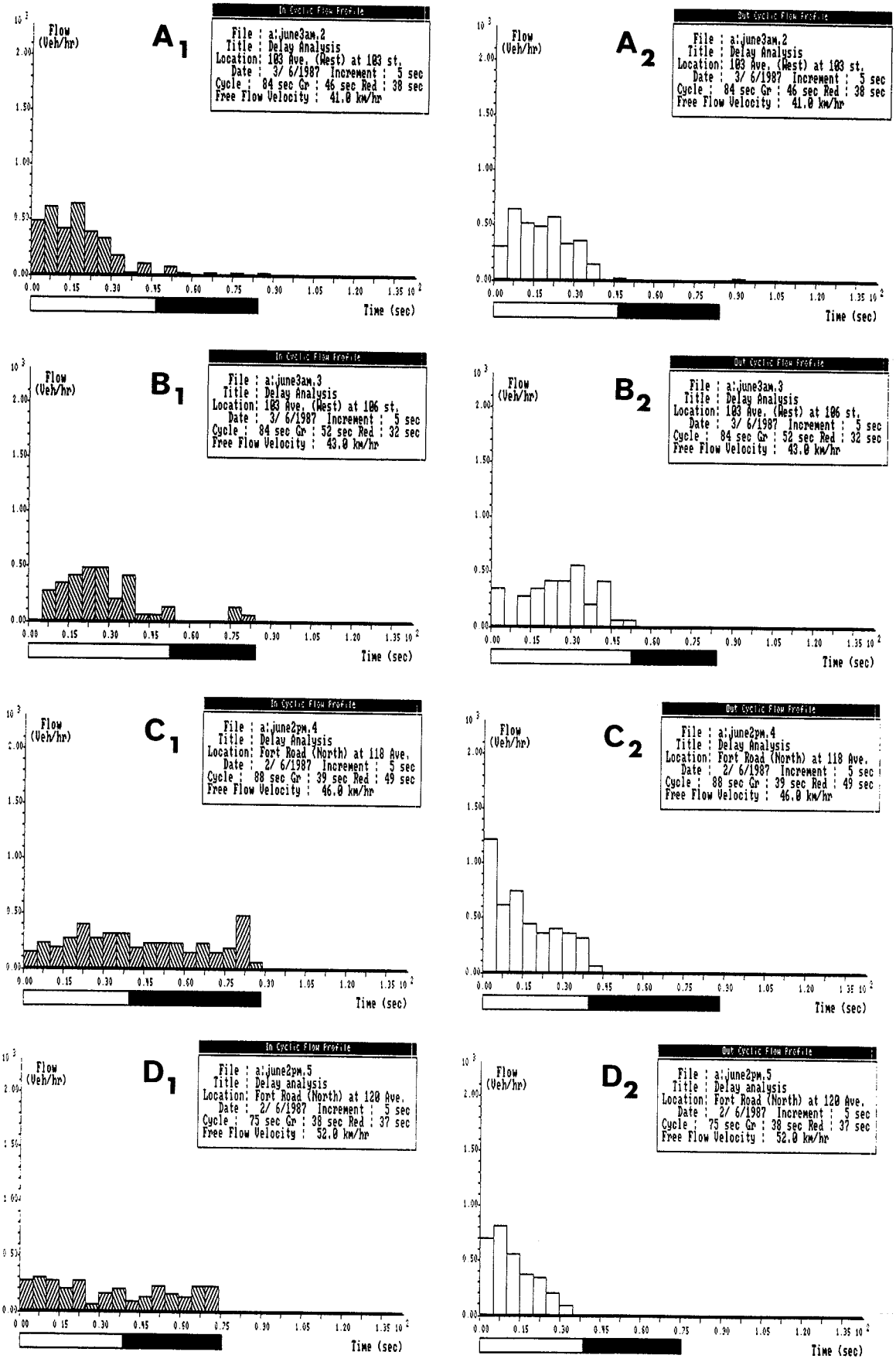
FIGURE 6 Examples of typical delay distributions. A and B—excellent signal coordination; C and D—poor signal progression.

signal timing, and average free flow speed). In addition, the average delay, mode, and standard deviation of the distribution are identified.

The distributions provide a good, relatively condensed illustration of the events in the surveyed lane. Attempts to condense the information further using a combination of average delay, standard deviation, and mode were not successful. Because most of these distributions are bimodal, simple statistical parameters cannot describe them adequately.

Sorting and Classification of Arrivals and Departures

For a detailed analysis, the user can employ a special software routine that organizes the arrivals and departures of individual vehicles into class intervals relative to the beginning of the green interval. Time "slices" of 5 sec are adequate in most cases, although 2-sec slices have been found to be useful for high volume situations.



NOTE: White and black bars below the diagrams relate the arrival or discharge times to the signal cycle.

FIGURE 7 Arrival and discharge cyclic flow diagrams for situations shown in Figure 6: A_1, B_1, C_1, and D_1—arrival patterns; A_2, B_2, C_2, and D_2—discharge patterns.

The arrival and departure information is organized into cyclic flow profiles (*3*), which are familiar from widely used programs such as TRANSYT (*5*). For the delay survey system, cyclic flow profiles for arrivals and departures can be printed either separately or together. These profiles are similar to those used in TRANSYT; however, the delay surveys use a base distance, while most TRANSYT versions queue vehicles directly on the stopline.

Figure 7 shows examples of cyclic flow profiles corresponding to the delay distributions in Figure 6. An examination of the patterns in Figures 7A and 7B illustrates why the progressions in Figures 6A and 6B are so good: virtually all vehicles arrive during the green interval. On the other hand, the arrival pattern along the second route (Figures 6C and 6D) is almost uniform throughout the cycle (Figures 7C and 7D).

The cyclic flow profiles are normalized like the delay probability distributions. The values are given in hourly rates of flow, which eliminate the impact of minor volume fluctuations during different survey periods.

CONCLUSIONS

The system described in this paper provides a practical, effective tool for the evaluation of implemented signal coordination designs. An initial analysis of delay distributions provides a starting point for the assessment. Where more serious problems are detected, cyclic flow profiles can be employed. The surveys are relatively simple and, since the results are presented in a normalized form, comparisons of signal performance at different intersections are easy as long as "typical" lanes can be identified. It is hoped that this system will encourage the use of before-and-after studies for signal coordination. The cities of Edmonton and St. Albert are currently using the system for this purpose.

It appears that a single numerical indicator of the quality of coordinated signal operation—such as the platoon ratio or a mode of the delay distribution—cannot sufficiently describe such a complex issue. Further research is required to supplement the relatively subjective inspection of delay distributions or cyclic flow profiles with more tangible indicators.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Natural Sciences and Engineering Research Council of Canada for providing funds for equipment and for supporting this research by operational grants. Paul Sabourin, a graduate student in the Electrical Engineering Department at the University of Alberta, was responsible for all hardware adjustments and data acquisition programs. H.C. Kua, a graduate student in Transportation Engineering at the same university, was the initial major user of the system and collected examples in St. Albert for this paper. We would also like to thank the reviewers of this paper for their valuable comments.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. D. Richardson, J. Schnablegger, B. Stephenson, and S. Teply. *Canadian Capacity Guide for Signalized Intersections* (S. Teply, ed.). Institute of Transportation Engineers, District 7, Canada, 1984.
3. D. I. Robertson. Cyclic Flow Profiles. *Traffic Engineering and Control*, June 1974.
4. S. Teply, J. D. Hunt, and V. Berka. *RCOORD User Manual.* Transportation Department, City of Edmonton, Alberta, Canada, 1981.
5. D. I. Robertson. TRANSYT Method for Area Traffic Control. *Traffic Engineering & Control*, October 1969.

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

Progression Adjustment Factors at Signalized Intersections

NAGUI M. ROUPHAIL

This paper presents a set of analytical models for estimating progression adjustment factors (PAFs) to delays at signalized, coordinated intersection approaches. The derived models are sensitive to the size and flow rate of platoons, which in turn are affected by the travel time between intersections. The procedure requires data that are readily available from time-space diagrams and flow counts. A comparison of the factors estimated in this study and their *Highway Capacity Manual* (HCM) counterparts reveals the limitations of the HCM method in predicting levels of service for coordinated approaches, especially under excellent or very poor progression scenarios. Finally, an interactive, computerized procedure is presented that carries out the necessary PAF calculations with minimal input requirements.

The research presented in this paper represents an application of a previously developed methodology for modeling traffic flow at coordinated intersections. The paper that describes this methodology (*1*) provides detailed derivations of the delay models used to develop the progression adjustment factors (PAFs).

A distinct feature of the 1985 *Highway Capacity Manual* (HCM) (*2*) is its use of stopped delay as the sole determinant of level of service (LOS) at signalized intersections. A key element in determining delay is the quality of progression afforded to the lane group. An examination of Tables 9.1 and 9.13 in the HCM reveals that progression significantly affects LOS. For example, a lane group operating at the midrange of LOS C under random arrival conditions (Arrival Type 3, Table 9.13) operates at LOS B under favorable progression (Arrival Type 5, Table 9.13) and LOS D under unfavorable progression (Arrival Type 1, Table 9.13) at a volume-to-capacity ratio of 0.60. In other words, there are three possible LOS designations for that lane group based on the degree of signal progression provided.

The PAFs in Table 9.13 of the HCM are expressed as the ratio of vehicle delay under a specific progression scenario to delay encountered with random arrivals. While the factors are applied to the overall delay (i.e., uniform plus random delay components), it is not yet clear from the literature whether the random (or overflow) delay component is truly affected by progression. Earlier studies by Hillier and Rothery (*3*) and Robertson (*4*) indicate that progression has little impact on the overflow delay. This is also evident in the TRANSYT-7F model (*5*), where the random delay component is insensitive to offset variations (see Sadegh and Radwan (*6*)). On the other hand, Akcelik (*7*) suggests reducing the random delay component by 50 percent under favorable progression. Pending future evidence to the contrary, it is assumed that progression effects are limited to the uniform delay component.

Department of Civil Engineering, University of Illinois at Chicago, P.O. Box 4348, Chicago, Ill. 60680.

A schematic of the traffic modeling concept for coordinated lane groups in the HCM is shown in Figure 1. In the figure, vehicle arrivals occur at two different rates in red (q_r) and green (q_g), and the proportion of arrivals that occur in the effective green phase is termed *PVG*. If arrivals are random (Line A), then $PVG = g/C$ or $PVG/(g/C) = 1$, where g equals effective green time in seconds and C equals cycle length. The total uniform delay per cycle is equivalent to the area under the arrival rate curves (the shaded area in Figure 1). It is evident that delay decreases as *PVG* increases, and vice versa. Thus, $PVG/(g/C)$ represents the relative traffic density in green. At a relative density of 1, lane group traffic arrives at random and its relative delay (compared with the isolated case) is also 1. In the HCM, the relative density in green is termed the "platoon ratio" (R_p), and the relative delay is referred to as the "progression adjustment factor." The platoon ratio is subdivided into five ranges, each designating an "arrival type" as shown in Table 9.13 of the HCM.

LITERATURE REVIEW

The rather limited calibration data base upon which the PAFs were estimated has spurred a flurry of research activities to calibrate and validate the factor values in Table 9.13 of the HCM. A comprehensive study is nearing completion at the Texas Transportation Institute that will develop empirical models of PAFs on the basis of delay data collected nationwide (*8*). Jovanis et al. (*9*) have reported the findings of a limited validation effort of PAFs in Illinois. Data obtained from 10 intersection approaches controlled by traffic-actuated controllers indicate (a) that PAFs are lower than those listed in the HCM and (b) that PAFs are much less sensitive to the platoon ratio than indicated. These results are depicted graphically in Figure 2. Thus, while a predictive association exists between PAFs and R_p, the Illinois data suggest that this association is not strong.

A recent paper by Courage et al. (*10*) compares the HCM's estimated PAFs with corresponding values generated in the TRANSYT-7F model (*5*). This evaluation was performed both for a hypothetical link and for an 85-link network. TRANSYT values were derived through a process of linking and delinking, whereby delay values at a given platoon ratio (via offset manipulation) are divided by their counterpart values when the approach is de-linked, or disconnected, from adjacent signals. The study found that, in general, there was good agreement between the two models: the mean percentage deviation between them was 2.36 percent. It was also observed that the platoon ratio as defined in the HCM is a better predictor of delays with heavy traffic volumes and that a wider

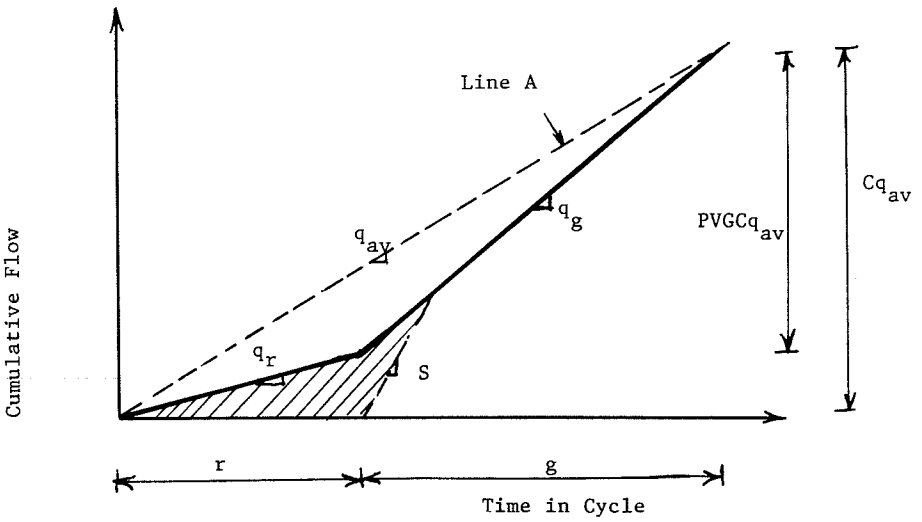


FIGURE 1 Model of progressed traffic in the HCM (*2*).

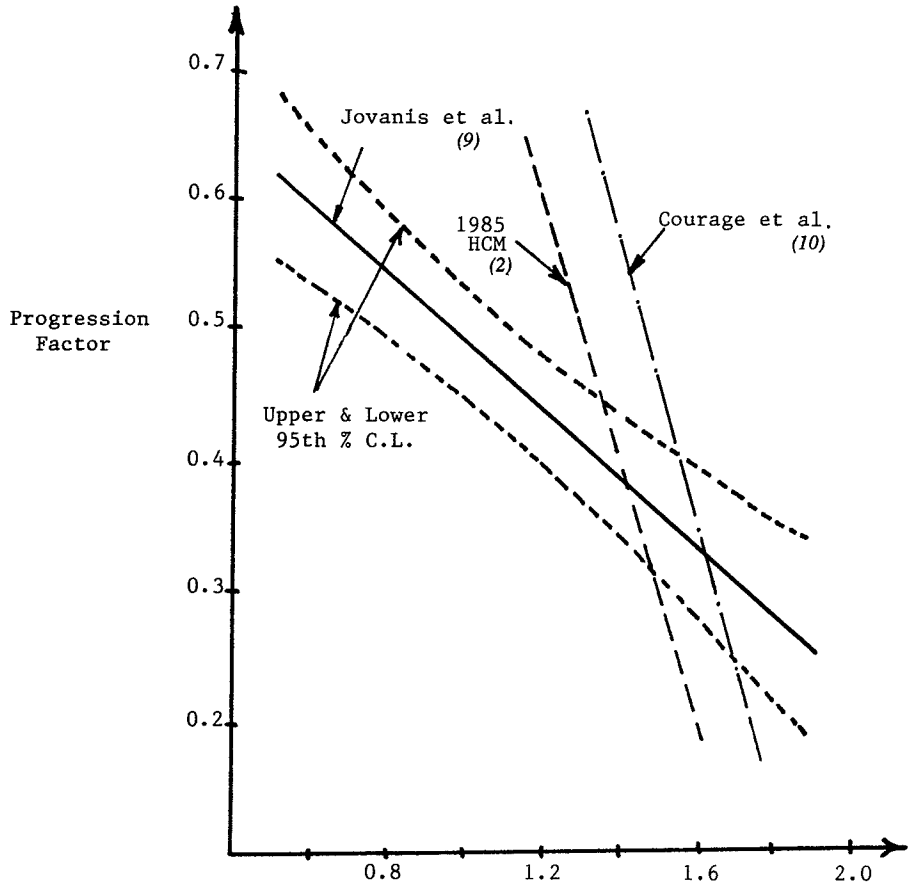


FIGURE 2 Estimation of progression adjustment factors from literature.

range of PAFs exists than the HCM recognizes. Therefore, some extrapolation of the HCM values may be warranted to cover exceptionally good or exceptionally poor progression. The study also considers the use of an estimator for R_p based on information derived from time-space diagrams (discussed in detail later in this paper). The regression line of PAF versus R_p from the Courage et al. study (*10*) is shown in Figure 2 for comparison with the Illinois study.

CRITIQUE OF PLATOON RATIO METHOD

Previous efforts have been directed primarily toward the development of improved PAF estimates based on observations of R_p and other explanatory variables and conforming with the basic structure of the HCM procedure, specifically Table 9.13. While the platoon ratio method is rational and simple, it suffers from many drawbacks, some of which may

explain why good correlations were not observed between PAFs and R_p. For example

- There is no mechanism in the HCM method for estimating delays under projected conditions. The value of *PVG* must be observed in the field before delays can be estimated. This problem was addressed in the study by Courage et al. (*10*), in which a "band ratio" (R_b) estimator is used to predict R_p. The formula for R_b is based on the assumption that the density of platoon arrivals in green is proportional to the bandwidth size and the amount of progressed traffic. Nonplatoon arrivals are assigned proportionally to the remainder of the green for nonprogressed traffic. The comparison of R_p and R_b provided in Figure 3 indicates a very close agreement. Note that all elements of R_b (such as cycle length, bandwidth, green times, and percent flow progressed) can be derived from a typical time-space diagram of the arterial.
- The coarse designations of arrival types generate delay estimates that are insensitive to a wide range of platoon ratios. For example, for a lane group with $g/C = 0.50$, Type 2 arrivals are applicable when 25.5 to 42.5 percent of all arrivals occur in green. Clearly, intermediate delay estimates are appropriate in this case. To resolve this problem, current and future research work will focus on developing continuous estimates of PAFs as demonstrated in the studies by Jovanis et al. (*9*) and Courage et al. (*10*).
- The underlying traffic model in the platoon ratio method (depicted in Figure 4a) is not responsive to delays caused by secondary queues. An alternate modeling concept (see Figure 4b) is to consider traffic as two separate and contiguous streams: a platoon of size B seconds and flow rate q_{pl} vehicles per second (veh/sec), and secondary flows of duration $C - B$ seconds and flow rate q_s veh/sec (*1*). This concept represents a simplification of the TRANSYT model histograms by averaging flow rates in only two distinct segments. Length B can be viewed as the bandwidth size commonly available in time-space diagrams or as the length of the platoon leaving an upstream intersection. An interesting comparison arises between Figures 4a and 4b. Assume in both models that the degree of saturation (X) and *PVG* (or R_p) are identical; thus, a unique PAF value would exist according to the HCM. In Figure 4b, two conditions are illustrated: an early platoon release in which a dense platoon arrives at the beginning of the green (a Type 5 arrival in the HCM), and a late platoon release in which arriving platoons do not interfere with secondary queues accumulated during the red phase. According to the HCM, this would be categorized as a Type 4 arrival (a dense platoon arriving in the middle of the green). Clearly, there is no unique delay solution even though there were no changes in R_p. This example also demonstrates a flaw in the definition of arrival types, since the delay for Arrival Type 5 (the best progression in the HCM) is higher than that for Arrival Type 4. Indeed, it is possible to demonstrate that, unless secondary flows are truly negligible, arrivals of Type 5 are seldom optimal in terms of delay. The difference between the two modeling concepts is more pronounced for relatively short (small B) and dense (high q_{pl}) platoons since the platoon ratio method tends to spread all arrivals in the green and red periods (see Figure 4a).
- The platoon ratio method does not consider the platoon structure. For instance, adjusting offsets and bandwidths on an arterial to achieve maximum progression is ineffective if platoons experience significant dispersion between intersections. The traffic model in Figure 4b allows for this variable by means of an explicit functional relationship between the flow rates (q_{pl} and q_s) and the average platoon travel time (t).

PROPOSED METHODOLOGY

The method proposed in this paper involves three basic steps:

1. The estimation of platoon size and flow rate,
2. The identification and calculation of delay models, and
3. The estimation of PAFs.

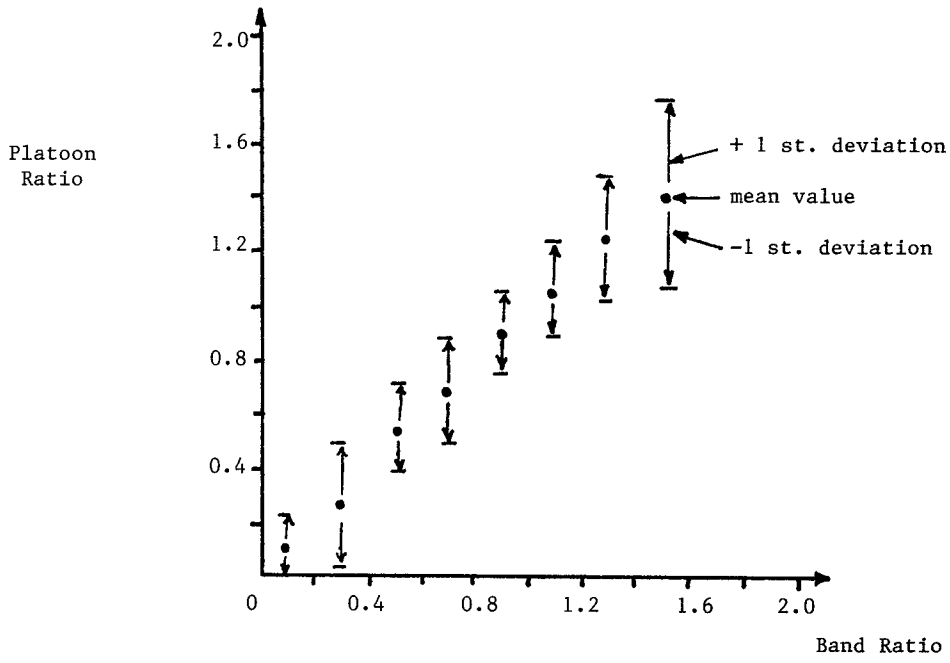
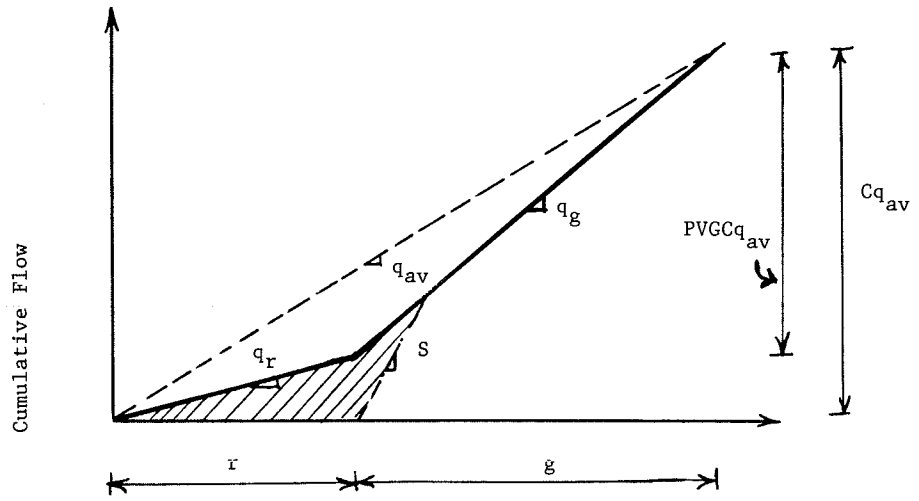


FIGURE 3 Estimation of platoon ratio from band ratio (*10*).

(a) HCM - (Platoon Ratio Method)



(b) Proposed Method

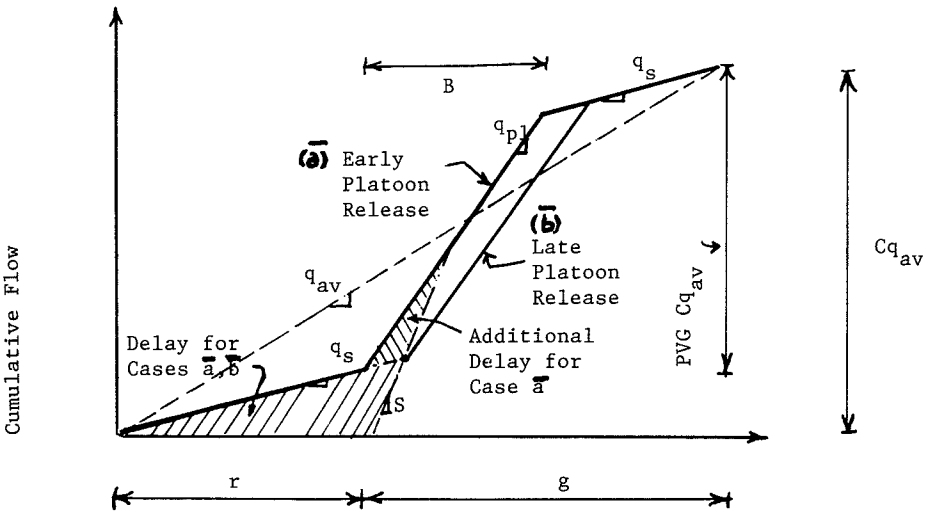


FIGURE 4 Comparison of model concepts.

Estimation of Platoon Size and Flow Rate

Flow rate estimates were derived in a previous paper (*1*) and are summarized below:

$$q_{pl} = q_{av} + (S - \alpha * q_{av}) * \exp(-0.01215 * t) \quad (1)$$

$$q_s = (C * q_{av} - B * q_{pl})/(C - B) \quad (2)$$

$$B = (C - g_u) * \alpha * q_{av}/(S - \alpha q_{av}) \quad (3)$$

where

q_{pl} = average platoon flow rate (veh/sec),
q_s = average secondary flow rate (veh/sec),
q_{av} = average lane group flow rate entering (veh/sec),
α = proportion of traffic that is progressed from the upstream intersection,
t = average platoon travel time (distance/progression speed),
S = saturation flow rate (veh/sec),
B = platoon size (equivalent to the bandwidth size from time-space diagrams or estimated from Equation 3), and
g_u = effective green time at the upstream intersection.

Identification and Calculation of Delay Models

Uniform delay models were derived in a previous paper by Rouphail (*1*) for the entire range of bandwidth, or platoon,

offsets. Offsets measure the time difference between the green start at the subject intersection and the arrival time of the platoon. Offset values thus range from $-r$ (when the platoon arrives r seconds before the green start) to $+g$ (when the platoon arrives g seconds after the green start). Four problem types emerged from those definitions, and they are designated as I,A, I,B, II,A, and II,B based on two selection criteria (see Table 1). Each problem type has a predetermined sequence of delay models for various offset ranges. Delay calculations for the numbered models depicted in Table 1 are given in Appendix I of the source paper (*1*). The paper also provides a thorough evaluation of the generated delay (d_p).

Estimation of Progression Adjustment Factors

In this study, PAFs were applied to both the uniform and random delay components. Random delays, d_r, were assumed to be identical to the HCM value:

$$d_r = 173*X^2*\{(X - 1) + [(X - 1)^2 + 16*X/c]^{\frac{1}{2}}\} \quad (4)$$

where X equals lane group volume-to-capacity (V/C) ratio and c equals lane group capacity (vph).

Uniform delays with random arrivals were also assumed to be identical to their HCM counterpart:

$$d_u = .38*C*(1 - g/C)/(1 - X*g/C) \quad (5)$$

The PAF is then expressed as

$$PAF = (0.76*d_p + d_r)/(d_u + d_r) \quad (6)$$

The 0.76 term converts the approach delays developed in this study to stopped delays consistent with the HCM definition (*6*).

NUMERICAL APPLICATION

The following example provides a comparison of PAF estimates obtained using the HCM method and the study approach.

An intersection approach services a demand rate of 720 vph in the peak period at a saturation level of 1,800 vphg. Cycle length is 60 sec, and effective green time is 30 sec. About 83 percent of this traffic originates from an intersection 1/3 mi upstream of the approach at a travel speed of 40 mph. The effective green time at the upstream intersection was observed to be 20 sec.

For the five arrival types listed in the HCM, the PAFs were obtained using the two methods and compared.

The following variables were defined:

q_{av} = 720/3,600 = 0.20 veh/sec,
S = 1,800/3,600 = 0.50 veh/sec,
t = 5,280/(3*0*1.467) = 30 sec,
α = 0.83,
C = 60 sec,
g = 30 sec, and
g_u = 20 sec.

TABLE 1 DESIGNATION OF PROBLEM TYPE AND DELAY MODELS (*1*)

Problem Type		I,A	I,B	II,A	II,B
Selection Criteria	1.	$B > r/(1 - q_{pl}/S)$	$B > r/(1 - q_{pl}/S)$	$B \leq r/(1 - q_{pl}/S)$	$B \leq r/(1 - q_{pl}/S)$
	2.	$B \leq g - rq_s/(S - q_s)$	$B > g - rq_s/(S - q_s)$	$B \leq g - rq_s/(S - q_s)$	$B > g - rq_s/(S - q_s)$
Applicable Delay Models					
Min BW Offset		-r	-r	-r	-r
Max BW Offset		$rq_s/(S - q_s)$	g - B	$\{q_s r - B(S - q_{pl})\}/(S - q_s)$	$\{q_s r - B(S - q_{pl})\}/(S - q_s)$
Delay Model #		2	2	1	1
Min BW Offset		$rq_s/(S - q_s)$	g - B	$\{q_s r - B(S - q_{pl})\}/(S - q_s)$	$\{q_s r - B(S - q_{pl})\}/(S - q_s)$
Max BW Offset		g - B	$C - B + rq_{pl}/(S - q_{pl})$	$rq_s/(S - q_s)$	$rq_s/(S - q_s)$
Delay Model #		3	4	2	2
Min BW Offset		g - B	$C - B + rq_{pl}/(S - q_{pl})$	$rq_s/(S - q_s)$	$rq_s/(S - q_s)$
Max BW Offset		$C - B + rq_{pl}/(S - q_{pl})$	g	g - B	g
Delay Model #		4	5	3	4
Min BW Offset		$C - B + rq_{pl}/(S - q_{pl})$		g - B	
Max BW Offset		g	N/A	g	N/A
Delay Model #		5		4	

Source: Rouphail (1988)

Step 1. Platoon Size and Flow Rate Estimates

From Equation 3, B was estimated as

$$B = 40*0.83*0.20/(0.50 - 0.83*0.2) = 20 \text{ sec} \quad (7)$$

and from Equations 1 and 2, respectively,

$$q_{pl} = 0.20 + (0.50 - 0.83*0.20) *\exp(-0.01215*30) = 0.43 \text{ veh/sec} \quad (8)$$

and

$$q_s = (0.2*60 - 0.43*20)/(60 - 20) = 0.085 \text{ veh/sec} \quad (9)$$

Step 2. Delay Model Identification and Calculation

From "Selection Criteria" in Table 1, the following applies:

$$30/(1 - 0.43/0.50) = 214 > 20 \text{ (Problem Type II)} \quad (10)$$

$$30 - 30*0.1/(.50 - .085) = 22.77 > 20 \text{ (Problem Type II,A)} \quad (11)$$

Thus, the following offset (Ofs) ranges and models apply (see Figure 5):

- $-30 \leq$ Ofs $\leq +2.88$ for Model 1,
- $+2.88 \leq$ Ofs $\leq +6.14$ for Model 2,
- $+6.14 \leq$ Ofs $\leq +10.0$ for Model 3, and
- $+10.0 \leq$ Ofs $\leq +30.0$ Model 4.

To compare the results of the various arrival types, offset values were set at -30, -15, 0, and $+15$ sec for HCM Arrival Types 1, 2, 5, and 4, respectively. Hence, Model 1 was used to estimate delays for Arrival Types 1, 2, and 5, while Model 4 was used for Arrival Type 4. Referring to Appendix I in the paper by Rouphail (*1*), the following results were obtained:

- For Ofs = -30 sec, d_p = 21.63 sec/veh (Arrival Type 1);
- For Ofs = -15 sec, d_p = 15.08 sec/veh (Arrival Type 2);
- For Ofs = 0 sec, d_p = 8.53 sec/veh (Arrival Type 5); and
- For Ofs = $+15$ sec, d_p = 10.39 sec/veh (Arrival Type 4).

Step 3. Estimation of Progression Adjustment Factors

Delays for a random arrival pattern were computed on the basis of Equations 4 and 5, yielding

$$d_u = 9.5 \text{ sec/veh} \quad (12)$$

and

$$d_r = 3.64 \text{ sec/veh} \quad (13)$$

The resulting PAFs, computed from Equation 6, are summarized in Table 2 along with their HCM counterparts. Surprisingly, the two methods were quite comparable at all offset values (or arrival types) except Arrival Type 1, which was

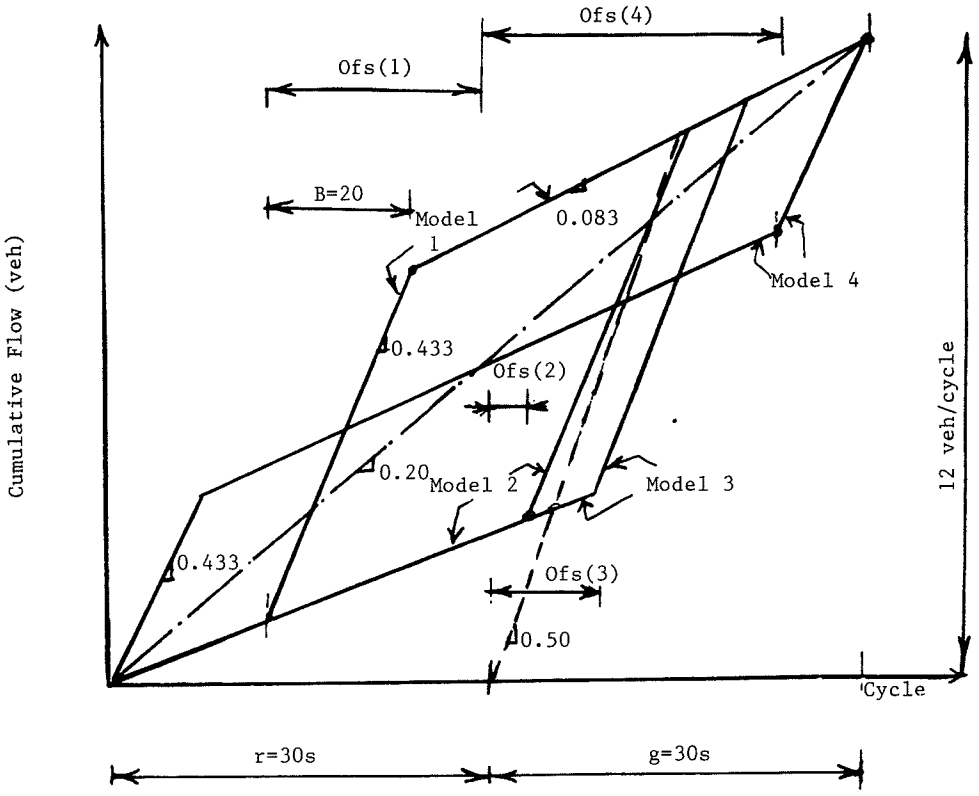


FIGURE 5 Platoon and secondary flow arrival models for numerical example.

about 10 percent higher under the proposed model. In addition, the range of PAFs (best to worst progression) was generally higher under the proposed model than with the HCM method. Interestingly, Courage et al. (*10*) observed this same phenomenon with the TRANSYT model. It appears that the platoon ratio method, because of its tendency to spread the arrivals over the entire green and red periods, is less sensitive to short, dense platoons, especially at both ends of the spectrum (Arrival Types 1 and 5). This is demonstrated in Table 3, which replicates the numerical problem but for a shorter platoon (by artificially reducing the red time at the upstream approach). In this case, g_u = 30 sec and the resulting platoon length (B) = 15 sec. Both the study and HCM methods show a smaller progression effect, since fewer vehicles arrive in the platoon compared with the original case. But, because the HCM method does not account for the actual size of the platoon, it underestimated the detrimental effect of poor progression (by 18 percent) as well as the beneficial effect of excellent progression (by 30 percent). In fact, the HCM method recognized only two arrival types in this case based on the value of the platoon ratio. Finally, this study shows that a platoon arriving in the middle of the green phase (Arrival Type 4) can result in lower delays for the lane group than if it was set to arrive at the beginning of the green (Arrival Type 5). This is because (a) the remaining green is long enough to clear the platoon and (b) the platoon does not interfere with secondary queues, which use up the initial portion of the green. The proposed model estimated the secondary queue clearance time to be about 10 sec; within that 5-sec "window" ($+10 \leq$ Ofs $\leq +15$), the platoon does not experience any stopped delay.

TABLE 2 COMPARISON OF PROGRESSION ADJUSTMENT FACTORS BY TWO METHODS WITH PLATOON SIZE OF 20 SECONDS

Platoon Offset[a] (sec)	Study Delay (sec/veh)	Study PAF	Platoon Ratio	Arrival Type[b]	HCM PAF[c]
−30.0	21.63	1.64	0.43	1	1.50
−15.0	15.08	1.14	0.71	2	1.22
0.0	8.53	0.64	1.57	5	0.67
+15.0	10.39	0.79	1.29	4	0.82

[a]Measured from the start of the green to the leading edge of the platoon or bandwidth (negative values indicate arrivals in red).
[b]Taken from Table 9.2 in the HCM.
[c]Taken from Table 9.13 in the HCM for X = 0.80.

TABLE 3 COMPARISON OF PROGRESSION ADJUSTMENT FACTORS BY TWO METHODS WITH PLATOON SIZE OF 15 SECONDS

Platoon Offset[a] (sec)	Study Delay (sec/veh)	Study PAF	Platoon Ratio	Arrival Type[b]	HCM PAF[c]
−30.0	19.51	1.48	0.62	2	1.22
−15.0	16.60	1.15	0.62	2	1.22
0.0	10.77	0.81	1.39	4	0.82
+15.0	8.30	0.63	1.39	4	0.82

[a]Measured from the start of the green to the leading edge of the platoon or bandwidth (negative values indicate arrivals in red).
[b]Arrival types are identified from Table 9.2 in the HCM based on the platoon ratio calculations. If the textual definition is used (for example, Arrival Type 1 occurs when a dense platoon arrives at the start of red), then PAF values would be identical to those in Table 2 of this paper.
[c]Taken from Table 9.13 in the HCM for X = 0.80.

SENSITIVITY ANALYSIS

A series of experiments was performed with the PAF models to test the effect of individual parameters on the final results. Of particular interest was the effect of the two parameters currently included in the HCM procedure: the lane group degree of saturation and the arrival type. Additionally, the models incorporate the effect of platoon size and dispersion via parameters B and t, respectively. The base condition consisted of the values given in the numerical example above with the following range of independent variables:

- Lane group degree of saturation (X)—varying from 0.60 to 1.0,
- Platoon travel time (t)—varying from 10 to 90 sec,
- Platoon size (B)—varying from 3 to 20 sec, and
- Proportion of lane group traffic progressed (α)—varying from 60 to 100 percent.

The results are depicted graphically in Figures 6 through 9 and discussed below.

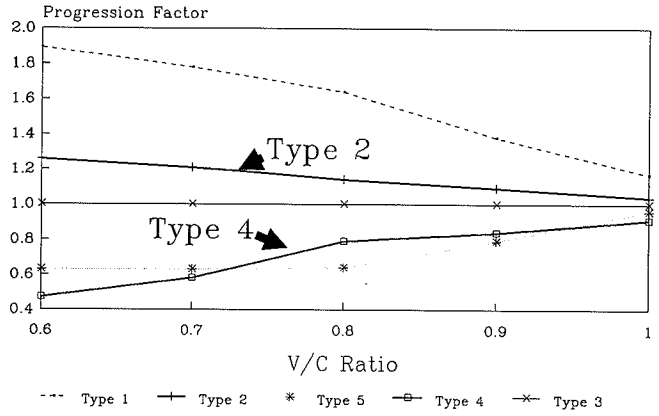


FIGURE 6 PAF versus V/C ratio by arrival type.

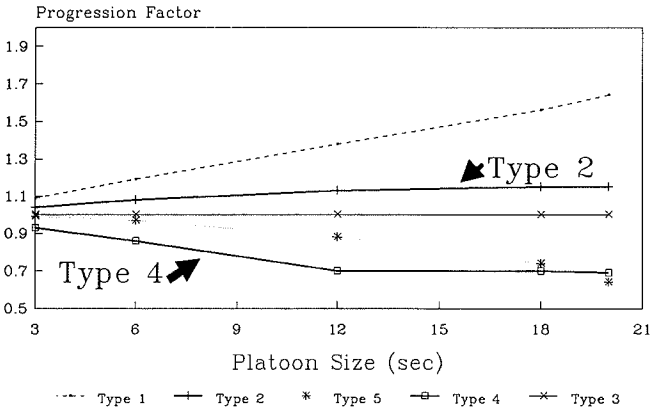


FIGURE 7 PAF versus platoon size by arrival type.

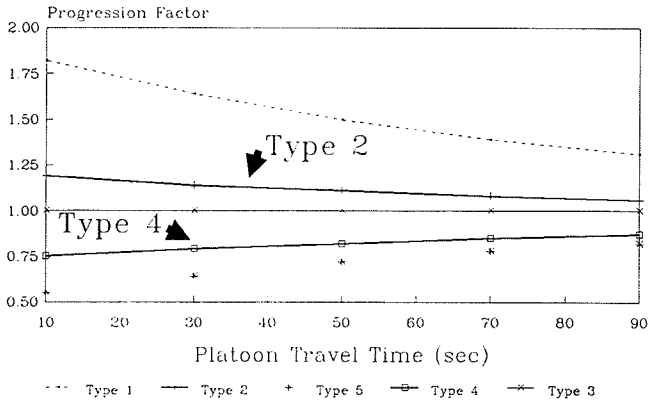


FIGURE 8 PAF versus travel time by arrival type.

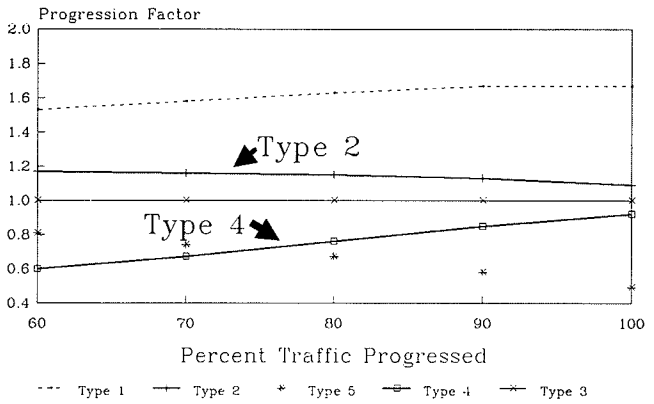


FIGURE 9 PAF versus percent progressed by arrival type.

Degree of Saturation

As the lane group V/C ratio increased, the effect of progression diminished significantly (see Figure 6). This is consistent with Table 9.13 in the HCM. The effects are more dramatic for Arrival Types 1 and 5, and less so for intermediate arrival types. These results must be viewed with caution since all progression adjustments are applied to the uniform delay component only; thus, as the V/C ratio increases, the relative effect of d_r on the PAF becomes more significant, and the overall progression effect tends to diminish.

Platoon Size

The size of the platoon was artificially manipulated by varying the red time at the upstream intersection. Since the models are only valid for undersaturated conditions, there were constraints on the length of the platoon (or, in other words, on the duration of the upstream red phase). The results are depicted in Figure 7. For Arrival Type 1, the combination of heavy platooning and poor offset selection increased the PAF substantially, while the opposite was true for Arrival Type 5. Furthermore, as long as the platoon length did not exceed one-half of the green phase, Arrival Type 4 consistently yielded lower PAF values than Arrival Type 5. This indicates an offset value that allows secondary queues to clear before the platoon arrival time and allows the platoon to clear the approach in the remaining portion of the green.

Platoon Travel Time

As suspected, the effect of progression decreased as platoon travel time increased. This is evident from Figure 8, which depicts the relationship between PAF and travel time. The effects were more pronounced at the extreme progression levels designated by Arrival Types 1 and 5.

Proportion of Progressed Traffic

This parameter represents the percentage of through traffic flow that is progressed from the upstream intersection. This factor is related to the platoon size since the amount of progressed traffic dictates the maximum size of the platoon at the upstream approach. Because of this constraint, it was assumed that an increase in α would correspond directly to an increase in platoon size. The results, which are shown in Figure 9, paralleled those in Figure 7. For Arrival Type 1, an increase in α resulted in higher delays, while the opposite was true for Arrival Type 5. As with the previous finding, Type 4 arrivals were optimal when the platoon size did not exceed half the length of the green phase. This corresponded to values of $\alpha \leq 70$ percent.

MODEL IMPLEMENTATION

The PAF estimation procedure described in this paper is implemented using an interactive microcomputer program on an IBM PC (or compatible) computer. The program, which is written in BASIC, prompts the user for the following input data:

- Lane group flow rate (veh/sec),
- Lane group saturation flow rate (veh/sec),
- Cycle length (sec),
- Effective green time for the lane group (sec),
- Distance and travel speed from the upstream intersection (ft and mph, respectively),
- Effective green time at the upstream intersection (to estimate platoon size) (sec), and
- Percentage of traffic that is progressed.

A sample input screen for the numerical example described earlier is depicted in Figure 10. The program estimates the platoon and secondary flow rates (q_{pl} and q_s, respectively) and determines the sequence of delay models that apply (according to Table 1). This information is displayed in Figure 11.

Finally, the user is prompted to enter an offset interval for displaying the results. For each offset value, the following information is given (see Figure 12):

- Average uniform delay per vehicle in the platoon (sec),
- Average uniform delay per vehicle in secondary flows (sec),

- Average uniform delay per vehicle in the lane group (sec),
- Average total delay per vehicle (uniform plus random) for the lane group (sec), and
- Progression adjustment factor.

In addition, offset values corresponding to Arrival Types 1, 2, 4, and 5 are identified for comparison with Table 9.13 of the HCM. Note that delays for Arrival Type 3 are assumed to be identical to HCM Equation 9.18.

The method described herein requires only two more input items than the HCM procedure—namely, platoon travel time and effective green at the upstream intersection—and is not dependent on field data or estimates of the platoon ratio. It is thus applicable for both design and operational analysis procedures at signalized intersections.

SUMMARY AND CONCLUSIONS

This paper has presented an application of a methodology for estimating stopped delays at signalized, coordinated intersections. The basic premise was that traffic arrives at two distinct flow rates inside and outside a platoon. This modeling concept was applied to the derivation of PAFs similar to those listed in Table 9.13 of the 1985 *Highway Capacity Manual* (*2*). A comparison of both methods revealed some shortcomings of the platoon ratio method adopted in the HCM. The following summarizes the study results:

- Averaging flow rates within and outside platoons provides better delay estimates than averaging flow rates in the red and green phases.
- The proposed models provide continuous estimates of the PAF as opposed to the discrete values currently used in the HCM method.
- Platoon adjustment factors derived from the study models directly incorporate the effects of upstream conditions on the platoon size and flow rate.
- The HCM factors tend to underestimate the effects of excellent and very poor progression (Type 1 and Type 5 arrivals), as previously determined by Courage et al. (*10*).
- Type 5 arrivals are not always optimal from a delay standpoint as stipulated in the HCM. If the green phase is long enough to clear the secondary queues accumulated in the red phase as well as the main platoon, then Type 4 arrivals may indeed produce shorter delays.

```
A  DELAY  MODEL FOR MIXED PLATOON AND RANDOM FLOWS
NAGUI M. ROUPHAIL , UNIVERSITY OF ILLINOIS , CHICAGO

ENTER AVERAGE THRU FLOW RATE ON APPROACH ,IN VEH/SEC AS QAV ? .20
ENTER SATURATION FLOW RATE FOR APPROACH IN VEH/SEC AS S? .50
ENTER CYCLE LENGTH C IN SECONDS ? 60
ENTER THE EFFECTIVE APPROACH GREEN TIME G IN SEC ? 30
ENTER PROGRESSION SPEED IN MPH ? 40
ENTER DISTANCE FROM UPSTREAM INTERSECTION ,FT ? 1760
ENTER PROPORTION OF APPROACH FLOW (QAV) THAT IS PROGRESSED, IN PERCENT ? 83
NOTE: UPSTREAM GREEN MUST BE >= 19.92  SEC TO MAINTAIN A V/C RATIO <=1
ENTER GREEN TIME AT UPSTREAM INTERSECTION ? 20
```

FIGURE 10 Sample program input screen.

```
                GENERATED DELAY MODELS AND THEIR BOUNDARIES

          ALL OFFSETS ARE MEASURED FROM START OF GREEN PHASE
       IF OFFSET < 0 , PLATOON LEADER  ARRIVES IN RED ,> 0 IN GREEN
     RANGE OF OFFSETS IS -  30  AT RED START  TO  +  30  AT GREEN END
                      PLATOON SIZE IS  20  SEC
    PLATOON[a] FLOW RATE =  .43 VEH/SEC AND SEC. FLOW RATE[b] = .08 VEH/SEC

OFFSET RANGE        FROM        TO        DELAY MODEL[c] #

OFFSET IN SEC       -30         2.88            1

OFFSET IN SEC        2.88       6.14            2

OFFSET IN SEC        6.14       10.11           3 *

OFFSET IN SEC        10.11      30              4

NOTE :FOR MODEL TYPE 3 , NO PLATOON DELAY OCCURS IN THIS RANGE

PRESS ANY KEY TO MOVE TO NEXT SCREEN

          Legend:   (a)  q_p1

                    (b)  q_s

                    (c)  See Table 1, Problem Type  II, A
```

FIGURE 11 Determination of delay model and parameters.

```
          DELAYS AT SPECIFIED  OFFSETS FOR  80 % SATURATION
   AVERAGE  DELAY FOR TYPE 3 ARRIVALS =  13.13789  SEC/VEH[a]
        PERCENTAGE OF APPROACH VOLUME OCCURING IN PLATOON=  71  %
         PLATOON RATIO FOR ZERO PLATOON OFFSET=  1.574751[b]

             AVERAGE  VEHICLE DELAY FOR SPECIFIC SUBGROUPS IN SEC/VEH

OFFSET       PLATOON    NON-PLAT.     UNIFORM      OVERALL     PROGR. FACTOR[c]

  -30          21.77        8.5         17.99         21.63          1.64
      ABOVE OFFSET REPRESENTS TYPE 1 ARRIVALS IN THE 1985 HCM

  -25          18.61        8.75        15.81         19.45          1.48

  -20          15.46        9.01        13.63         17.26          1.31

  -15          12.31        9.27        11.44         15.08          1.14
      ABOVE OFFSET REPRESENTS TYPE 2 ARRIVALS IN THE 1985 HCM

  -10          9.14         9.52        9.26          12.9           .98

   -5          6            9.78        7.07          10.71          .81

    0          2.85         10.04       4.89          8.53           .64
      ABOVE OFFSET REPRESENTS TYPE 5 ARRIVALS IN THE 1985 HCM

    5          .12          10.26       3             6.64           .5

   10          0            10.27       2.92          6.55           .49

   15          5.53         9.82        6.75          10.39          .79
      ABOVE OFFSET REPRESENTS TYPE 4 ARRIVALS IN THE 1985 HCM

   20          11.07        9.37        10.59         14.23          1.08

   25          16.48        8.93        14.34         17.97          1.36

   30          21.77        8.5         17.99         21.63          1.64
      ABOVE OFFSET REPRESENTS TYPE 1 ARRIVALS IN THE 1985 HCM

DO YOU WISH TO TEST ANOTHER SET OF OFFSETS? Y/N ?

        Legend:  (a)  Computed from Eq. 9.18 in HCM (2).
                 (b)  R_p for Arrival Type 5 in HCM.
                 (c)  Progression Factors - Compare with Table 9.13 in HCM.
```

FIGURE 12 Sample output display screen.

- While the procedures described in this paper are amenable to manual computations, they are readily applied in an interactive programming environment. Only eight input values are needed to run the program, two of which are not currently used in the HCM method but can be easily estimated.
- The Study method requires no field data collection of platoon ratios and thus is applicable to both operational analysis and design procedures for signalized intersections.

REFERENCES

1. N. Rouphail. Delay Models for Mixed Platoon and Secondary Flows. *ASCE Journal of Transportation Engineering*, Vol. 114, No. 2, 1988, pp. 131–152.
2. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
3. J. Hillier and R. Rothery. The Synchronization of Traffic Signals for Minimum Delay. *Transportation Science*, Vol. 1, No. 2, 1967, pp. 81–94.
4. D. Robertson. *TRANSYT: A Traffic Network Study Tool*. Report 253. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1969.
5. Transportation Research Center, University of Florida at Gainesville. *TRANSYT-7F User's Manual*. FHWA, U.S. Department of Transportation, 1983.
6. A. Sadegh and A. Radwan. Comparative Assessment of 1985 HCM Delay Model. *ASCE Journal of Transportation Engineering*, Vol. 114, No. 2, 1988, pp. 194–208.
7. R. Akcelik. Traffic Signals: Capacity and Timing Analysis. *Australian Road Research Report*, No. 123, Vol. 11, No. 1, 1981.
8. E. Chang and D. Fambro. *NCHRP Report 3-28C: Effects of Quality of Signal Progression on Delay*. TRB, National Research Council, Washington, D.C., 1987.
9. P. Jovanis, P. Prevedouros, and N. Rouphail. *Design and Operation of Signalized Intersections in Illinois*. Final Report IHR-012, Illinois Department of Transportation, Springfield, Illinois, 1987.
10. C. Courage, C. Wallace, and R. Alqasem. Modeling the Effect of Traffic Signal Progression on Delay. In *Transportation Research Record 1194*, TRB, National Research Council, Washington, D.C., 1988.

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

Applications of 1985 *Highway Capacity Manual* for Estimating Delays at Signalized Intersections

FENG-BOR LIN

The 1985 *Highway Capacity Manual* (HCM) contains a procedure for estimating stopped delays at signalized intersections. These estimates are to be used to assess the levels of service at an intersection. The assessed levels of service, in turn, provide a basis for making decisions concerning geometric designs and signal operations. To facilitate sound decision making, the HCM procedure must be able to produce accurate delay estimates. This paper evaluates the reliability of the HCM procedure, based on field data, and discusses needed modifications. The evaluation reveals that the procedure tends to overestimate stopped delays at reasonably well-timed signal operations. The discrepancies between the HCM estimates and the observed delays can be very large even when correct cycle lengths and green durations are used as inputs. Such large discrepancies are attributable in part to the progression adjustments recommended in the HCM procedure. Given actual cycle lengths and green durations, the procedure's ability to correctly identify the levels of service is good. However, a reliable method is needed for estimating average cycle lengths and green durations for traffic-actuated signal operations.

The 1985 *Highway Capacity Manual* (HCM) (*1*) is widely used to assist in the planning, design, and operation of highway facilities. In chapter 9 of the HCM, a procedure is provided for estimating stopped delays at signalized intersections. A major element of this procedure is a delay function, which is represented by

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

where

d = average stopped delay per vehicle for a lane group (sec/veh),
C = cycle length (sec),
g = effective green for the lane group being considered (sec),
X = volume-to-capacity ratio for the lane group, and
c = capacity for the lane group (vph).

Another major component of the HCM procedure is a set of progression adjustment factors, which are multiplied by the values obtained in Equation 1 to produce the final delay estimates. These adjustment factors attempt to account for the variations in stopped delays as a function of the type of signal control, directional movements, volume-to-capacity ratio, and vehicle arrival type.

The delays estimated from this procedure form the basis for assessing the levels of service at signalized intersections. The level of service assessments, in turn, help engineers make decisions concerning geometric designs and signal operations. Therefore, it is imperative that the HCM procedure be able to consistently produce reliable delay estimates. Little information on the reliability of this procedure has previously appeared in literature.

A major purpose of this paper is to present the findings of a study that compared observed stopped delays with estimates obtained from the HCM procedure. The paper also discusses needed enhancements to the procedure.

Twenty sets of field data were used to evaluate the HCM procedure. These data were collected in the following four urban areas in the state of New York: Canton, Potsdam, Watertown, and Syracuse. The types of signal control represented in the data sets include fixed-time control, semi-actuated control, and full-actuated control. When using the HCM procedure for actuated signal operations, the average observed cycle lengths and effective green durations were used as inputs into Equation 1. Stopped delays were estimated on the basis of 15-min flow patterns.

DATA COLLECTION AND SYNTHESIS

Stopped delays were measured for single-lane movements at seven intersections. To compare the HCM estimates with observed delays, the cycle lengths, green durations, yellow durations, and saturation flow rates were also recorded using video cameras with built-in stopwatches. For the data collection, each traffic lane was divided into two zones. Zone I covered a distance extending from the stopline to a location about 200 ft upstream. Zone II included the entire length upstream of Zone I. Two observers were used in tandem to provide separate accounts of the vehicle movements in these zones. This arrangement was necessary because of the presence of rather long queues at most of the study sites. When possible, the actual time at which a vehicle came to a stop in a queue and the time when a vehicle started moving after the green onset were recorded. For the eight or nine queuing vehicles that could usually be stored in Zone I, this task of data collection was not difficult. However, keeping track of the individual movements of those vehicles further upstream was difficult at times. Therefore, the numbers of stopped vehicles in Zone II were often recorded at 5-sec intervals. The recording interval was occasionally lengthened to 10 sec when very long queues were encountered.

Following the HCM procedure, the field data were synthesized into 15-min flow patterns for analysis. This resulted

Department of Civil and Environmental Engineering, Clarkson University, Potsdam, N.Y. 13676.

in a total of 20 sets of data. With the exception of one 15-min flow pattern involving opposed left turns, the capacities for the subject movements exceeded the arrival rates. However, spillovers of queuing vehicles from one cycle to another were not uncommon for most of the subject movements.

A summary of the data concerning the signal control, directional movement, flow rate, and capacity of each subject movement is given in Table 1. The average observed stopped delays and the stimates obtained from the HCM procedure are shown in Table 2.

OBSERVED DELAYS VERSUS HCM ESTIMATES

The HCM procedure attempts to estimate the true average stopped delay for a given condition. In contrast, the observed delay for 15-min flow is only a sample, which would invariably deviate from the true average. While discrepancies between the observed delays and the HCM estimates can be expected, the HCM procedure must be able to produce unbiased estimates.

Table 2 shows that the HCM estimates sometimes deviate significantly from the observed delays. For Movements 8 and 11, the HCM estimates were only about one-half of the observed delays. These two movements were associated with a traffic-actuated signal phase at a major intersection in Canton. This signal phase was often prematurely terminated because of the inability of its queuing vehicles to extend the green intervals continuously. In several observed cycles, the green intervals lasted only for about 15 sec when more than 10 queuing vehicles were still moving toward the intersection. Because the HCM procedure is intended for assessing reasonably efficient signal operations, it cannot adequately reflect the delays under such undesirable signal operations. Therefore, the estimates for Movements 8 and 11 do not necessarily signify a real limitation of the HCM procedure.

TABLE 1 SUMMARY OF MOVEMENT-RELATED DATA

Movement	Turn[1/] Type	Signal[2/] Control	Cycle, C sec	Effective Green, g sec	Arrival Rate, q vph	Arrival Type	Capacity vph
1	PL	semi	50.1	19.5	484	2	660
2	OL	fixed	120.0	39.4	360	2	340
3	OL	fixed	120.0	39.4	231	2	340
4	SR	fixed	90.0	28.0	528	2	565
5	SR	fixed	90.0	28.0	442	3	534
6	PL	fixed	90.0	28.6	432	2	521
7	PL	fixed	90.0	28.6	506	3	560
8	ST	full	103.1	39.7	523	3	561
9	SL	full	90.6	20.1	248	3	337
10	ST	full	86.7	31.7	448	3	516
11	ST	full	73.6	30.0	520	3	563
12	SR	full	159.0	40.0	458	1	506
13	SR	full	168.2	40.0	449	1	506
14	ST	full	125.0	19.2	193	2	243
15	ST	full	107.0	41.0	635	2	682
16	ST	full	116.0	35.0	438	1	485
17	ST	full	99.8	26.9	390	2	411
18	ST	full	116.0	33.3	465	2	478
19	ST	full	112.4	33.5	498	1	536
20	ST	full	114.7	33.3	479	2	536

[1/] OL = opposed left turns; SR = straight plus right turns; PL = protected left turns; SL = straight plus left turns

[2/] semi = semi-actuated; fixed = fixed-time; full = full-actuated

TABLE 2 COMPARISON OF OBSERVED STOPPED DELAYS AND LEVELS OF SERVICE (LOS) WITH HCM ESTIMATES

Movement	Observed		HCM Estimates		
	Delay sec/veh	LOS	Delay sec/veh	LOS	Overestimate sec/veh
1	15.0	B	12.9	B	-2.1
2	65.6	F	87.7	F	22.1
3	31.1	D	30.2	D	-0.9
4	41.1	E	47.0	E	4.9
5	26.1	D	29.1	D	3.0
6	30.9	D	29.1	D	-1.8
7	27.1	D	35.2	D	8.1
8	64.2	F	33.6	D	-30.6
9	27.7	D	25.9	D	-1.8
10	37.4	D	25.2	D	-12.2
11	44.2	E	26.4	D	-17.8
12	64.3	F	69.4	F	5.1
13	60.6	F	71.2	F	10.6
14	45.2	E	48.9	E	3.7
15	31.0	D	36.3	D	5.3
16	40.8	E	52.6	E	11.8
17	36.8	D	47.7	E	10.9
18	48.6	E	53.1	E	4.5
19	38.4	D	54.1	E	15.7
20	34.7	D	40.2	E	5.5

For most of the other movements, the HCM estimates were within 10 sec/veh of the observed delays. In fact, the HCM estimates for 11 of the 20 movements deviated from the observed values by less than 6 sec/veh. This implies that, with properly timed signals and accurate information on cycle lengths, green durations, and flow characteristics, the HCM procedure has a reasonable chance of producing good estimates for assessing the levels of serve at signalized intersections. Excluding Movements 8 and 11, the HCM procedure resulted in incorrect classifications of the level of service for only three movements.

However, as shown in Figure 1, the HCM procedure appears to have a tendency to overestimate stopped delays. The existence of such systematic errors warrants attention.

SOURCE OF ESTIMATION ERRORS

Many factors can contribute to the estimation errors produced by the HCM procedure. One potential contributing factor is

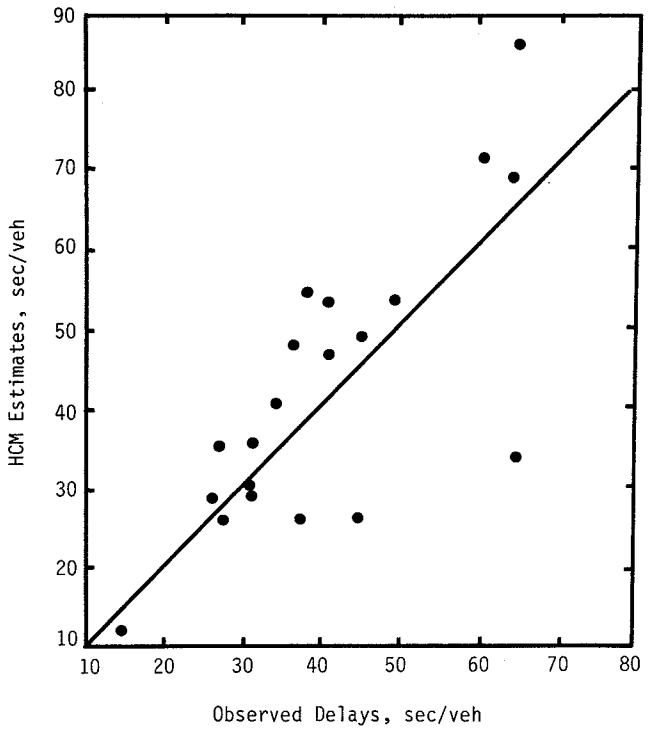


FIGURE 1 Observed stopped delays versus HCM estimates.

TABLE 3 ESTIMATED DELAYS WITH AND WITHOUT PROGRESSION ADJUSTMENT

Movement	Adjustment Factor	Estimated Delays, sec/veh			
		With	Error*	Without	Error*
4	1.19	47.0	4.9	39.5	-1.6
8	0.85	33.6	-30.6	39.5	-24.7
9	0.85	25.9	-1.8	30.5	2.8
10	0.85	25.2	-12.2	29.7	-7.7
11	0.85	26.4	-17.8	30.9	-13.3
12	1.20	69.4	5.1	57.8	6.5
13	1.20	71.2	10.6	59.2	-1.4
14	0.98	48.9	3.7	49.9	4.7
15	0.95	36.3	5.3	38.2	7.2
16	1.20	52.6	11.8	43.8	3.0
17	0.95	47.7	10.9	50.2	13.4
18	0.94	53.1	4.5	56.5	7.9
19	1.19	54.1	15.7	45.5	7.1
20	0.96	40.2	5.5	41.9	7.2

*Estimate - observed delay

the use of the progression adjustment factors for modifying the estimates obtained from Equation 1. Table 3 compares delays estimated with and without applying the progression adjustments. The table reveals that, without the adjustments, most of the large estimation errors can be reduced substantially. By treating the estimates obtained from Equation 1 as the final estimates, the estimation errors for the movements shown in Table 3 can be reduced from an average of 10 to 7.7 sec/veh. Therefore, the progression adjustments given in the HCM require revision and, until revised factors are available, it may be advisable not to apply the adjustment factors to vehicular movements at isolated intersections.

Another factor that may contribute to the estimation errors is the use of a constant coefficient of 0.38 in the first term of Equation 1. This constant implies that, for uniform arrivals with no spillovers of queuing vehicles from one cycle to another, the stopped delays equal 76 percent of the corresponding approach delays.

Referring to Figure 2, the approach delay of a vehicle is defined in this paper as the elapsed time between the expected time of arrival (t_1) at the stopline and the time of departure (t_4) from the stopline. The corresponding stopped delay is measured from time t_2 (at which the vehicle comes to a stop) to time t_3 (when the vehicle begins to move after the green onset). The total approach delay of the vehicles arriving at a uniform rate in a cycle is represented by the area of the triangle ABC in Figure 3. The slope of AC of the triangle equals an arrival rate of q vph. The slope of BC represents a saturation flow rate of S vph in green duration. The base of the triangle, which has a length of r sec, is equal to the effective red interval (cycle length minus effective green).

With no spillovers of queuing vehicles from one cycle to another, it can be shown that the average approach delay for

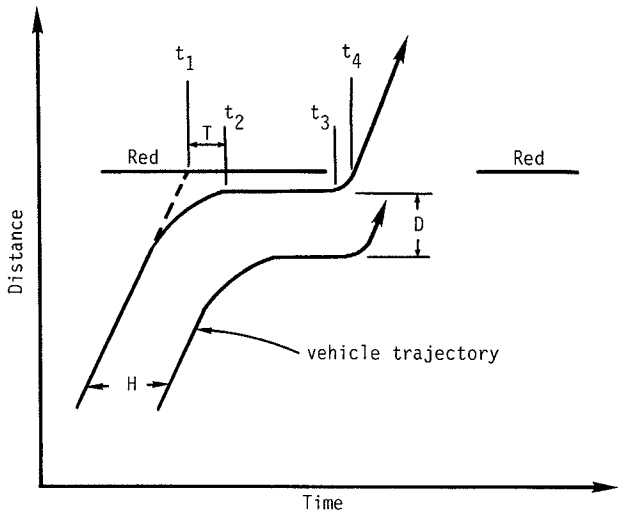


FIGURE 2 Schematic of approach delay and stopped delay.

a uniform arrival pattern can be estimated from

$$d_a = \frac{C(1 - g/C)^2}{[1 - (g/C)X]/2} \quad (2)$$

Conceptually, it is reasonable to assume that, for a given cycle length, the stopped delays will account for a larger proportion of the approach delays as the available green intervals

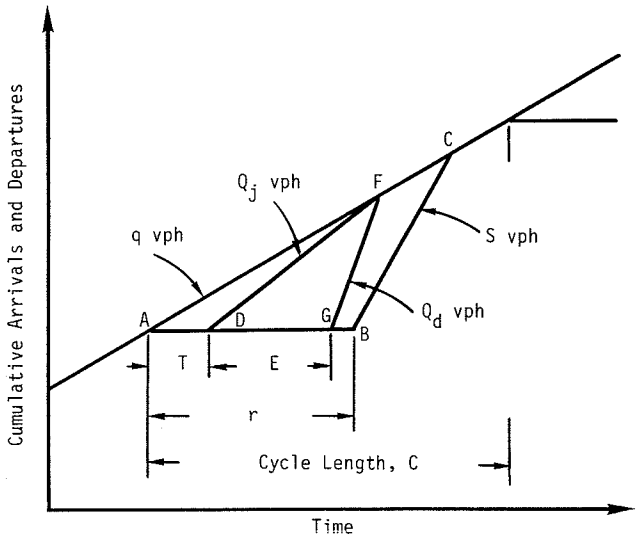


FIGURE 3 Determination of delays per cycle for uniform arrivals.

are shortened. In other words, the constant coefficient of 0.38 used in Equation 1 should be replaced by a variable.

To examine the magnitude of estimation errors attributable to the use of a constant coefficient, assume that vehicles approach the intersection at speed V and, if necessary, decelerate a constant rate B to come to a stop. Based on this assumption, the first queuing vehicle, which is expected to reach the stopline at t_1 (Figure 2) if no deceleration is necessary, will come to a stop at $t_1 + V/(2B)$. The time lag between the expected arrival time and the stop time is denoted as T in Figures 2 and 3. The value of this time lag is

$$T = V/(2B) \tag{3}$$

The second queuing vehicle has an expected arrival time of $t_1 + H$, where H is the uniform headway, in seconds, of the arriving vehicles. The value of H is

$$H = 3{,}600/q \tag{4}$$

This vehicle will come to a stop behind the first queuing vehicle at $t_1 + H - D/V + V/(2B)$, where D is the spacing between two stationary queuing vehicles, measured from the front end of one car to the front end of the other.

Similarly, the third queuing vehicle will join the queue at $t_1 + 2H - 2D/V + V/(2B)$. In other words, the rate at which vehicles come to a stop is one vehicle every $H - D/V$ seconds. If this rate is represented by Q_j vph, then

$$Q_j = 3{,}600/(H - D/V) \tag{5}$$

After the green onset, the queuing vehicles will start moving successively at an average rate of one vehicle per R sec, where R is the driver reaction time. This implies that the queuing vehicles depart from the stationary queue at a rate of

$$Q_d = 3{,}600/R \tag{6}$$

where Q_d is the rate of departure, in vph, from the stationary queue.

Therefore, the total stopped delay for the vehicles arriving in one cycle can be represented by the smaller triangle DGF in Figure 3. The base of this triangle can be approximated by a time period of

$$E = C - G - Y - R - T \tag{7}$$

where Y is the yellow interval and G is the green interval. (All units in Equation 7 are in seconds.)

The average stopped delay can then be determined as

$$d_s = \frac{(E^2/C)(Q_j/q)}{(1 - Q_j/Q_d)/2} \tag{8}$$

Based on Equations 2 and 8, it can be shown that the ratio of stopped delay to approach delay can vary from less than 0.4 (when the green interval accounts for a large portion of the cycle) to more than 0.7 (when the green interval is relatively short). This relationship is intuitively correct and can be confirmed by computer simulation.

Figures 4 and 5 compare the simulated delays with the estimates obtained from Equation 2 and Equation 8, respectively. These estimates are based on a driver reaction time of $R = 1.2$ sec (*2*), a yellow interval of $Y = 4$ sec, a vehicle approach speed of $V = 50$ ft/sec, a deceleration rate of $B = 6.25$ ft/sec^2, and a saturation flow rate of $S = 1{,}630$ vphg. The simulation model used in this study generates uniform arrivals and processes the vehicles downstream by updating their positions and speeds once per second of real time. Unlike the analytical model, the simulation model allows the arriving vehicles to move toward and out of the intersection according to probabilistic flow characteristics observed in the field (*2*). The simulation runs performed in this study analyze the delays under various conditions for 1 hr. As shown in Figures 4 and

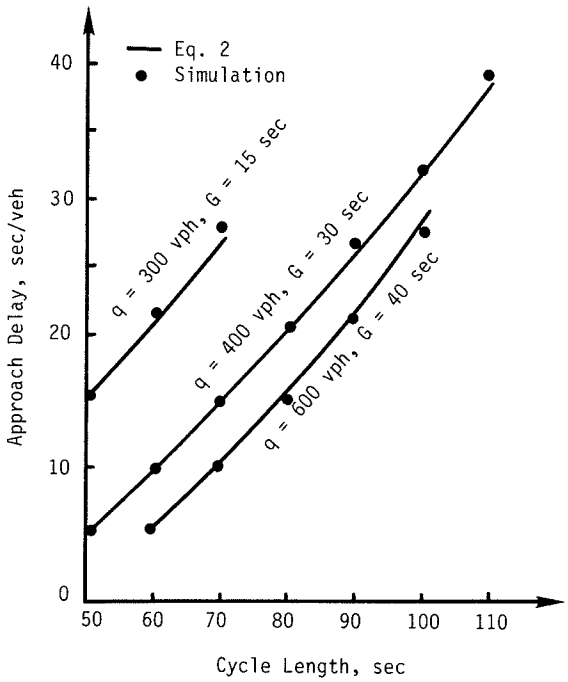


FIGURE 4 Simulated approach delays versus delays determined from Equation 2 for uniform arrivals.

5, the agreement between the simulated delays and the theoretical values is remarkably good.

Figure 6 compares the stopped delays estimated from Equation 8 and the first term of Equation 1 for uniform arrivals. The first term of the HCM delay function tends to overestimate delays when they are under 20 sec/veh. Beyond this level, it tends to underestimate the delays. However, because the estimation errors are mostly less than 3 sec/veh, the first term of the HCM delay function can be considered sufficiently accurate and the modification of Equation 1 should be focused on the second term. To alleviate the tendency to overestimate delays, the second term can be revised to lower its sensitivity to the value of the volume-to-capacity ratio (X).

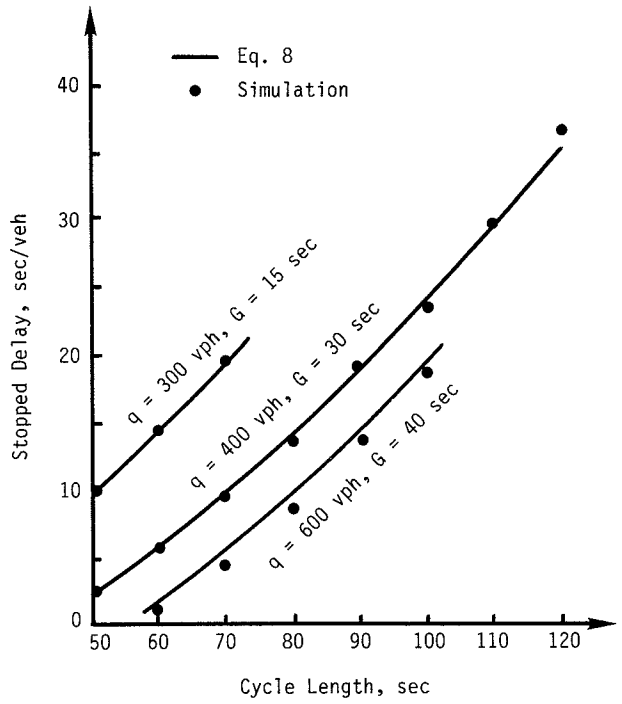


FIGURE 5 Simulated stopped delays versus delays determined from Equation 8 for uniform arrivals.

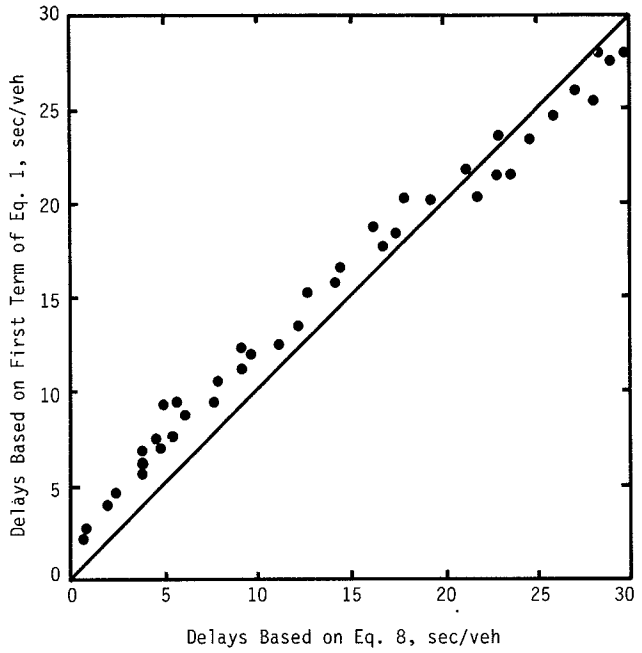


FIGURE 6 Comparison of delays estimated from Equation 8 and the first term of Equation 1 for uniform arrivals.

A WEAK LINK

Although the HCM estimates compare favorably with the observed delays and the levels of service for most of the subject movements, these comparisons use the actual cycle lengths and green durations as inputs into the HCM procedure. In reality, the applications of the HCM procedure are likely to involve unknown cycle lengths and green durations at traffic-actuated signals. This is a major flaw in the HCM procedure.

The HCM suggests that the average cycle length for a traffic-actuated signal operation be estimated from

$$C = LX_c/(X_c - \Sigma(v/S)_{ci}) \quad (9)$$

where

C = cycle length (sec),
L = loss time per cycle (sec),
X_c = critical volume-to-capacity ratio for the intersection,
v = volume of a lane group (vph),
S = saturation flow of a lane group (vphg), and
$(v/S)_{ci}$ = critical flow ratio for lane group i.

In using this equation, a value must be assumed for X_c. The suggested values are 0.85 for semi-actuated control and 0.95 for full-actuated control. These values may be realistic in closely packed platoons. However, at intersections where a signal phase is associated with a number of traffic lanes, the green intervals can be extended for long periods even though the capacity of each individual lane or lane group is far from being fully utilized. This is because traffic-actuated logics combine individual lane flows into a single, equivalent flow to the gaps between successive actuations of detectors. As a result, traffic-actuated operations may have a much smaller X_c than the values suggested for Equation 9. Under such a condition, the use of the suggested values would lead to large errors in estimating average cycle lengths.

The signal operation observed at the T-intersection between Central Avenue and Vly Street in Albany underlines this potential problem. This T-intersection was controlled with a two-phase, semi-actuated signal. The observed total flow rate on Central Avenue was 2,492 vph, which included a critical flow of 1,260 vph and a corresponding saturation flow of 3,260 vphg. The side street had a critical flow of 143 vph and a saturation flow of 1,500 vphg. If Equation 9 is used with a loss time of 10 sec and an X_c of 0.85 as suggested, the estimated average cycle length would be 23 sec. In contrast, the observed average cycle length was 103.4 sec and the corresponding X_c value was about 0.53.

It can be determined from Table 1 that the volume-to-capacity ratios for the subject movements under full-actuated control ranged from 0.73 to 0.95. Because most of these movements represented the critical movements of their respective

signal phases, the X_c ratios for the related intersections would have been only as high. Therefore, for the purpose of broadening the applicability of Equation 9, X_c should be treated as a function of cycle length, green interval, flow rate, and saturation flow. This equation can be made even more meaningful for the operational analysis of signal operations if it is enhanced to account for the impact of timing settings and detector configurations.

CONCLUSIONS

Given accurate information on cycle lengths, green durations, and saturation flows, the HCM procedure produces delay estimates that are mostly within 10 sec/veh of the observed values for movements regulated by well-timed signals. However, the discrepancies between the HCM estimates and the observed delays for several movements examined in this study are quite large. Furthermore, it appears that the HCM procedure has a tendency to overestimate delays. Nevertheless, the HCM procedure can correctly identify the levels of service if accurate cycle lengths and green durations are used in Equation 1.

The large discrepancies between the HCM estimates and some of the observed delays can be reduced significantly if no progression adjustments are applied to the estimates obtained from Equation 1. Obviously, the progression adjustments given in the HCM need revision. The first term of Equation 1 was found to be sufficiently accurate for estimating the levels of service. Therefore, the tendency to overestimate delays may be mitigated by reducing the sensitivity of the second term of Equation 1 to the value of X.

Applications of the HCM procedure for assessing traffic-actuated signal operations often involve unknown cycle lengths and green durations. The current method for estimating these factors is a major weakness of the HCM procedure. To provide realistic estimates, the value of X_c in Equation 9 must be treated as a variable rather than as a constant. In addition, to facilitate comparisons of alternative signal operations, Equation 9 should be enhanced to account for the variations in cycle lengths with respect to timing settings and detector configurations.

ACKNOWLEDGMENT

The author wishes to thank Sangaranathan Vijaykumar of the Civil and Environmental Engineering Department at Clarkson University for his assistance in collecting the data presented in this paper.

REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
2. F. B. Lin and D. Cooke. Modeling of Queue Dissipation for Signal Control. *ASCE Journal of Transportation Engineering*, Vol. 112, No. 6, Nov. 1986, pp. 593–608.

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

Accuracy of Delay Surveys at Signalized Intersections

S. TEPLY

Delay is an excellent tool for evaluating the operation of signalized intersections. However, it is not easily determined. This paper examines two approaches for measuring delay—a time-space diagram and a queuing diagram—and explains various problems related to each. The paper concludes that, while delay cannot be precisely measured, it can be a useful engineering tool if it is calculated properly.

A number of evaluation criteria are available to assess how well a lane at a signalized intersection operates under a given set of geometric and timing conditions. These criteria include the volume-to-capacity ratio, the reserve capacity, the probability of discharge, the load factor, and the average delay. Delay has a prominent place among these measures of effectiveness because it relates directly to the experience of the drivers and because its meaning is generally understood. The 1985 *Highway Capacity Manual* (*1*) uses delay as the sole basis for determining the level of service. The 1984 *Canadian Capacity Guide for Signalized Intersections* (*2*) identifies delay as the most powerful means of evaluating intersection operation, although it also recommends the use of other simultaneous measures. Nevertheless, although delay is an excellent evaluation tool, it is not easily determined.

Hurdle (*3*) discusses the principles behind analytical delay formulas. In a similar sense, this paper explains problems related to field measurement of delay in undersaturated conditions and shows that what is calculated is not always what is measured. Since direct measurements are not possible, surveys employ indirect techniques based on the time-space concept or on queuing theory. A detailed examination of both approaches shows that one of the major problems lies in the inherent vagueness of delay definition and in the different interpretations of the concept of delay. The paper explains relevant problems and concludes that, despite these difficulties, delay remains a good traffic engineering tool, provided it is calculated, measured, and used in a consistent fashion.

DELAY DETERMINATION IN A RECONSTRUCTED TIME-SPACE DIAGRAM

This method of delay measurement has not been commonly used, although it has been in existence for over 20 years. However, new electronic and computing developments have recently made it more appealing (*4*).

Figure 1 shows trajectories of two vehicles in a typical time-space diagram as used in signal design. In the figure, distance is horizontal and time is vertical. The driver of the first vehicle was forced to stop by the red signal, and the second vehicle moved unimpeded through the stopline during the green signal. In other words, the first vehicle experienced some delay while the second vehicle obviously had no delay. The wavy nature of the trajectories indicates speed noise (the fluctuation of speeds).

Trajectory 1 represents a vehicle that approached the intersection during a green interval. When it was some distance upstream, the signal changed to amber (Points a,A) and then to red. The driver slowed down and came to a complete stop (Points b,B). Vehicle 1 was stopped until shortly after the beginning of the green interval when the driver started accelerating (C) and continued the trip. The time during which the vehicle did not move ($B-C$) is called "stopped" delay (shown as d_s). This portion of the delay is rarely timed directly in intersection surveys; however, it can be as an indicator in floating car surveys. No travel time considerations are involved in stopped delay.

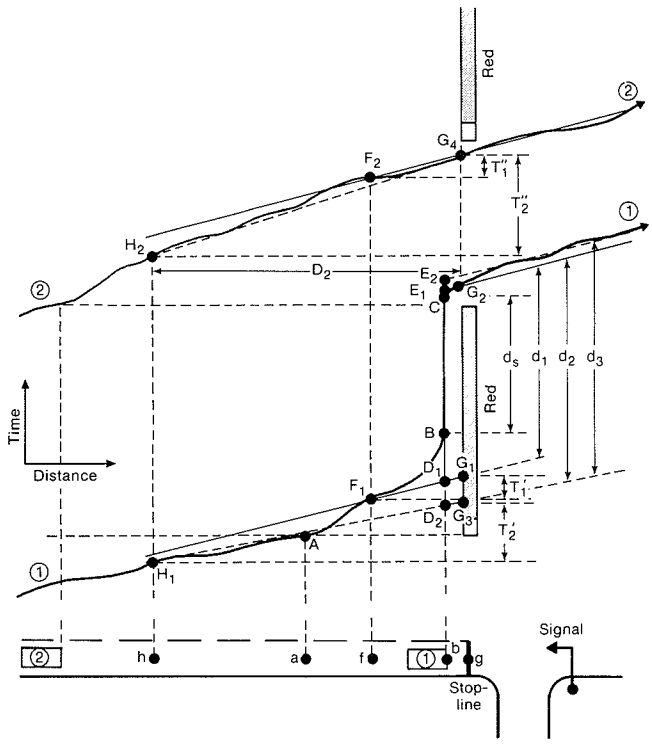


FIGURE 1 Schematic time-space diagram depicting delay factors at a signalized intersection approach.

Department of Civil Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2G7.

A common definition of delay is the time "lost" by the impact of the signal. (This paper does not consider delays caused by traffic or its overflow.) This delay involves some portion of the slowing down and speeding up process and can be represented by an approximation of the vehicle trajectory using straight lines. In the following discussion, this delay is called "overall delay" to distinguish it from stopped delay and from "total delay," which represents the sum of the delays experienced by individual vehicles.

An arbitrary point (f) with a corresponding point in time (F_1) can be employed to identify the approach part of the trajectory. The time when the vehicle crosses the stopline (G_2) is used for the discharge portion of the trajectory. The overall delay experienced by Vehicle 1 is represented by d_1 in Figure 1. This delay can be determined as the time difference between points F_1 and G_2 minus the free-flow travel time between point f and the stopline (g). This travel time, shown as T_1', is the vertical difference between F_1 and G_1. Nevertheless, it cannot be measured directly because Vehicle 1 did not cross the stopline until after it experienced delay. Other vehicles, however, can be employed instead. For example, Vehicle 2 was not impeded by other traffic or by the signal; consequently, its travel time (T_1'') between point f and the stopline (time F_2-G_4) can be used as a surrogate for T_1'. T_1' is usually determined as an average travel time for vehicles similar to Vehicle 2.

It would appear that the overall delay has now been accurately determined. However, a closer examination of real-life situations (as shown in Figure 2) reveals a problem with the definition of reference point f (or F_1 and F_2 in time). Speeds of vehicles approaching a traffic signal are rarely constant, especially when drivers can observe the change from green to amber and red. Referring again to Figure 1, after passing point a, the driver of Vehicle 1 slowed down considerably when the signal changed. As a result, the previous F_1-G_1 straight-line approximation of Trajectory 1 does not apply. If another arbitrary point (h) further upstream were used, a linear approximation of the approach process would be different. Instead of G_1, point G_3 would probably be used. The average speed of Vehicle 1 between h and b, however, would be faster than that between f and b. Consequently, a larger portion of the deceleration would be included, and the delay (d_2), calculated on the basis of a more distant upstream point, would be longer than the previously determined delay (d_1).

An additional problem lies in the determination of the free-flow travel time between the reference point and the stopline for vehicles similar to Trajectory 2. Again, the average speed between h and g (the slope of the straight line between H_2 and G_4) is different than the average speed between f and g, (the slope between F_2 and G_4). After a longer duration of a green interval, drivers can hardly be expected to behave in the same way as drivers approaching a red signal. As a result, the measured travel time (T_2'') is somewhat different than the T_2' needed to determine the delay. Since points f and h are arbitrary and their choice influences the delay, the delay values cannot be considered absolute.

The issue becomes even more complex when considering the speed behavior of vehicles departing from a standing queue. For the straight-line approximation of vehicle trajectories, it is normally assumed that vehicles will travel from time point G_2 at the free-flow speed determined for the approach. Although this assumption may be true for vehicles that stopped

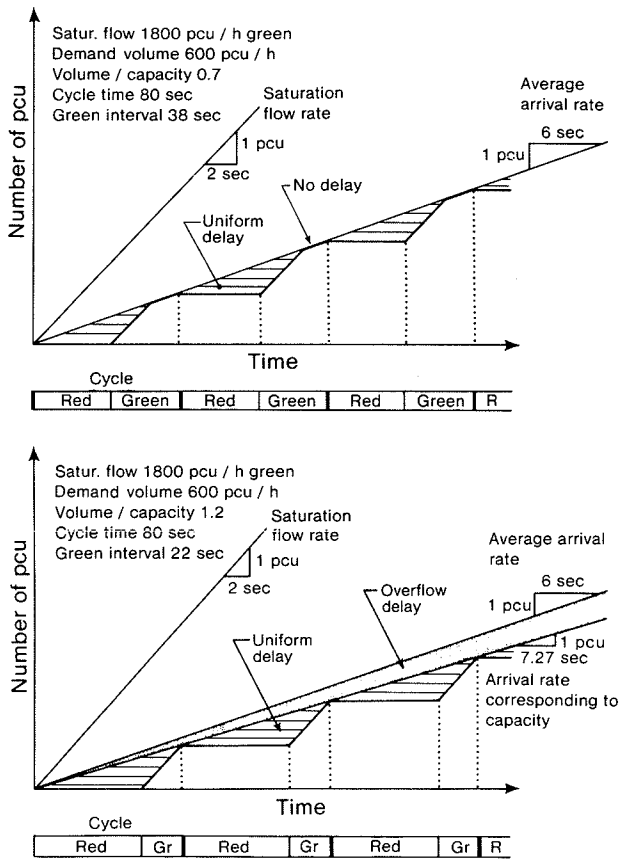


FIGURE 2 Signalized interaction delay interpretation in a deterministic queuing diagram (2).

farther upstream of the stopline, it apparently does not hold for the first few vehicles stopped during red. The discharge trajectories of these vehicles should not be approximated by lines that are parallel to the approach trajectories and pass through the point in time when the vehicle crossed the stopline (G_2). Rather, shifted lines should be used to represent the resumed movement downstream of the intersection. As a consequence, point E_1 should be replaced by E_2. The resulting overall delay (d_3) represents the longest but possibly the most correct value because it truly reflects the time lost because of signal operation.

The principles of an overall delay survey system that employs vehicle trajectories are discussed by Teply and Evans in a paper in this Record.

DELAY DETERMINATION USING A QUEUING DIAGRAM

Description

Delay measurements are usually based on the principles of deterministic queuing theory, as shown in Figure 2, rather than on assessing the arrival and discharge times of individual vehicles.

In contrast to time-space diagrams, the queuing diagram does not have a distance dimension. Vehicles "appear" at points on the arrival line and "disappear" at horizontally cor-

responding points on the discharge line (the saturation flow line). The horizontal distance between these points represents the delay for individual vehicles. Since these delays fill in the space between the arrival and discharge lines, the sum of all delays for all vehicles (the total delay) is represented by the area between both lines. The type of delay depends mostly on the definition of a vehicle's arrival and discharge times, as discussed in the previous section.

The focus of the following discussion is on uniform delay although most aspects also apply to overflow delay.

The vertical difference between both lines in the queuing diagram represents the number of vehicles accumulated at any given time (in other words, a queue). Because of the missing distance dimension, however, the diagram actually portrays a "stacking" of vehicles at the stopline, with arrivals at the top and departures at the bottom. This fact is important for delay measurements.

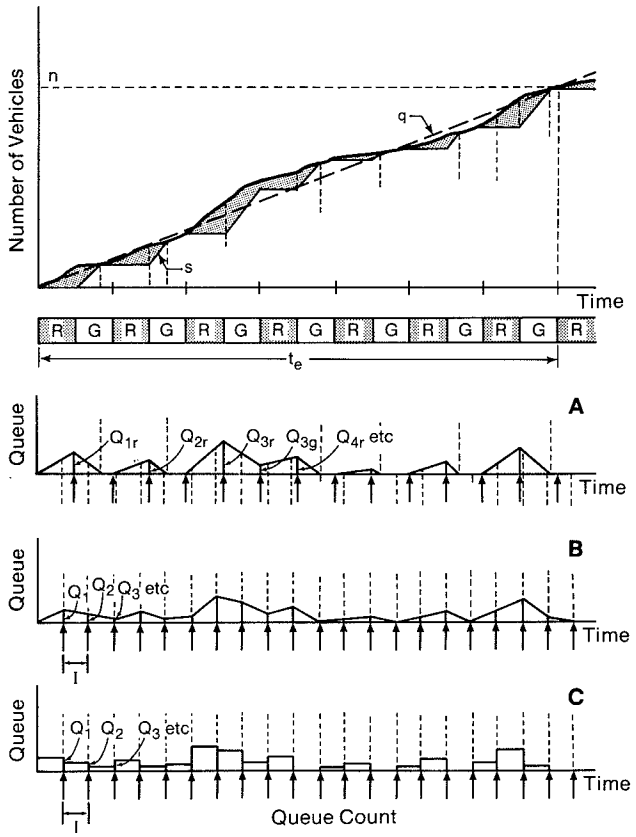
There are several ways in which a queuing diagram can be interpreted and used to determine delay. The top portion of Figure 3 depicts a fixed time signal operation with a relative steady arrival rate (q), subject to some random fluctuations. The rate of discharge during green intervals is represented by a constant saturation flow (s).

Queues (the vertical dimensions in the top portion of Figure 3) can be transferred to the diagram in Figure 3A. A linear representation of the development and decay of queues can employ only the most significant points in time, such as the ends of the green and red intervals and periods without queues. As a result, the total area of Figure 3A is approximately equal to the area between the arrival and discharge lines in the top portion of the figure. The average delay during the survey period can then be determined as the sum of the areas in Figure 3A divided by the total number of vehicles, i.e.,

$$D = \frac{\Sigma D_i}{n} \tag{1}$$

Since delay is usually surveyed over a limited number of cycles, it is not completely impractical to determine the individual areas in Figure 3A. Figure 4A shows the basis for this calculation. Apparently, areas in the top portion of Figure 4 equal areas in Figure 4A, i.e.,

$$\Delta ABC = \Delta A_1B_1C_1 \tag{2}$$



NOTE: Upward-pointing arrows indicate the points in time when a queue count is taken.

FIGURE 3 Transformations of a typical signalized interaction queuing diagram for various types of delay surveys: A, counting queues at the beginnings and ends of red intervals; B, counting queues at regular intervals "out-of-step" with the signal cycle, assuming linear growth or decay of queues during individual intervals; C, counting queues at regular intervals "out-of-step" with the signal cycle, assuming constant queues during individual intervals (I).

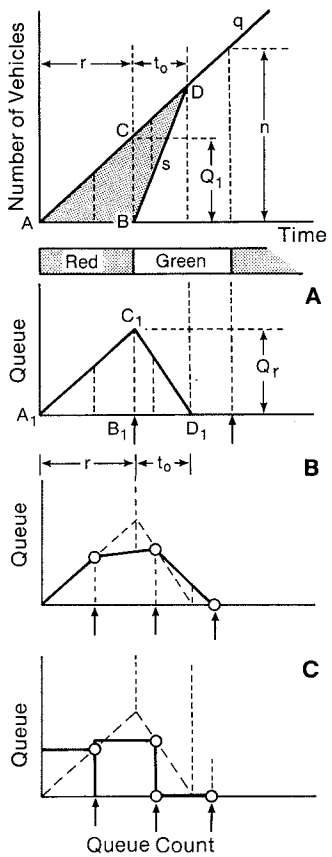


FIGURE 4 Delay calculations based on queue surveys shown in Figures 3A, 3B, and 3C.

and

$$\Delta BCD = \Delta B_1C_1D_1 \tag{3}$$

To determine area ΔBCD, the value of t_o is needed. This factor represents vehicles that joined the queue during the green interval. It can be calculated from the top portion of Figure 4 as follows:

$$q(r + t_o) = st_o \tag{4}$$

That is,

$$t_o = \frac{qr}{s - q} \tag{5}$$

where

r = red interval (sec);
s = saturation flow (veh/hr), measured directly; and
q = arrival rate (veh/hr).

The delay in Figure 3A is then the sum of all areas in the diagram divided by the total number of vehicles (n), i.e.,

$$d = \frac{1}{n}\left[\frac{1}{2} r\, Q_{1r} + \frac{1}{2} r\, Q_{1r}\right] + \left[\frac{1}{2} r\, Q_{2r} + \frac{1}{2} r\, Q_{2r} \frac{q}{s - q}\right] + \left[\frac{1}{2} r\, Q_{3r} + \frac{1}{2} g\,(Q_{3r} + Q_{3g})\right] + \ldots \tag{6}$$

where Q_{ir} equals queues at the end of red intervals and Q_{ig} equals queues at the end of green intervals, both expressed in number of vehicles.

The variable q can be averaged by

$$q = \frac{n}{t_e} \tag{7}$$

where t_e is the evaluation time.

To simplify these equations, it is usual to account not for a full value of t_o but only for discharge of vehicles in the queue at the end of the red interval, i.e., $t_o' = Q/s$. Another alternative is to assume that the discharge takes place during the total green interval, i.e., $t_o' = Q/g$. These changes will cause delay to be somewhat underestimated or overestimated.

Most delay surveys employ delay counts at regular intervals, out of step with the cycle time, to provide for random sampling. Two of these techniques are illustrated in Figures 3B, 3C, 4B, and 4C. Based on Figures 3B and 4B, average delay is then

$$d = \frac{1}{2n} I \Sigma\,(Q_i + Q_{i+1}) \tag{8}$$

where

I = duration of the survey interval,
n = total number of vehicles, and
Q_i = queue at the beginning of survey interval i (0, 1, 2 . . .).

The most common technique, described in the *Highway Capacity Manual* and illustrated in Figures 3C and 4C, assumes that a vehicle observed in the queue has been stopped for an average of the interval between counts. The average delay is then

$$d = \frac{I \Sigma Q_i}{n} \tag{9}$$

All three methods shown in Figures 3 and 4 can be considered reasonable approximations of the delay since the over- and underestimations appear to cancel each other. Naturally, they cannot be expected to produce identical results. In fact, in situations such as high arrival fluctuations, platooning, or overflows, the results will vary significantly.

Problems

Earlier sections of this paper have illustrated the vagueness of delay definitions and the differences in measurement techniques. There are, however, additional problems with using the basic deterministic queuing diagram for delay calculation and surveys.

The following questions are asked most frequently:

- Does the stacking of vehicles at the stopline introduce errors in delay determination?
- Since delay surveys are based on the counting of vehicles stopped in a queue, should overall delay or stopped delay be measured?
- What is the relationship between overall delay and stopped delay?

The Effect of Stacking Vehicles

When determining the arrival rate (q) in the queuing diagram, the average value is generally based on the number of vehicles that pass through an intersection approach lane during the survey period, as shown in Figure 3 and Equation 7. The time-space diagram in Figure 5 confirms that this rate is correct at a point some distance upstream of the queue. In the straight-line approximation of vehicle trajectories, however, vehicles are joining the end of the queue at a different rate. Since the end-of-queue shock wave travels backward, the rate of end-of-queue arrivals must be greater than the upstream rate of flow.

The arrival flow at the end of the queue can be determined easily from the spacing of arriving vehicles, their speed, and the spacing of vehicles in the queue. In the specific example in Figure 5,

- Arrival flow (q) = 900 veh/hr (one vehicle every 4 sec);
- Saturation flow = 1,800 veh/hr (one vehicle in 2 sec);
- Arrival speed = 50 km/hr = 13.9 m/sec;
- Arrival spacing = 4 × 13.9 = 55.6 m;
- Spacing of vehicles in the queue = 6.0 m; and
- Distance to be travelled by consecutive vehicles to reach the end of the queue = 55.6 − 6.0 = 49.6 m.

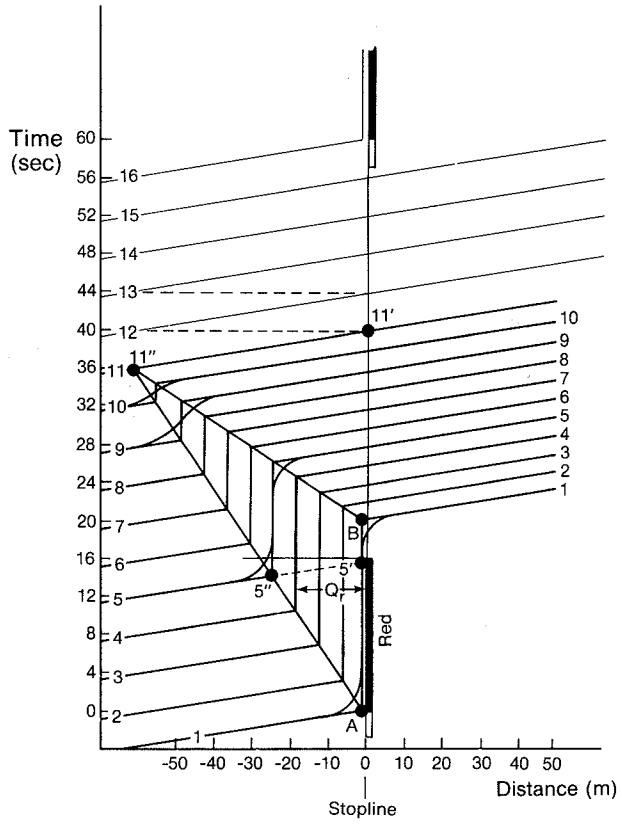


FIGURE 5 Time-space diagram of traffic conditions analyzed in Figures 6 and 7.

The rate of vehicle arrivals at the end of the queue can be calculated as follows:

$$\frac{1}{q_e} = \frac{49.6 \text{ m/veh}}{13.9 \text{ m/sec}} = 3.57 \text{ sec/veh} \quad (10)$$

Therefore, the rate of flow at the end of the queue is

$$q_e = \frac{3{,}600 \text{ sec/hr}}{3.57 \text{ sec/veh}} = 1{,}008 \text{ veh/hr} \quad (11)$$

An identical result can be obtained by calculating the speed of the shock wave and the rate at which vehicles cross it.

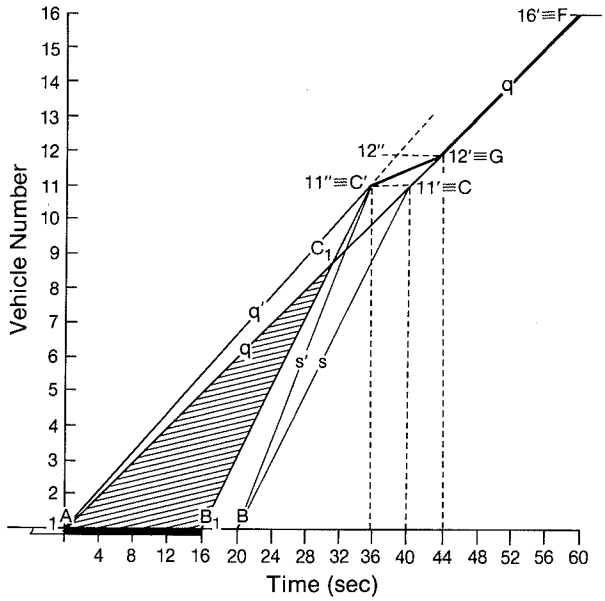
This rate of flow can be transferred into the queuing diagram in Figure 6. As a result of this rate at the end of queue, Vehicle 11 appears in the diagram not at Second 40 (10 headways of 4 sec each) but at Second 35.7 (10 headways of 3.57 sec each).

Provided there is no overflow (no spillover of the queue into the next cycle), vehicles can be considered discharged once they have moved from the queue and resumed their speed. Consequently, the appropriate saturation flow is the discharge from the queue (at the front shock wave), not the one at the stopline. It can be calculated in the same way as the flow arriving at the end of queue. Vehicles in the queue start moving not when the previous vehicle is 2 sec × 13.9 m/sec = 27.8 m away but 6.0 m less, i.e., 21.8 m. Therefore, the headway is 21.8 ÷ 13.9 = 1.57 sec, and the resulting saturation flow at the front of the queue is 3,600 ÷ 1.57 = 2,293 veh/hr (as opposed to the 1,800 veh/hr at the stopline). This is clearly shown in Figure 6 since Vehicle 11 did not encounter any delay and, as a result, the saturation flow line at the front of the queue must connect with the arrival rate at the end of the queue at Point 11″. The flow rate at the front of the queue results from the following calculation: for 10 headways (i.e., 11 vehicles), 35.7 − 20.0 sec were needed, or 1.57 sec per headway. Therefore, the flow was 2,293 veh/hr.

When there is no queue, arriving and discharging vehicles must be counted at the stopline. Naturally, the discharge rate equals the arrival rate, and no vehicle is delayed. The return of the count to the stopline from the upstream end of the queue causes a longer headway between the last vehicle in the queue and the first unimpeded vehicle. In Figures 5 and 6, Vehicle 11 arrived at (and was discharged from) the queue at Second 35.7 while Vehicle 12 arrived at (and was discharged from) the stopline at Second 44. In other words, the headway was more than 8 sec. After that, the count continued at the stopline and the original arrival rate of one vehicle in 4 sec was resumed. The consequences for the queuing diagram are shown in Figure 6.

While the events with the vertical stacking of vehicles at the stopline were represented by the shape *ABCF*, the diagram corrected for the effect of horizontal queuing is represented by *ABC'GF*. The distortion is caused by the changing point of reference. Nevertheless, the areas of the two triangles that identify the total delay, ΔABC and $\Delta ABC'$, are identical. As a result, the simpler form *ABCF* yields correct delay values.

Both analytical formulas and survey calculations based on queue counts, however, usually fully or partially neglect the fact that vehicles accelerating from the stop upstream of the stopline encounter delay until they reach the previous full speed (see Figures 1 and 5). For most delay interpretations, it is assumed that the saturation flow discharge starts at the



NOTE: Points A and B correspond to straight-line first vehicle trajectory representation in Figure 5. For uniform delay formulas, Point B_1 is usually used.

FIGURE 6 The effect of "stacking" of vehicles at the stopline.

end of the red interval or an "effective red," which is usually only slightly longer than the actual red interval. Consequently, the triangle included in delay calculations in Figure 6 is reduced to ΔAB_1C_1, which is smaller than the correct ΔABC. The resulting average delay is even shorter than that shown in Figure 1 as the difference between d_3 and d_1 (or d_2, depending on the reference point).

Nevertheless, in typical situations, analytical delay formulas and direct measurement of vehicle arrivals and discharges yield adequate approximations of uniform delay values.

Overall Delay versus Stopped Delay

As shown in Figure 5, Vehicle 5 would not be considered a stopped vehicle since it was not yet part of the stopped queue at the end of the red interval. It was still in motion some distance from the end of the queue, although the straight-line approximation of its trajectory indicates that it belongs to the queue. As a result, in the queuing diagram in Figure 7, only four vehicles constitute the queue at the end of the red interval. This observation can be generalized: when watching a stopped queue grow, the rate of arrivals given by the line AC does not apply. The rate-of-arrival line should connect the time points at which vehicles stopped—line A_1C_1 in Figure 7. Although the rate-of-arrival line is not used in this technique, the survey calculations attempt to approximate the individual areas below it, as shown in Figures 3 and 4.

Similar reasoning applies to the front shock wave. Vehicles actually depart from the queue sooner than indicated by straight-line approximation of their trajectories in the time-space diagram, which roughly corresponds to line B_1C_1 in Figure 7. As mentioned above, however, the correct line (BC) is usually not considered.

The consequences of these considerations depend on the way a delay survey is carried out and interpreted.

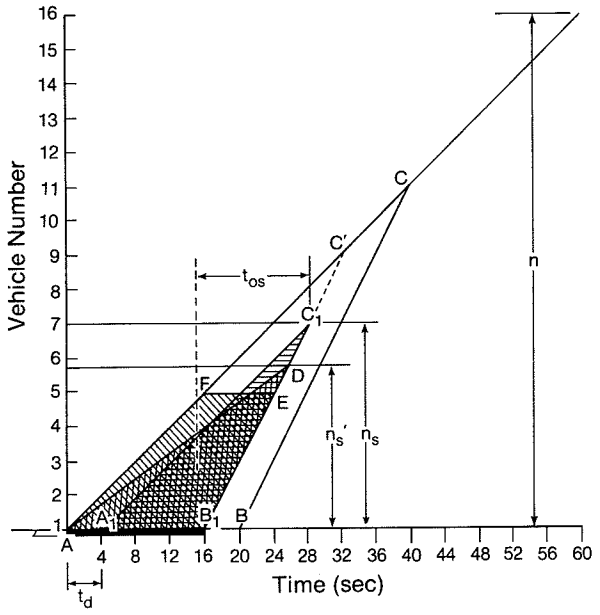


FIGURE 7 The effect of counting stopped queues on the value of delay.

Counting Queues at Regular Intervals Out of Step With Signal Cycle This method is discussed in the *Highway Capacity Manual* and by Buehler et al. (*4*). Only those vehicles standing in the queue are included. Therefore, the correct total delay in Figure 7 given by the triangle ΔABC is reduced to triangle $\Delta A_1B_1C_1$, which includes only stopped delay.

The area of $\Delta A_1B_1C_1$ in Figure 7 can be approximately calculated as

$$D_s = \frac{1}{2}(r - t_d)n_s \tag{12}$$

where

D_s = total delay (veh-sec),
t_d = deceleration delay (sec), and
n_s = number of vehicles stopped (veh).

Since $q(r - t_d + t_{os}) = st_{os}$, where t_{os} = time needed to dissipate the standing queue (including vehicles stopped during the dissipation) (sec),

$$t_{os} = \frac{q(r - t_d)}{s - q} \tag{13}$$

As mentioned, the acceleration delay is usually not included, even in the analytical formulas.

Therefore,

$$n_s = st_o = \frac{sq(r - t_d)}{s - q} \tag{14}$$

and

$$D_s = \frac{sq(r - t_d)^2}{2(s - q)} \tag{15}$$

The relationship between overall delay and stopped delay is then

$$\frac{D}{D_s} = \left[\frac{sqr^2}{2(s - q)}\right]\left[\frac{sq(r - t_d)^2}{2(s - q)}\right]^{-1} = \frac{r^2}{(r - t_d)^2} \tag{16}$$

When the ratio between overall delay and stopped delay is surveyed and calculated in this fashion, it depends not on arrival or discharge rates but solely on the duration of the red interval and on the deceleration delay t_d.

The deceleration delay can be considered relatively constant in most urban situations. For instance, at a 50 km/hr approach speed and a deceleration rate of 3.0 m/sec^2, the deceleration delay would be 4.6 sec. (A similar value can also be assumed for the neglected acceleration delay.)

For very long red intervals, the ratio would apparently approach 1.0, which would mean that the surveyed values (of stopped delay) should be close to those calculated (of overall delay). For short red intervals, however, the ratio may be a very high value. For example, if the effective red interval (which usually includes amber) is 18 sec and t_d = 5.0 sec, the ratio is 2.25. This means the calculated delay would be 2.25 times longer than the surveyed one.

These examples also indicate that it is not appropriate to use a fixed coefficient in an analytical formula to convert the delay into stopped delay.

Counting Queues at End of Red and Green Intervals This technique is not as common as the previous one, but it is used when a cycle-by-cycle reconstruction of the queuing diagram is needed. As explained above, when counting the number of vehicles stopped in a queue, the correct area of the triangle ΔABC in Figure 6 is reduced. The survey calculations (as shown in Figures 3 and 4) assume, however, that the area can be approximately reconstructed from the queue at the end of the red interval (and, if there were an overflow delay, from the queue at the end of the green interval).

Figure 7 illustrates the type of distortion that takes place. Since only four vehicles have been included in the queue, the calculated rate of arrival is not five vehicles. Therefore, instead of four headways in 16 sec, only three headways would be used, which corresponds to one vehicle in 5.3 sec. This average headway represents 675 veh/hr, not the correct value of 900 veh/hr.

The area of total delay that would be used for calculations is then given by the triangle ΔAB_1D. The resulting average delay value would again reflect stopped delay rather than overall delay, with a similar impact on their ratio as in the previously mentioned technique.

Even if this problem could be eliminated by including vehicles that are about to join the standing queue, survey crews usually have difficulty accounting for vehicles that join a moving queue after the beginning of the green interval. Such vehicles are represented, for instance, by Trajectories 6 through 10 in Figure 5. They do not have to come to a full stop as their simplified straight-line trajectories suggest. In Figure 7, excluding Vehicles 6 through 10 would "cut off" the top of the delay triangle and cause the total delay to be represented by the area AB_1EF. This shape would also correspond to the queue dissipation calculation mentioned with the survey techniques that count queues at the ends of green and red intervals.

When calculated with an area of AB_1EF, the total delay is

$$D_c = \frac{qr^2(1 + y)}{2} \tag{17}$$

where

$$y = \frac{q}{s} \tag{18}$$

The ratio of overall delay to the delay calculated in this fashion is then

$$\frac{D}{D_c} = \left[\frac{qr^2}{2(1 - y)}\right]\left[\frac{qr^2(1 + y)}{2}\right]^{-1} = \frac{1}{1 - y^2} \tag{19}$$

Contrary to previous stopped delay considerations, the result is completely independent of the signal timing.

Since $y = q/s$, the distortion depends solely on the relative magnitudes of the arrival and saturation flows. When the arrival rate is very small, the ratio would be close to 1.0; for conditions close to capacity, the ratio would be very large.

The cases represented in Equations 16 and 19 are rather extreme. In most situations, the measurements are subject to both errors. As a result, the ratio of the calculated overall delay to the measured stopped delay would depend on the duration of the red interval as well as on arrival and saturation flows.

PRACTICAL COMPARISONS

To test the findings of this paper, delays were surveyed at intersections along St. Albert Trail in the city of St. Albert. Two different techniques were used:

1. A vehicle trajectory survey to represent overall delay values (described in a paper by Teply and Evans in this Record); and
2. A queue length count at 15-sec intervals, assuming constant queues during individual intervals (*1*), which yields rough stopped delay values.

The southbound direction was assessed during the morning peak period and the northbound direction during the afternoon peak. These directions experience heavy commuter flows at the intersection of St. Albert Trail and Hebert Road.

The results of the comparison are shown in Figure 8. With the exception of the intersection at Hebert Road, all average delays were relatively short. As expected, the delay based on queue counts was usually shorter than that based on vehicle trajectories. Several irregularities, however, are obvious and are discussed below.

Giroux Road (a.m. and p.m.)

For this intersection, the average delay surveyed by vehicle trajectories was shorter than that based on overall queue counts. The intersection has a very short red interval in the surveyed direction along St. Albert Trail. Local drivers slow down well upstream of the intersection, which was not accounted for in the trajectory survey distance base. Consequently, the overall delay was underestimated. On the other hand, the stops are brief—shorter than the counting interval. As a result, the assumption that vehicles were stopped for the whole duration was incorrect, and stopped delays were overestimated. In addition, the flow in the surveyed direction was light.

St. Vital and St. Anne (a.m.)

These intersections produced similar results but for a different reason. Signal progression at these intersections was difficult, and platoons of vehicles arrived just at the end of the red interval and were forced to stop for several seconds. As a result, vehicles were not stopped for the whole interval between counts. (The coordination at these intersections has since been re-designed).

Hebert Road (a.m. and p.m.)

Because of the magnitude of its delay, this intersection yielded the best illustration of the measurement problem. The ratio of measured average overall and stopped delays was 1.35 in

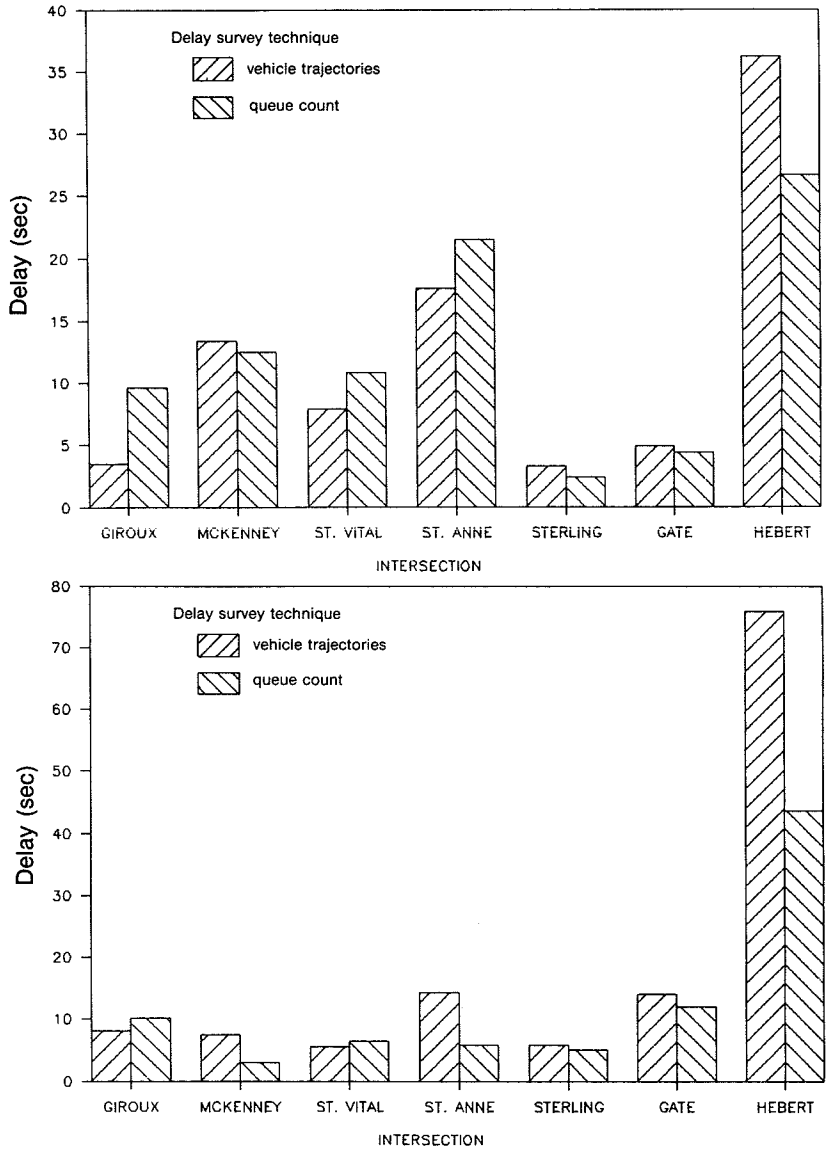


FIGURE 8 Comparison of delays measured by two different survey methods. The two diagrams are based on different scales.

the a.m. peak and 1.73 in the p.m. period. The volumes were very heavy (over 2,000 veh/hr in the peak directions in three lanes). The ratio of overall delay and stopped delay determined using Equation 16 was 1.2 for both periods. The ratio calculated using Equation 19, which reflected the effect of vehicles that joined the queue during the green interval, was between 2.1 and 2.4 for both a.m. and p.m. periods. Assuming that delay surveys based on the reconstructed vehicle trajectories yield a nearly correct average delay value, the ratios of measured values (1.35 and 1.73) fall within the range of expected magnitudes (1.2 and 2.1 or 2.4). Naturally, a perfect match can hardly be expected for the reasons discussed in this paper.

CONCLUSIONS

The discussion in this paper shows that delay cannot be precisely measured. As a result, a perfect match between an analytical delay formula and measured delay values cannot be expected.

Specific findings, some of which are illustrated in Figure 9, can be summarized as follows:

- Uniform delay formulas slightly underestimate overall delay because they neglect a portion of the acceleration delay.
- For a similar reason, typical delay surveys underestimate delays.
- Delay surveys based on the reconstruction of individual vehicle trajectories yield a reasonable approximation of average overall delay at signalized intersections. The values are not exact because of the arbitrary nature of the reference distance.
- The fact that uniform delay formulas and delay surveys based on queue counts do not account for the rate of arrival at the end of the queue and the rate of discharge at the front of the queue does not distort the resulting delay values.

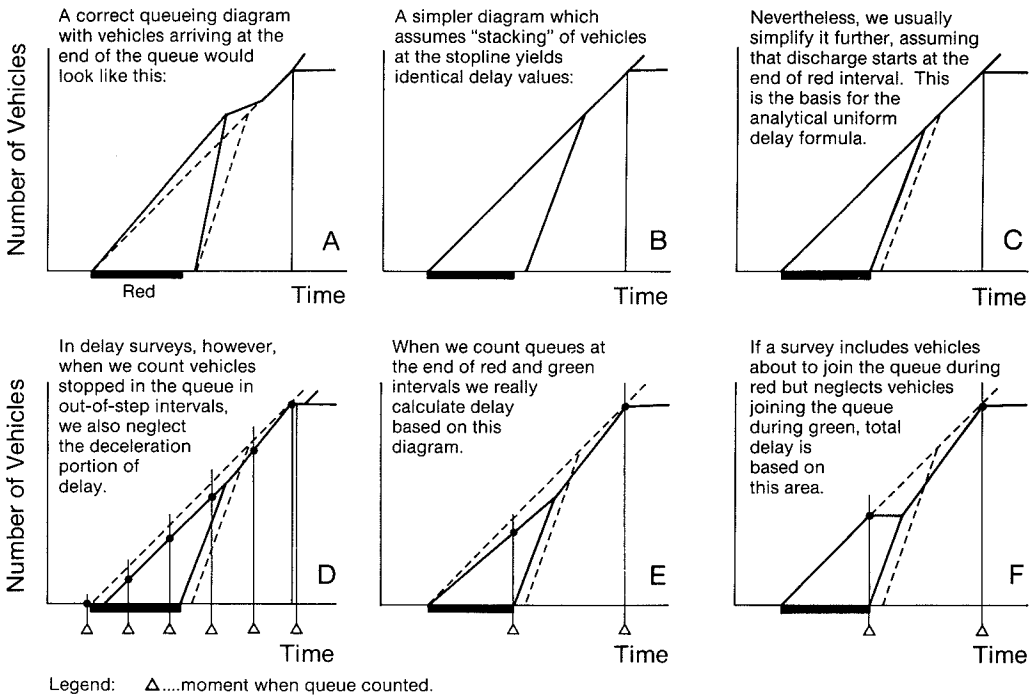


FIGURE 9 Schematic representation of the findings related to the applications of queuing diagrams.

- Different delay survey methods based on queue counts cannot be expected to yield identical results because of different assumptions in the application of queuing theory.
- Delay surveys based on stopped queue counts produce stopped delay values. In situations with low volumes and short red intervals, these techniques may overestimate stopped delay to the point of exceeding overall delay values. (Overall delay includes deceleration and acceleration delays.)
- The ratio between delay and stopped delay is not constant and depends mostly on the duration of the red interval. As a result, a fixed coefficient for all timing conditions is not appropriate.
- The ratio between measured values of overall delay and stopped delay is also not constant. In most situations, it depends not only on the duration of the red interval but also on the relationship between arrival and saturation flows. As a consequence, an accurate estimation of this ratio is difficult.
- In some situations, such as where the red interval is very short and drivers are familiar with the signal operation or in coordinated systems where vehicles encounter delays shorter than the queue count interval, the value of stopped delay calculated from the queues may be longer (incorrectly) than the overall delays determined from vehicle trajectories.

ACKNOWLEDGMENTS

The support of this research by the Natural Sciences and Engineering Research Council of Canada through equipment and operational grants is gratefully acknowledged. Several graduate students at the University of Alberta contributed to the project: Hung-Chuan Kua was responsible for the surveys, which were organized with the assistance of the city of St. Albert; the hardware system used for the vehicle trajectories surveys was adapted by Paul Sabourin; and the software was developed by Gary Evans.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. P. Richardson, J. Schnablegger, B. Stephenson, and S. Teply. *Canadian Capacity Guide for Signalized Intersections* (S. Teply, ed.). Institute of Transportation Engineers, District 7, Canada, 1984.
3. V. F. Hurdle. Signalized Intersection Delay Models—A Primer for the Uninitiated. In *Transportation Research Record 971*, TRB, National Research Council, Washington, D.C., 1984.
4. M. G. Buehler, T. J. Hicks, and D. S. Berry. Measuring Delay by Sampling Queue Backup. In *Transportation Research Record 615*, TRB, National Research Council, Washington, D.C., 1976.

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

Comparison of Macroscopic Models for Signalized Intersection Analysis

LAWRENCE T. HAGEN AND KENNETH G. COURAGE

The signalized intersection methodology presented in the 1985 *Highway Capacity Manual* (HCM) introduced a new delay model, which naturally invites comparison with the delay models contained in existing traffic signal timing design and analysis techniques. This paper compares the HCM delay computations with those performed by the Signal Operations Analysis Package (SOAP) and by TRANSYT-7F Release 5. The paper focuses on the effect of the degree of saturation, the peak-hour factor, and the period length on delay computations and on the treatment of left turns opposed by oncoming traffic. All of the models agreed closely at volume levels below the saturation point. When conditions became oversaturated, the models diverged; however, they could be made to agree by the proper choice of parameters. The computed saturation flow rates for left turns opposed by oncoming traffic also agreed closely. However, the treatment of protected plus permitted left turns produced substantial differences. It was concluded that neither SOAP nor the HCM treats this case adequately. Therefore, an alternative model based on a deterministic queuing process was proposed and evaluated.

The 1985 *Highway Capacity Manual* (HCM) (*1*) introduced a new analytical technique for modeling the operation of traffic at signalized intersections. This represents a significant change from the previous HCM methodology. The change was required to support the newly adopted use of vehicular delay as a measure of the level of service experienced by drivers at traffic signals. Previous techniques were based on the relationship between traffic volume and intersection capacity.

Traffic signal operations have received extensive analytical treatment in the past. It should not be surprising that several computerized models already exist; it is also natural that the introduction of a new model would invite comparisons with the established methodology. This paper attempts to make such a comparison using two traffic signal timing design and analysis programs that have been employed extensively in the United States. The comparisons are limited to isolated traffic signal operations. In other words, no consideration is given to the effect of coordination with adjacent signalized intersections, which has been discussed in a previous paper (*2*).

The two existing programs considered are the Signal Operations Analysis Package (SOAP) (*3*) and the Traffic Network Study Tool (TRANSYT) (*4*). SOAP is concerned only with single, isolated intersections, while TRANSYT treats such intersections as a special case in a traffic signal network, which is its primary area of application. The version of TRANSYT used is TRANSYT-7F, Release 5.

Transportation Research Center, University of Florida, 512 Weil Hall, Gainesville, Fla. 32611.

DELAY MODEL COMPARISON

Delay is an important measure in traffic engineering since it represents a direct cost of fuel consumption and an indirect cost of the time lost to motorists. Delay results when traffic is impeded by factors beyond the motorist's control. It may be due to interference from other motorists or attributable to the traffic control devices themselves.

Webster's Delay Model

Before comparing the SOAP and TRANSYT models with the HCM model, it is important to understand what they represent and where they originated. Both models are variations of one first proposed by Webster in 1958 (*5*). Webster's delay model for intersections controlled by pretimed signals is given in the following equation:

$$D = \frac{C(1-\lambda)^2}{2(1-\lambda X)} + \frac{X^2}{2q(1-X)} - 0.65\left(\frac{C}{q^2}\right)^{\frac{1}{3}}[X^{(2+5\lambda)}] \quad (1)$$

where

D = average delay per vehicle on the subject approach or movement,
λ = proportion of the cycle that is effectively green for the approach or movement (g/C),
X = degree of saturation [volume/capacity (V/C)],
C = cycle length,
q = flow rate (the average number of vehicles passing a given point on the road in the same direction per unit of time), and
g = effective green time.

The first term of Webster's equation represents the average delay per vehicle, assuming the vehicles arrive at a uniform rate (q) throughout the cycle. The second term attempts to account for the fact that the vehicles arrive randomly, not uniformly. This term increases rapidly with the degree of saturation. The third term is an adjustment factor developed semiempirically to provide a better mathematical fit to the theoretical curve. The upper limit for Webster's model was generally considered to be slightly less than full saturation.

Enhancements to Webster's Model

All techniques considered in this paper use a two-component version of Webster's original three-component model. The first term accounts for the delay experienced by traffic with uniform arrivals from cycle to cycle, and the second term provides for the additional delay caused by randomness in the arrival patterns.

The first terms differ as follows:

- SOAP uses Webster's first term directly.
- The HCM uses Webster's first term divided by 1.3 to reflect only the stopped delay portion of the total delay.
- TRANSYT divides the cycle into 60 equal steps and accumulates arrivals over each step. This produces results identical to Webster's first term for isolated operation.

So, in effect, all of the models use the same formulation for the first term.

The second terms have the following differences:

- SOAP and TRANSYT both use Robertson's model (*4*), which was originally incorporated into TRANSYT to account for oversaturation. The length of the oversaturated period is an important factor in the result produced by this model. Neither SOAP nor TRANSYT uses a peak-hour factor (PHF) since peaking characteristics are treated implicitly in the model.
- The HCM uses a different formulation to account for oversaturation. In this case, the period of oversaturation is assumed to be 15 min. The PHF has a strong influence on the results.

So, each model uses one parameter that is not recognized by the other. By choosing the proper combination of values for these parameters, it should be possible to produce more or less equivalent results.

Example Problem

A simple example is presented to determine how the HCM and SOAP differ when they are analyzing the same conditions. The example deals with the intersection of two two-lane, two-way streets with only through movements on all approaches.

The delay models were implemented over a wide range of volumes for the same capacity. The results for undersaturated conditions are summarized in Figure 1, which shows that the two models are very similar for PHF = 1.0. For PHF = 0.9, the delay values from SOAP and the HCM are significantly different because the PHF increases the flow rate only in the HCM.

Figure 2 reveals a noticeable difference in delay values for oversaturated conditions because SOAP uses a 60-min period and the HCM is based on a 15-min period. Therefore, SOAP

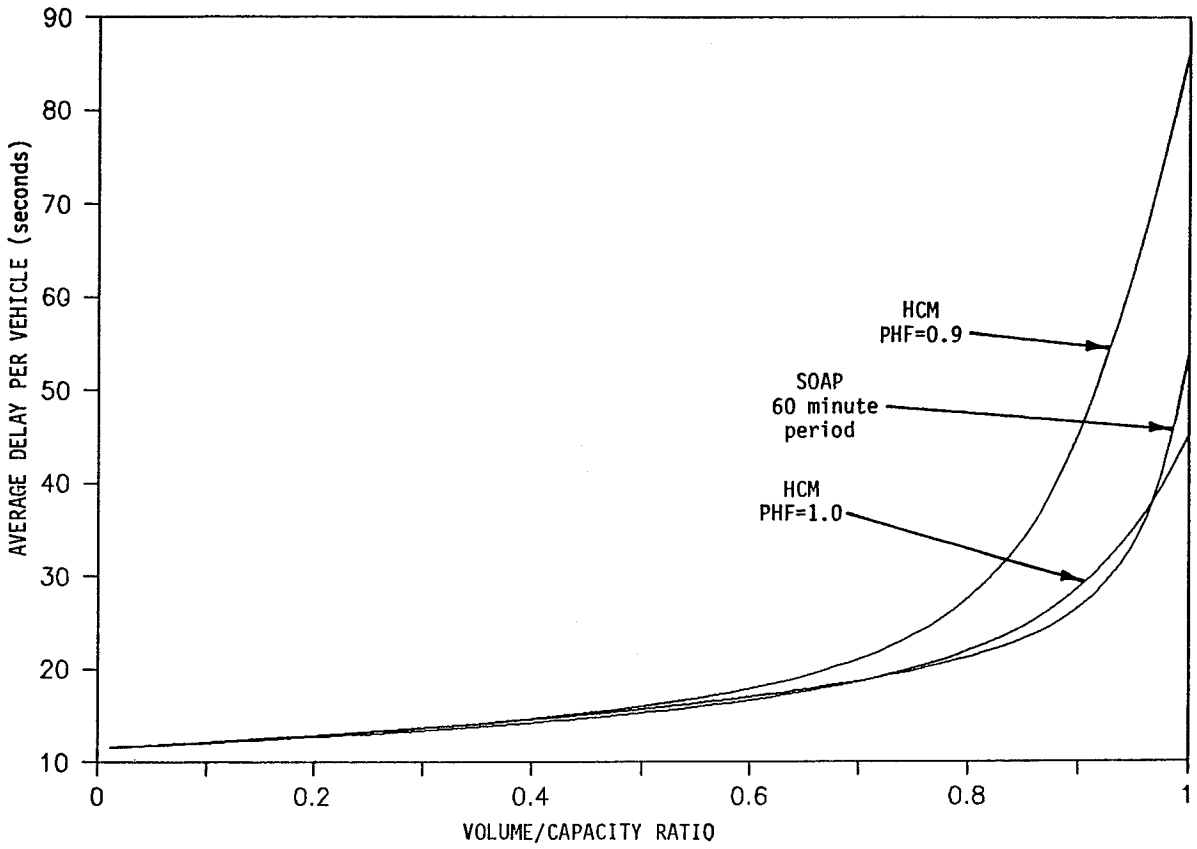


FIGURE 1 Comparison of SOAP and HCM delay values for undersaturated conditions.

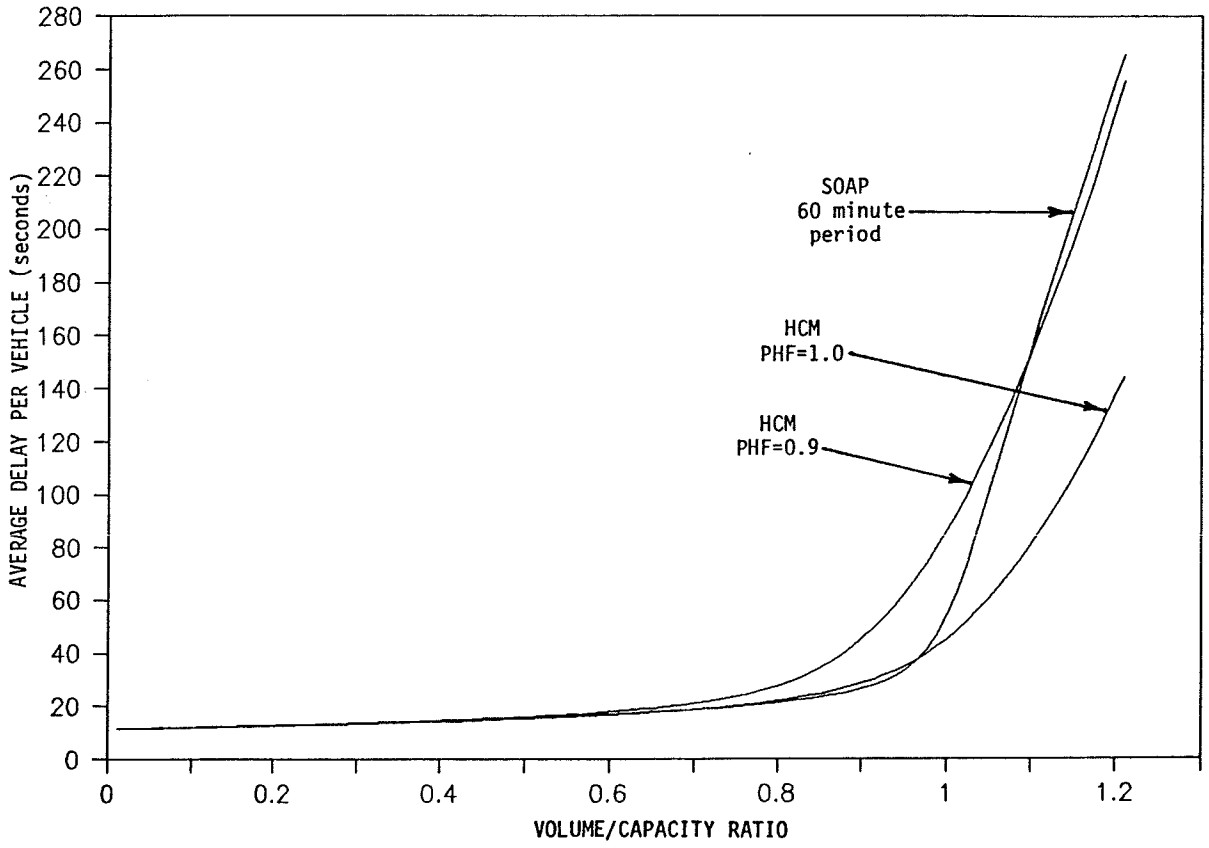


FIGURE 2 Comparison of SOAP and HCM delay values for oversaturated conditions.

assumes that the oversaturation lasts for a much longer period of time, which results in a much higher value for the delay per vehicle due to a constantly increasing residual queue.

Figure 3 compares of SOAP's delay for a 60-minute period with the HCM's delay for various PHF values. For V/C values less than 1.0, SOAP agrees well with the HCM for PHF = 1. For V/C > 1.0, the values from SOAP agree well with the HCM for PHF = .90. The HCM shifts the curves to the left by increasing the 15-minute flow rate for the same hourly volume. This is accomplished through the use of the PHF.

In Figure 4, the HCM's delay values for PHF = 1 are compared with SOAP's delay for various period lengths. The figure shows that SOAP's delay values for a period length of 40 min agree well with the HCM for large V/C values. It is also apparent that the period length is a significant factor in SOAP's delay equation.

Figures 3 and 4 demonstrate that the HCM's delay values for PHF = 1.0 compare well with SOAP's delay values for 60-min period lengths provided the intersection is undersaturated. For oversaturated conditions, however, a small difference in PHF or period length leads to significantly different results.

PERMITTED LEFT TURNS

One of the most difficult traffic movements to analyze at a signalized intersection is the permitted left turn. Through and unobstructed right-turning movements can essentially proceed at their saturation flow rate throughout the green phase. On the other hand, permitted left turns must filter through gaps in the oncoming traffic at a much lower saturation flow rate. The saturation flow rate for opposed left turns is primarily a function of the opposing traffic volume.

To analyze this complex situation, both SOAP and the HCM compute the unsaturated green time for the opposed left turns based on the opposing volumes. In the HCM, a supplemental worksheet for permitted left turns is used. This worksheet contains a series of complex equations to calculate the effects of the opposing traffic on the left-turning traffic. It also computes the impact of the left-turning traffic on through vehicles in shared lanes.

Example Problem

An example is used to demonstrate the comparison of the permitted left-turn models. It is similar to the example for the previous section except that these approaches include exclusive left-turn lanes.

The intersection configuration is shown in Figure 5. To ensure that any differences that arise are due to differences in the models themselves, it is assumed that there are no trucks, buses, pedestrians, or grades at the subject intersection.

All of the streets shown in the figure are three-lane, two-way streets with a center left-turn lane and through lanes on the outsides. The intersection has no grades on any approach. Since this paper has already established that the through

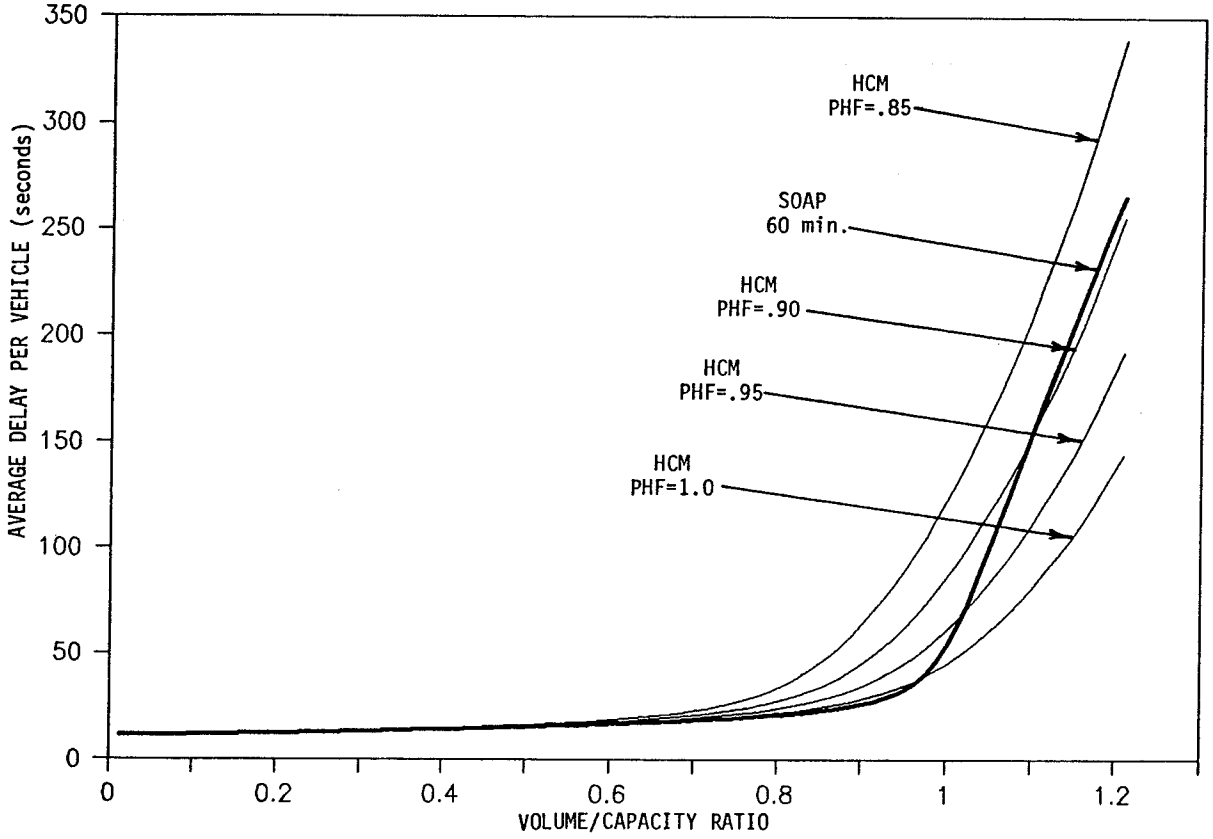


FIGURE 3 Comparison of SOAP's delay model for a 60-minute period and the HCM delay model for various peak-hour factor values.

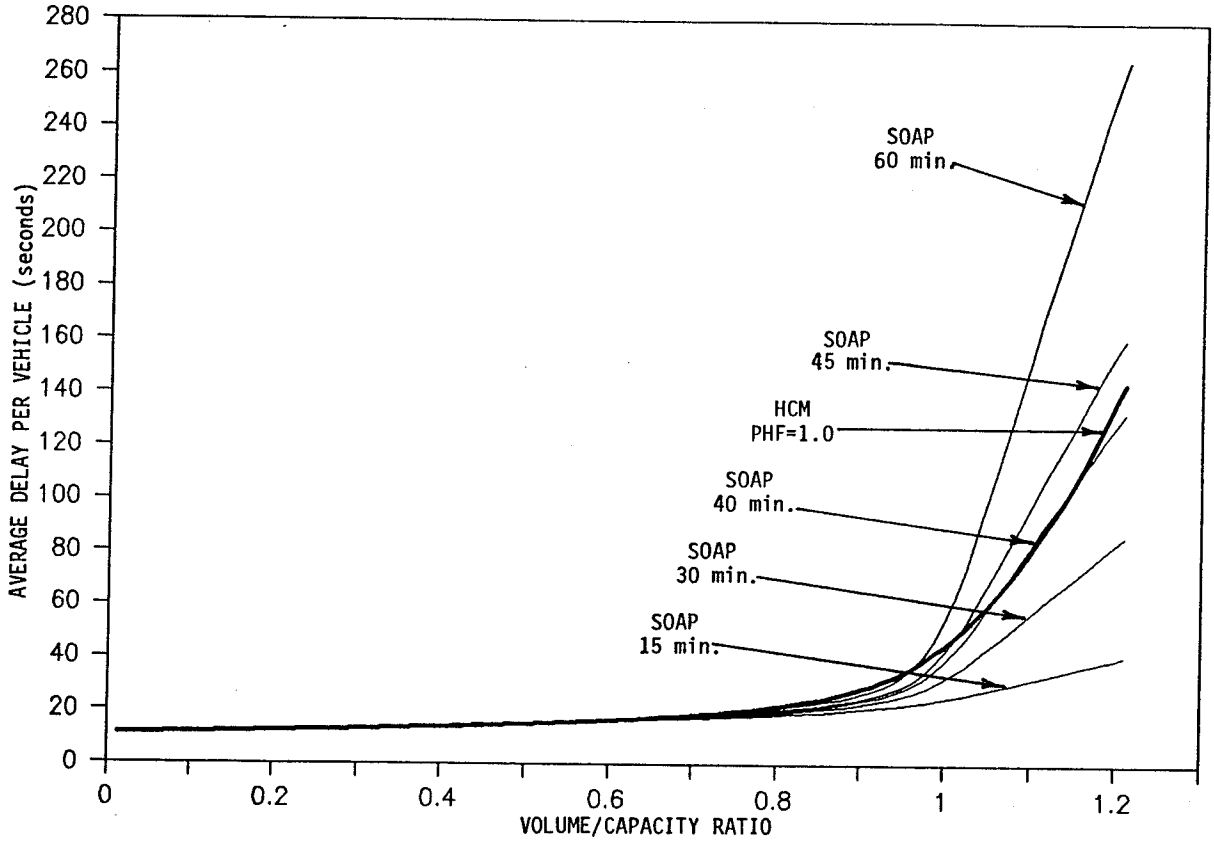


FIGURE 4 Comparison of the SOAP and HCM delay models for various period lengths.

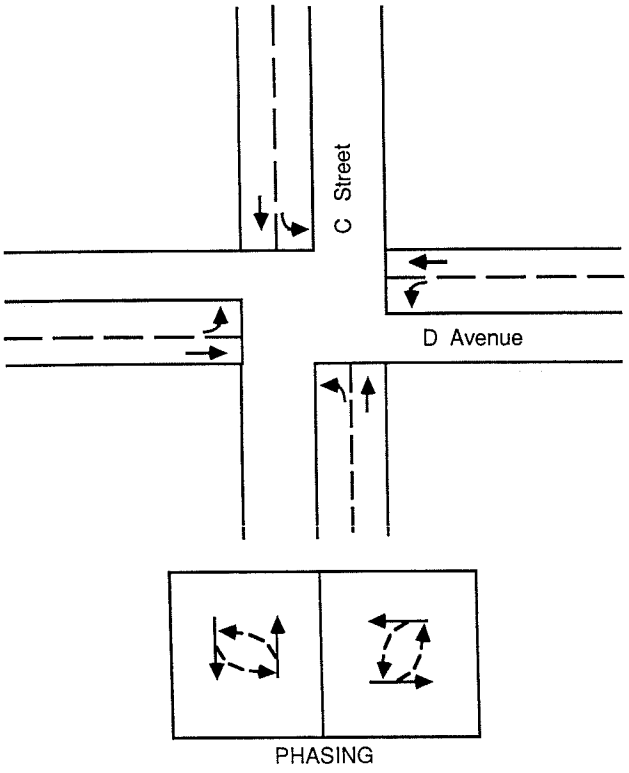


FIGURE 5 Intersection layout and phasing pattern for the permitted left-turn example.

movements behave similarly in the two models, the example only considers the effects of opposing traffic on the left-turning capacity.

Comparison of Results

Figure 6 compares the permitted left-turn capacities for the two programs. While the models have different shapes, they are still quite similar. Both of them converge to two sneakers per cycle when the opposing traffic reaches 100 percent saturation. The values are alike because both models use the same equation for unsaturated green time and because the models for the permitted saturation flow rate are very similar.

The HCM defines the left-turn factor for a shared left-through lane or an exclusive left-turn lane by the following equation:

$$f_m = \frac{g_f}{g} + \frac{g_u}{g}\left\{\frac{1}{1 + P_L(E_L - 1)}\right\} + \frac{2}{g}\{1 + P_L\} \quad (2)$$

where

f_m = left-turn factor,
g_f = portion of the green phase during which through vehicles may move in a shared lane until the first left-turn vehicle arrives,
g_u = portion of green not blocked by clearing of the opposing queue,
P_L = proportion of left turns in a shared lane, and
E_L = through vehicle equivalent for opposed left turns.

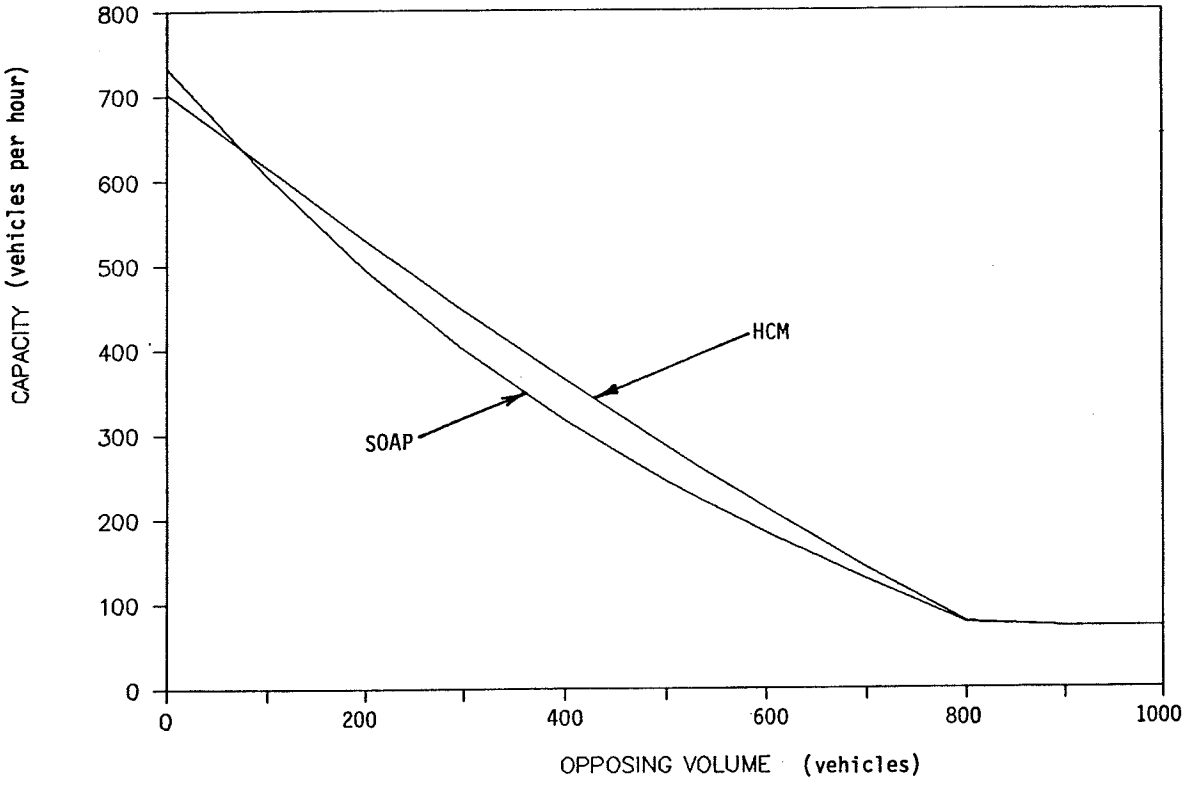


FIGURE 6 Comparison of permitted left-turn capacities for SOAP and the HCM.

However, for exclusive left-turn lanes, $P_L = 1.00$ and $g_f = 0$. Therefore, the equation simplifies to the following:

$$f_m = \frac{g_u}{g}\frac{(1{,}400 - V_o)}{1{,}800} + \frac{4}{g} \quad (3)$$

where V_o is the opposing volume.

In SOAP, the saturation flow rate for unprotected left turns is computed based on the Netsim model, which was developed by Nemeth and Mekenson (*6*). The Netsim model considers the opposing traffic and the number of opposing lanes as follows:

For a single opposing lane:

$$S_L = 1{,}404 - 1.632\,V_o + .0008347\,V_o^2 - .0000002138\,V_o^3 \quad (4)$$

and

For multiple opposing lanes:

$$S_L = 1{,}393 - 1.734\,V_o + .0009173\,V_o^2 - .0000001955\,V_o^3 \quad (5)$$

where S_L is the permitted left-turn saturation flow rate.

The unsaturated flow rate is defined by the following equation in SOAP:

$$g_u = g_o - g_{so} \quad (6)$$

where

g_u = unsaturated green time
g_o = green time for the opposing traffic, and
g_{so} = saturated green time for the opposing queue.

In Equation 6, g_{so} is defined by

$$V_o(C - g_o)/(S_o - V_o) \quad (7)$$

where S_o is the saturation flow rate of the opposing traffic. Therefore,

$$g = g_o - V_o(C - g_o)/(S_o - V_o) \quad (8)$$

which simplifies to

$$Y_o = V_o/S_o \quad (9)$$

where Y_o is the flow ratio of the opposing traffic.

This equation is exactly the same as the unsaturated green model for the HCM provided that $g_o = g$. In cases where $g_o \neq g$, the HCM does not specify which green time to use; however, it seems to suggest using the green time of the permitted movement.

It was concluded from this example that the left-turn saturation flow rates for SOAP and the HCM are more or less equivalent. Because TRANSYT-7F bases its treatment of permitted left turns on a time scan deterministic simulation, no comparisons with that model were performed.

PROTECTED PLUS PERMITTED LEFT TURNS

When a left turn is made on both a protected phase and a permitted phase, the modeling process becomes complex. The three techniques differ substantially:

- The HCM requires a user-specified split of the left-turning volume between the permitted and protected phases.
- SOAP makes an internal adjustment to increase the effective green time in proportion to the saturation flow rates for the two phases. This eliminates the need for a user-specified volume split, by equalizing the degree of saturation on both phases.
- TRANSYT accumulates arrivals on a step-by-step basis and releases the left-turning vehicles based on the instantaneous volume of the conflicting movement. This is a detailed deterministic simulation process.

Of the three models, the TRANSYT technique appears to offer the most rational formulation, but it can only be implemented by a relatively complex iterative algorithm that is not suited to the HCM format.

Problems With Existing Models

The treatment of this situation is one of the significant weaknesses of the HCM. An iterative technique is suggested, beginning with the assumption that all flow occurs during the protected phase. If this results in a V/C ratio that is too high, some volume may be assigned to the permitted interval (up to the capacity of the permitted phase).

However, vehicles assigned to the permitted phase are subsequently ignored in the delay calculations. Furthermore, the g/C ratio used in the delay calculations is the g/C total of both the protected and permitted phases. This underestimates the delay of these left-turning vehicles and results in an overly optimistic level of service for the intersection as a whole. FHWA's *Highway Capacity Software User's Manual* (HCS) (*7*) allows three options for the protected plus permitted phasing:

1. Assign no vehicles to the permitted phase,
2. Assign the maximum number of vehicles to the permitted phase, or
3. Assign the vehicles to the permitted phase such that the V/C ratios for the permitted and the protected phases are equal.

Of these three options, the third seems the most defensible. The first option ignores the additional capacity of the permitted movement and therefore uses a V/C value that is too high. The second option ignores a volume equal to the capacity of the permitted movement calculated as if the left-turn movement was unopposed (i.e., the upper limit of the permitted capacity). The equation for the capacity of the permitted phase multiplies the saturation flow rate for permitted left turns by the g/C ratio for the entire permitted phase rather than using the ratio of g_u (the unsaturated green) to the cycle length. As noted by Bonneson and McCoy (*8*), this yields a permitted capacity that is unreasonably high unless the opposing volume is negligible.

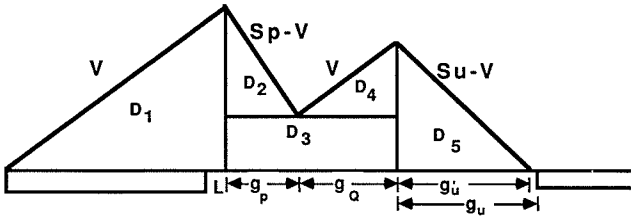
Unlike the HCM, SOAP does not ignore the permitted turning vehicles. It modifies the saturation flow rate in the protected phase to account for the flow during the permitted phase when computing the capacity of the protected plus permitted left turn. The saturation flow rate for the permitted left-turning movement is computed in the same manner as that for permitted-only left turns. This rate is then multiplied by the *g/C* ratio for the permitted movement to obtain the permitted capacity. The protected capacity is then adjusted for the permitted capacity, and the flow is assumed to occur in the protected interval. The saturation flow rate of the protected interval is modified, and this higher saturation flow rate is then used in conjunction with the *g/C* ratio of the protected interval to determine the left-turn delay. This method seems more reasonable than ignoring the permitted flow; however, by using the *g/C* ratio of the protected interval only, the delay values are much higher.

A Deterministic Queuing Model

Both SOAP and the HCM could improve their analysis of protected and permitted operations by using a deterministic queuing approach of the type shown in Figure 7. This queuing model computes only the uniform delay component. It assumes that the second term of the HCM delay equation is appropriate and proposes a new first term for the case of protected plus permitted left turns. The model further assumes that the methodology of the supplemental worksheet for permitted left turns applies to the permitted portion of the protected plus permitted left movement. The queuing model's principal advantage over the other models is that it accounts realistically for the complexity of this type of movement.

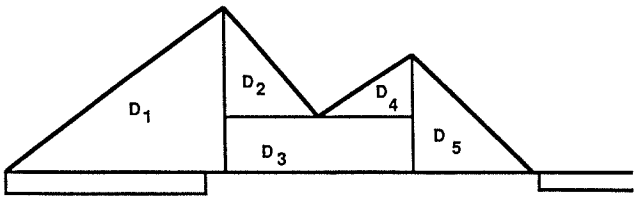
Figure 8 shows the components of the protected plus permitted left-turn movement:

- The red portion of the cycle, when the arriving vehicles are queued;
- The protected left-turn interval when the queued vehicles discharge at their saturation flow rate;



$$d_1 = .76 \left\{ \frac{1}{2}(R+L)^2 V + \frac{1}{2}(S_p - V) g_p'^2 + (g_p + g_Q)\left[(R+L)V - (S_p - V)g_p\right] + \frac{1}{2} V g_Q^2 + \frac{1}{2}(S_u - V) g_u'^2 \right\} / VC$$

FIGURE 7 Delay model for protected plus permitted left turns.



$$D_1 = \frac{1}{2}(R+L)^2 V$$

$$D_2 = \frac{1}{2}(S_p - V) g_p'^2$$

$$D_3 = (g_p + g_Q)\left[(R+L)V - (S_p - V)g_p\right]$$

$$D_4 = \frac{1}{2} V g_Q^2$$

$$D_5 = \frac{1}{2}(S_u - V) g_u'^2$$

FIGURE 8 Components of the protected plus permitted delay model.

- The beginning of the permitted interval, when the vehicles must yield to the oncoming traffic that is discharging from the opposing queue at the saturation flow rate; and
- The portion of the permitted interval during which the vehicles must accept gaps in the oncoming traffic that is entering the intersection at the arrival rate.

To help implement this complex model, a worksheet similar to the one for permitted left turns was developed (see Figure 9). The model involves the following steps:

1. *Input Variables*. The first nine rows of the worksheet summarize the data needed to compute the delay for protected plus permitted left turns:
   a. Cycle length, C (sec);
   b. Red time, R (sec);
   c. Start-up lost time, L (sec);
   d. Effective green time for the protected interval, g_p (sec);
   e. Effective green time for the permitted interval, g (sec);
   f. Left-turn flow rate, V (vph);
   g. Saturation flow rate for the protected left turns, S_p (vphg);
   h. Opposing volume, V_o (vph); and
   i. Opposing saturation flow rate, S_o (vphg).
2. *Computations*. The lower portion of the worksheet contains a series of equations to compute the delay for the various portions of the cycle:
   a. Flow ratio of the opposing traffic (Y_o):

$$Y_o = \frac{V_o}{S_o}$$

   b. Unsaturated green time (g_u):

$$g_u = \frac{g - CY_o}{1 - Y_o}$$

SUPPLEMENTAL WORKSHEET FOR PROTECTED PLUS PERMITTED LEFT TURNS				
Input Variables	**EB**	**WB**	**NB**	**SB**
Cycle Length, C (sec)				
Red Time, R (sec)				
Start-Up Lost Time, L (sec)				
Effective Green Protected Interval, g_p (sec)				
Effective Green Permitted Interval, g (sec)				
Left-Turn Flow Rate, V (VPH)				
Saturation Flow Rate for Protected Lefts, S_P				
Opposing Volume, V_o (VPH)				
Opposing Saturation Flow Rate, S_o (VPHG)				
COMPUTATIONS	**EB**	**WB**	**NB**	**SB**
$Y_o = V_o / S_o$				
$g_u = \frac{g - C Y_o}{1 - Y_o}$				
$g_Q = g - g_u$				
$f_{LT} = \frac{g_u}{g} \frac{1400 - V_o}{1800} + 4/g$				
$S_u = f_{LT} * S_P$				
$g_p' = \frac{V(R+L)}{S_p - V}$				
$g_u' = \frac{(R+L)V - (S_p - V) g_p' + g_Q V}{S_u - V}$				
$D_1 = \frac{1}{2}(R+L)^2 V$				
$D_2 = \frac{1}{2}(S_p - V) g_p'^2$				
$D_3 = (g_p + g_Q)\left[(R+L)V - (S_p - V) g_p\right]$				
$D_4 = \frac{1}{2} V g_Q^2$				
$D_5 = \frac{1}{2}(S_u - V) g_u'^2$				
$d_1 = .76(\sum_{i=1}^{5} D_i) / VC$				
NOTES: IF $g_p' > g_p$ THEN $g_p' = g_p$; IF $D_3 < O$ THEN $D_3 = O$; IF $g_u' > g_u$ THEN $g_u' = g_u$				

FIGURE 9 Worksheet for protected plus permitted left turns.

The green time used in this equation is the green time for the opposing traffic ($g_u \geq 0$).

c. Green time that is blocked by the clearance of the opposing queue (g_Q):

$$g_Q = g - g_u$$

d. Permitted left-turn factor (f_{LT}):

$$f_{LT} = \frac{g_u}{g}\frac{(1{,}400 - V_o)}{1{,}800} + \frac{4}{g}$$

e. Saturation flow rate for the permitted left portion (S_u):

$$S_u = f_{LT} * S_p$$

f. The portion of the protected green time that is used, g_p' is computed as follows:

$$g_p' = \frac{V(R + L)}{S_p - V}$$

(Note: g_p' must be in the range $0 \leq g_p' \leq g_p$.)

g. The portion of the unsaturated green time that is used, g_u':

$$g_u' = \frac{(R + L)V - (S_p - V)g_p' + g_Q V}{S_u - V}$$

(Note: g_u' must be in the range $0 \leq g_u' \leq g_u$.)

h. First delay term (D_1):

$$D_1 = \frac{1}{2}(R + L)^2 V$$

i. Second delay term (D_2):

$$D_2 = \frac{1}{2}(S_p - V)g_p'^2$$

j. Third delay term (D_3):

$$D_3 = (g_p + g_Q)\,[(R + L)V - (S_p - V)g_p]$$

(Note: $D_3 = 0$ if $D_3 < 0$.)

k. Fourth delay term (D_4):

$$D_4 = \frac{1}{2} V g_Q^2$$

l. Fifth delay term (D_5):

$$D_5 = \frac{1}{2}(S_u - V)g_u'^2$$

m. The total delay must be converted to the average stopped delay per vehicle to be consistent with the HCM methodology. The average stopped delay (d_1) is computed from the following equation:

$$d_1 = .76\frac{\left(\sum_{i=1}^{5} D_i\right)}{VC}$$

Since the queuing model attempts to account for all components of the protected plus permitted left movement, it should yield much more realistic delay values than SOAP or the HCM.

This model assumes protected plus permitted left turns (leading left-turn protection). A similar model can be used for permitted plus protected turns (lagging left-turn protection). The worksheet for lagging protection is shown in Figure 10.

A comparison of leading protection and lagging protection is shown in Figure 11. The figure shows that the delay values for lagging protection would be much higher than those for leading protection. While the first, second, fourth, and fifth terms are identical, the third term is much greater for lagging protection. This is true because one cycle of the leading protection case can essentially be viewed as two short cycles (red-green-effective and red-green), and shorter cycles yield reduced delays in the first term of the delay equation. However, neither SOAP nor the HCS computes any difference in delay between leading and lagging protection. This issue is critical since the delay values can differ significantly and many models used to design and evaluate signal timing do not consider this difference.

Example Problem

To compare the left-turn models used in SOAP and the HCS with the queuing model for protected plus permitted left turns, a simple example is considered (see Figure 12). This is the same intersection as in the permitted left-turn example, but the phasing has been changed to add left-turn protection. For this example, all through movements have a volume of 500 vph.

Comparison of Results

The results from SOAP, the HCS, TRANSYT, and the queuing model are shown in Figure 13 for leading protection and in Figure 14 for lagging protection.

Neither SOAP nor the HCM compares favorably with the proposed model. The first protected plus permitted treatment from the HCS becomes oversaturated quickly since the permitted capacity is totally ignored. For the second HCS option, the delay stays constant over the full range of left-turning volume. This is because the permitted capacity is overestimated by incorrectly using the entire permitted green time (as mentioned above). The third case from the HCS is more reasonable than the other HCS solutions but is significantly lower than either TRANSYT or the queuing model. Again, this is due to the overestimation of the permitted capacity.

SUPPLEMENTAL WORKSHEET FOR PERMITTED PLUS PROTECTED LEFT TURNS				
Input Variables	**EB**	**WB**	**NB**	**SB**
Cycle Length , C (sec)				
Red Time , R (sec)				
Start - Up Lost Time , L (sec)				
Effective Green Protected Interval , g_p (sec)				
Effective Green Permitted Interval , g (sec)				
Left - Turn Flow Rate , V (VPH)				
Saturation Flow Rate for Protected Lefts , S_p				
Opposing Volume , V_o (VPH)				
Opposing Saturation Flow Rate , S_o (VPHG)				
COMPUTATIONS	**EB**	**WB**	**NB**	**SB**
$Y_o = V_o / S_o$				
$g_u = \frac{g - C Y_o}{1 - Y_o}$				
$g_Q = g - g_u$				
$f_{LT} = \frac{g_u}{g} \frac{1400 - V_o}{1800} + 4/g$				
$S_u = f_{LT} * S_P$				
$g_p' = \frac{V (R + L + g_Q) - g_u (S_u - V)}{S_p - V}$				
$g_u' = \frac{(R+L+g_Q) V_p}{S_u - V}$				
$D_1 = \frac{1}{2} (R + L)^2 V$				
$D_2 = \frac{1}{2} (S_p - V) g_p'^2$				
$D_3 = (R + L) g_Q V + (S_p - V) g_u g_p'$				
$D_4 = \frac{1}{2} V g_Q^2$				
$D_5 = \frac{1}{2} (S_u - V) g_u'^2$				
$d_1 = .76 (\sum_{i=1}^{5} D_i) / VC$				
NOTES: IF $g_p' > g_p$ THEN $g_p' = g_p$; IF $g_u' > g_u$ THEN $g_u' = g_u$; IF $g_u' \leq g_u$ THEN $g_p' = 0$				

FIGURE 10 Worksheet for permitted plus protected left turns.

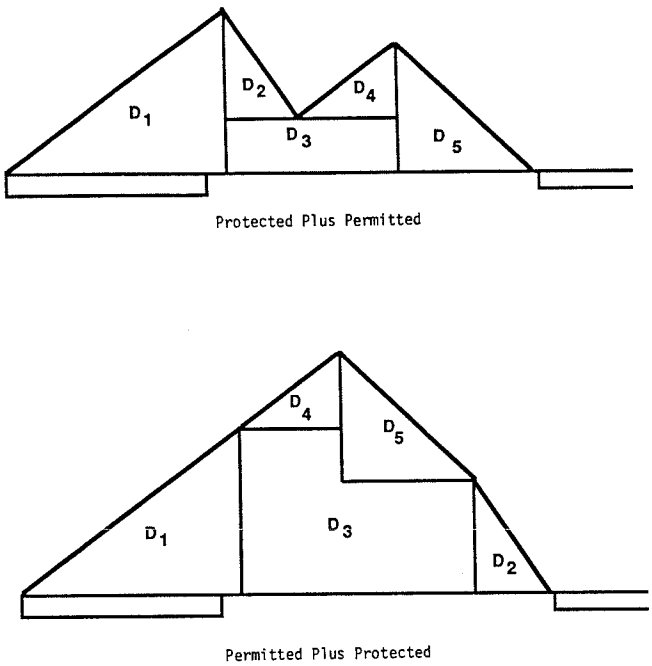


FIGURE 11 Comparison of protected plus permitted versus permitted plus protected queuing characteristics.

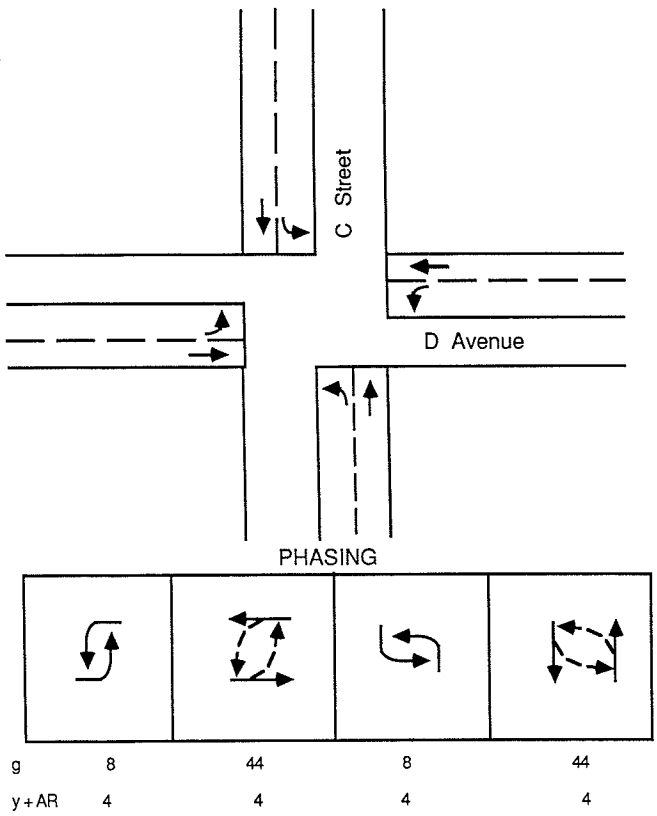


FIGURE 12 Intersection layout and phasing pattern for the protected plus permitted left-turn example.

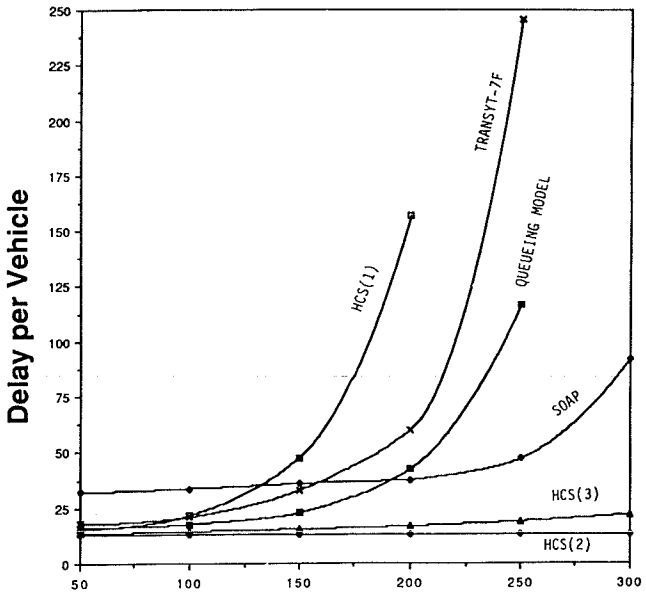


FIGURE 13 Leading left-turn delay.

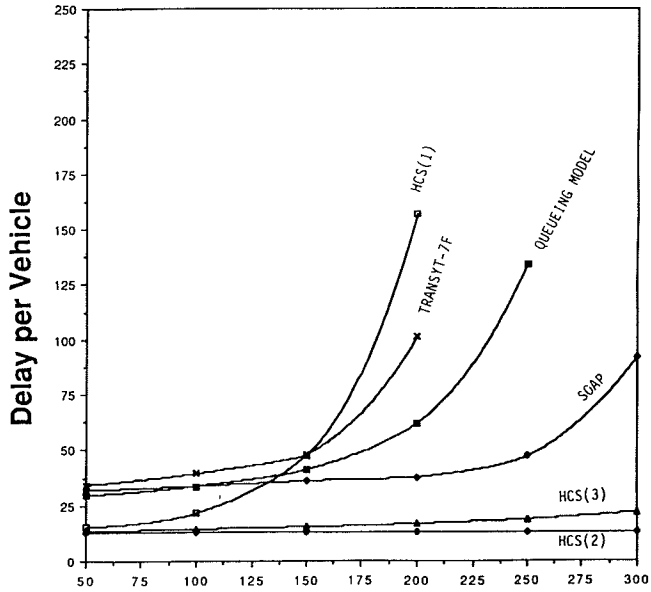


FIGURE 14 Lagging left-turn delay.

The queuing model does seem to compare well with TRANSYT. The general shape of the curves is similar and, more important, both models reflect the difference between leading and lagging protection. In contrast, both SOAP and the HCS results are identical (incorrectly) for both left-turn protection schemes.

The general concept of the queuing model is not new. A similar formulation was suggested by Bonneson and McCoy (*8*) and the PASSER II-87 signal optimization program (*9*). It is also the basis for delay calculation in SICAP (*10*), a computer program that implements the HCM calculations on a microcomputer.

CONCLUSIONS AND RECOMMENDATIONS

This research resulted in some interesting observations. It is apparent that the delay model used in the 1985 *Highway Capacity Manual* agrees very well with the SOAP and TRANSYT models for undersaturated operation. There were significant differences when conditions were oversaturated; however, it was possible to make the delay values agree by the proper choice of period length and peak-hour factor. It is also clear that neither of these models should be used to estimate the delay for long periods of oversaturation or high degrees of oversaturation.

The TRANSYT model, because of its detailed breakdown of the cycle, appeared to treat the case of permitted and protected left turns in a more reasonable manner than either SOAP or the HCM. While it is not practical to incorporate the TRANSYT model in either of the other techniques, it is relatively easy to incorporate the queuing model described in this paper. The queuing model offers two important advantages over the HCM signalized intersection methodology:

1. It eliminates the need for user-specified volume split.
2. It differentiates between leading and lagging protection.

With this in mind, it is recommended that the models described in this paper for protected plus permitted left turns and permitted plus protected left turns be incorporated into the HCM methodology.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. K. G. Courage, C. E. Wallace, and R. Alqasem. Modeling the Effects of Traffic Signal Progression on Delay. In *Transportation Research Record 1194*, TRB, National Research Council, Washington, D.C., 1988.
3. *SOAP84 User's Manual.* Report FHWA-IP- 85-7. FHWA, U.S. Department of Transportation, 1985.
4. D. I. Robertson and P. Gower. User Guide to TRANSYT Version 6. Supplementary Report 255. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1977.
5. F. V. Webster. Traffic Signal Settings. Her Majesty's Stationary Office, London, England, 1958.
6. Zoltan Nemeth and J. R. Mekemson. *Guidelines for Left Turn Treatments at Signal Controlled Intersections.* Report FHWA/OH-83/003. FHWA, U.S. Department of Transportation, 1983.
7. *Highway Capacity Software User's Manual.* FHWA, U.S. Department of Transportation, 1987.
8. J. Bonneson and P. McCoy. Operational Analysis of Exclusive Left-Turn Lanes with Protected/Permitted Phasing. In *Transportation Research Record 1114*, TRB, National Research Council, Washington, D.C., 1987.
9. *PASSER II-87 User's Manual.* Texas State Department of Highways and Public Transportation, Austin, Texas, 1988.
10. J. A. Bonneson. *SICAP User's Manual.* Bryan, Texas, 1985.

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

Capacity of Shared Left-Turn Lanes—A Simplified Approach

HERBERT S. LEVINSON

This paper develops a simple and readily usable formula for estimating the capacity of a shared left-turn lane. It also derives a left-turn blockage factor for use in the formula. The capacity of a through lane at a signalized intersection is reduced by factors that reflect (a) opposing left turns or (b) the blockage effect of left turns on through traffic in the same lane. The capacity of a shared lane represents the minimum of these computations. Thus, the capacity of a shared lane is reduced by the number of left turns using the lane, as well as the traffic volumes in the opposing direction. Opposing left turns reduce the capacity of both shared left-turn lanes and through lanes. When there are five or more left turns per cycle in a shared lane, they preempt that lane for all practical purposes. Short signal cycles appear preferable where shared left-turn lanes dominate.

Computing the capacity of shared left-turn lanes continues to be a complex and elusive task. It is generally agreed that (a) left-turning vehicles have longer headways than through traffic and (b) opposing through vehicles reduce left-turn capacity. There is growing recognition that left turns may impede (or block) through vehicles in the shared lane and that the combined effects of opposing traffic flow and left turns reduce the capacity of a shared lane. While these factors are contained in various capacity guides, the methodologies are complex. There are no simple, practical relationships for ready use.

This paper develops an analytical approach to the capacity of left-turn lanes that considers the nature, magnitude, and interaction of the opposing traffic. It derives simple and readily usable formulas for estimating the capacity of a shared left-turn lane, and it shows how the formulas can be applied. The formulas relate capacity to traditional variables such as green time, cycle length, starting-clearing losses, and vehicle headways (or saturation flows). However, they also include the number of left turns, the proportion of through traffic that is blocked by left turns, and the opposing traffic flow.

The formulas are based on the assumption that only one vehicle can cross a conflict point at any given time. Thus, they build on the critical movement analysis of vehicles across a conflict point. They also build on the 1985 *Highway Capacity Manual* (HCM) concept of "free" and "blocked" green time (*1*). They are interactive since they reflect the impact of the opposing traffic flow.

BACKGROUND

The basic formula for intersection capacity under ideal conditions relates capacity to green time, lost time, and vehicle headways (or saturation flow). The number of through passenger vehicles per lane per cycle can be obtained from the following formula:

$$c = \frac{G - \text{lost time per cycle}}{h_i} = \frac{g}{h_i} \quad (1)$$

where

G = total green + yellow time per cycle,
g = effective green time per cycle,
h_i = ideal headway at saturation flow (sec),
lost time = starting and clearing lost time (sec), and
c = capacity (veh/lane/cycle).

This formula is modified in practice to reflect prevailing roadway and traffic conditions. A series of adjustment (or reductive) factors are applied to account for lane width, traffic composition, and turn interferences.

The general formula becomes

$$c = \frac{g}{h_i} \prod F_i \quad (2)$$

where ΠF_i equals the product of adjustment factors.

This formula can be rearranged to separate the reductive effects of left turns from other effects. One possible form for the formula is as follows:

$$c = \frac{g}{h_i} \prod F_i' - \frac{\text{lost time per cycle due to left turns}}{h_i} \quad (3)$$

Therefore,

$$c = \frac{g}{h} - \text{lost vehicles per cycle due to left turns} \quad (4)$$

where

$\prod F_i'$ = product of all reductive factors other than left turns, and
h = actual or adjusted headway based on these factors.

Hence,

$$c = \frac{g}{h_i} \prod F_i' \quad (4')$$

This concept reduces the capacity for time losses resulting from left-turn delays rather than adjusting the saturation flow.

Department of Mechanical & Industrial Engineering, Polytechnic University of New York, Brooklyn, N.Y. 11201.

The following sections develop a series of analytical expressions for computing the lost capacity due to left turns.

CONCEPTS AND ASSUMPTIONS

Two sets of formulas were derived for what might be termed a "conflict/blockage" approach. They were keyed to traffic operations through a typical right-angle, two-phase, signal-controlled intersection (see Figure 1) and reflect the following assumptions:

- The time required across a point of conflict represents the sum of the time required by conflicting lane volumes. Conversely, a given green time is shared by the through and conflicting left-turn flows.
- Where through vehicles conflict with opposing left turns, the through vehicles move first.
- Where left turns block through vehicles in the same lane, cars move in the following sequence:
  1. Unblocked through and opposing traffic,
  2. Remainder of opposing traffic,
  3. Left turn(s), and
  4. Blocked through traffic.
- Traffic in each direction moves within a platoon.
- Where through traffic and opposing left turns conflict, some of the left turns will move on the clearance phase if no other time is available. These are defined as "sneakers."
- On opposing single-lane approaches, opposing left turns will move at the same time.
- Right turns form part of the through traffic.
- The headways for through and unblocked left-turning vehicles are the same.

These assumptions result in two sets of criteria for estimating the reductive effect of left turns on capacity (see Figure 2). The capacity of any given approach lane reflects the minimum capacity obtained by applying either the conflict or blockage criteria:

- *Conflict.* Green time is shared between traffic in one direction and left turns in the opposing direction.
- *Blockage.* Green time is lost as a result of delayed left turns waiting for opposing through traffic to clear.

FORMULA DEVELOPMENT

Conflict

The formula for estimating the effect of opposing left turns on the capacity of a lane is straightforward. If there were no opposing left turns, the capacity would be g/h. However, since left turns preempt some of the available green time, this amount should be reduced by the number of opposing left turns per cycle (l_2) and increased by the number of sneakers per cycle (s_2). This leads to the following formula:

$$c\ (\text{conflict}) = \frac{g}{h} - l_2 + s_2 \tag{5}$$

subject to

$$s_2 \leq l_2 \leq \frac{g}{h}$$

$$\left(\text{Since } l_2 \leq \frac{g}{h},\ s_2 \text{ is also} \leq \frac{g}{h}\right)$$

This formula is a restatement of the critical lane analysis and applies to both shared and nonshared lanes. It also applies when the opposing left turns operate from an exclusive lane.

Blockage

The capacity of a shared left-turn lane can be viewed as having three basic components in terms of likely blockage effects. These are

1. The through vehicles that precede the first left turn,
2. The through vehicles that follow and are blocked by the left turn, and
3. The left turns.

The expected capacity per cycle is the sum of these three capacities, each weighted by its relative share of the total. This relationship can be expressed as follows:

$$c_s = p_1c_1 + p_2c_2 + p_3c_3 \tag{6}$$

where

c_s = expected capacity of the shared lane (cars per cycle) (blockage),
c_1 = capacity of shared-lane through vehicles that are not delayed by left turns (cars per cycle),

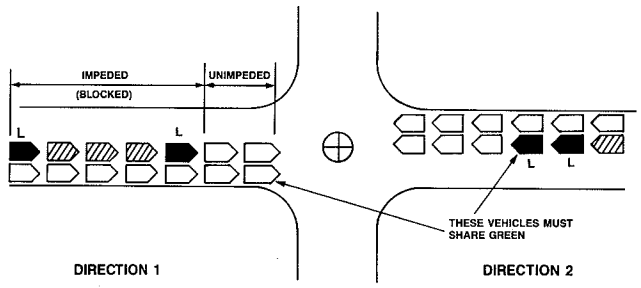


FIGURE 1 Intersection operations.

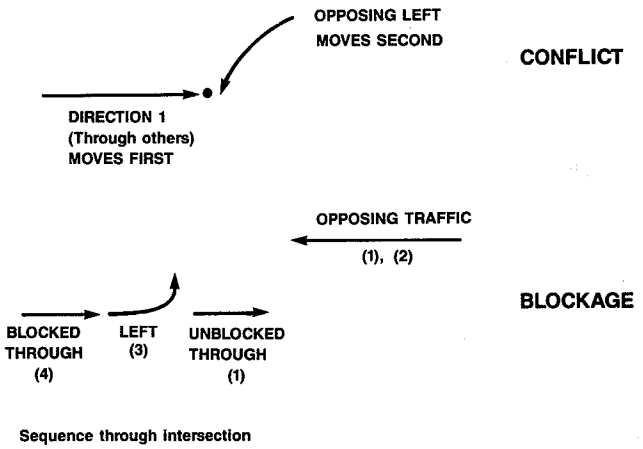


FIGURE 2 The concepts of conflict and blockage.

c_2 = capacity of shared-lane through vehicles that are delayed by left turns (cars per cycle),
c_3 = capacity of shared-lane left turns (left turns are delayed by opposing traffic) (cars per cycle),
p_1 = proportion of capacity used by unimpeded through traffic,
p_2 = proportion of capacity used by impeded (or delayed) through traffic, and
p_3 = proportion of capacity used by left turns.

The proportion of capacity p_1, p_2, and p_3 can be expressed in terms of the total capacity (c_s), and left turns per cycle (l), and the proportion of through vehicles delayed by left turns (K) as follows:

$$p_3 = \frac{l}{c_s} \tag{7a}$$

$$p_2 = \left(1 - \frac{l}{c_s}\right)(K) \tag{7b}$$

$$p_1 = \left(1 - \frac{l}{c_s}\right)(1 - K) \tag{7c}$$

Substituting into Equation 6 yields the following formula:

$$c_s = \left(1 - \frac{l}{c_s}\right)(1 - K) \cdot c_1 + \left(1 - \frac{l}{c_s}\right) K \cdot c_2 + \frac{l}{c_s} c_3 \tag{8}$$

The capacities c_1, c_2, and c_3 are estimated as follows:

$$c_1 = \frac{g}{h} \quad \text{(no blockage)} \tag{9a}$$

$$c_2 = \left(\frac{g}{h} - o_L\right) \quad \text{(blockage)} \tag{9b}$$

$$c_3 = \left(\frac{g}{h} - o_L\right) \quad \text{(blockage)} \tag{9c}$$

where o_L equals the opposing traffic per lane per cycle.

The blockage condition assumes that the effective green time is shared by the opposing traffic and the blocked vehicles. It also assumes that the opposing traffic moves first and the blocked vehicles move next. This assumption is similar to that used in critical conflict analysis.

Equation 8 was solved algebraically. Equations 9a, 9b, and 9c were then substituted for c_1, c_2, and c_3, respectively. Further simplification produced the following quadratic formula for the expected capacity of a shared left-turn lane:

$$c_s = \frac{\left(\frac{g}{h} - Ko_L\right) + \left[\left(\frac{g}{h} - Ko_L\right)^2 - 4o_L\, l(1 - K)\right]^{1/2}}{2} \tag{10}$$

This formula states that the capacity of a shared lane depends on (a) the effective green time, (b) the vehicle headways, (c) the proportion of through vehicles blocked by left turns, (d) the number of left turns per cycle, and (e) the opposing traffic per lane. The blockage factor (K) depends on the left turns per cycle.

If there are no left turns, $l = 0$ and $K = 0$. The formula simplifies to

$$c_s = \frac{g}{h} \tag{11}$$

If all through vehicles are blocked by left turns, $K = 1$ and $(1 - K) = 0$. The formula simplifies to

$$c_s = \frac{g}{h} - o_L \tag{12}$$

Estimating Two Factors

Application of the formula calls for knowing the values of o_L and K. According to the HCM, for multilane approaches o_L equals [the opposing traffic divided by the number of lanes] times a lane utilization factor. When both approaches of a street have only one lane, o_L becomes the sum of the through and right-turn traffic.

The left-turn blockage factor (K) depends on the number of left turns on the shared approach. It was derived by the following two methods:

1. *Simulation*. Random number tables were used to generate the position of left turns and through vehicles in 10-car platoons. Fifty platoons were analyzed for each of four cases, in which left turns represented 10, 20, 30, and 40 percent of the total traffic in the platoon, respectively. The average number of through vehicles and the proportion of through vehicles delayed were then computed.
2. *Positional Probabilities*. Probability theory was used to estimate the likelihood of the first, second, third, and ith vehicles in line being left turns. Conditional probabilities were computed, assuming that the first left turn was the ith car in line based on sampling with replacement. The probability (p) that the first $(i - 1)$ cars were through vehicles and the ith vehicles were left turns is given by the formula:

$$\left(\frac{T}{l + T}\right)^{i-1} \cdot \frac{l}{l + T} \tag{13}$$

where T equals through vehicles per cycle.

The computed values of K are shown in Table 1 and Figure 3. The simulation and sampling methods yield similar results.

An inspection of Table 1 shows that the K values can be approximated by a single set of numbers that are a function of the number of left turns per traffic signal cycle or platoon. As shown in Table 2 and Figure 4, the suggested K values based on left turns per cycle provide a reasonable approximation of left-turn blockage for most conditions:

- When there is one left turn per cycle, approximately 40 percent of the through vehicles in the shared lane would be blocked.
- When there are three left turns per cycle, approximately 70 percent of the through vehicles in the shared lane would be blocked.

TABLE 1 COMPUTED VALUES FOR K

Left Turns as Percent of Total Traffic in Shared Lane	Length of Platoon (veh/cycle)			
	5	10	15	20
1	.029	.053	.076	.094
10	.247	.388 [.322][a]	.495	.576
20	.390	.584 [.607]	.690	.757
30	.527	.694 [.647]	.783	.834
40	.608	.761 [.747]	.834	.874
50	.670[b]	.806	.866	.900
60	.720	.838	.889	.916
70	.770	.860[b]	.905	.929
80	.800	.880	.917	.938
90	—	.900	.929[b]	.945
95	—	—	—	.950

[a]Brackets signify a simulation.
[b]Computed value adjusted slightly.

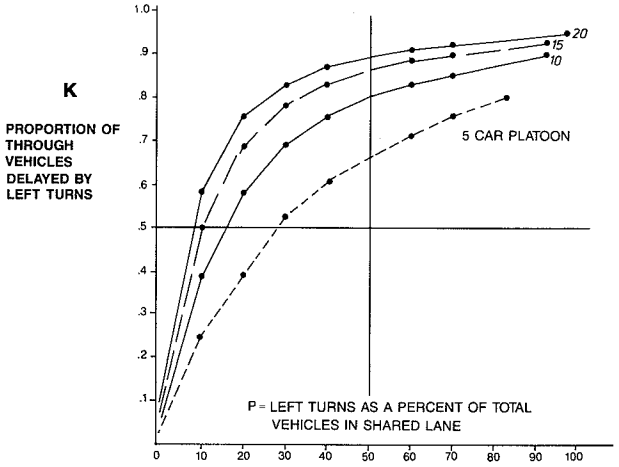


FIGURE 3 K as a function of p.

• When there are five left turns per cycle, approximately 80 percent of the through vehicles in the shared lane would be blocked.

These values simplify the computational procedures and are in a form suitable for use in Equation 10.

It is interesting to note that the K values are generally less than those obtained by the relationship

$$K = \frac{l}{l + 1}$$

Simplifying the Formula

The basic quadratic formula for the capacity of a shared lane as a result of blockage is straightforward to use. However, it can be closely approximated by the following simplified formula:

$$c_s = \frac{g}{h} - 1.2Ko_L \tag{14}$$

In this equation, $1.2K \leq 1$. If $1.2K > 1$, use 1.

TABLE 2 SUGGESTED K VALUES FOR APPLICATION

Left Turns Per Cycle	Computed	Suggested Value Rounded
0.5	0.25	0.25
1	0.39	0.40
2	0.58	0.60
3	0.69	0.70
4	0.76	0.75
5	0.81	0.80
6	0.84	0.84
7	0.86	0.86
8	0.88	0.88
9	0.89	0.89
10	0.90	0.90

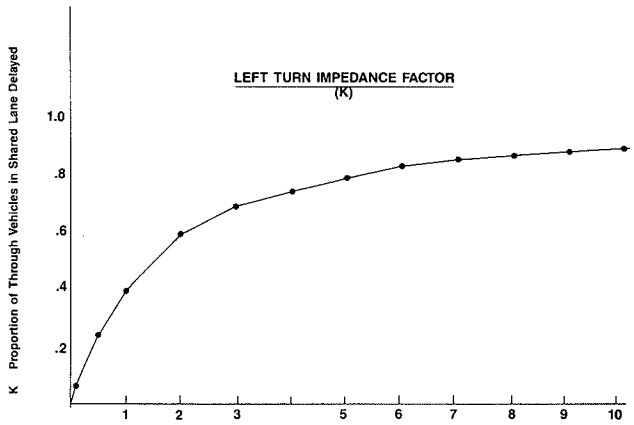


FIGURE 4 Left-turn impedance factor (K).

Sample comparisons between the basic quadratic formula and the simplified formula are shown in Table 3. In most cases, the differences are small—usually less than 0.5 cars per cycle.

The simplified formula can be expressed as

$$c_s = \frac{g}{h} - Bo_L \tag{15}$$

where B, the modified blockage factor, equals $1.2K$ or 1, whichever is smaller.

Table 4 and Figure 5 show how the modified blockage factors (B) compare with the initial K factors. These exhibits can be used in applying Equation 15.

FORMULAS FOR APPLICATION

The capacity of any traffic lane will be reduced by two factors:

1. The time required by the opposing left turns, and
2. The time loss that results from blockage by left turns in the same direction.

The lane capacity will represent the minimum capacity resulting from these two conditions. Suggested formulas reflecting these conditions are shown in Table 5. These for-

TABLE 3 SELECTED COMPARISONS OF THE BASIC AND SIMPLIFIED FORMULAS

Left Turns Per Cycle	Vehicles Per Cycle: Basic Quadratic Formula (Eq. 13)	Vehicles Per Cycle: Simplified Formula (Eq. 15)
$g/h = 12, o_L = 6$		
0.5	10.28	10.20
1	9.21	9.12
2	7.78	7.68
3	7.03	6.46
4	6.59	6.60
5	6.24	6.24
6	6.00	6.00
$g/h = 12, o_L = 8$		
0.5	9.69	9.60
1	8.20	8.16
2	6.16	6.24
3	4.94	5.28
4	4.00	4.80
$g/h = 15, o_L = 6$		
0.5	13.33	13.20
1	12.31	12.12
2	10.96	10.68
3	10.27	9.46
4	9.89	9.60
5	9.57	9.24
6	9.34	9.00
7	9.20	9.00
8	9.04	9.00
9	9.00	9.00
$g/h = 15, o_L = 8$		
0.5	12.76	12.60
1	11.38	11.16
2	9.53	9.24
3	8.56	8.28
4	8.00	7.80
5	7.54	7.32
6	7.22	7.00
7	7.00	7.00
$g/h = 15, o_L = 10$		
0.5	12.19	12.00
1	10.42	10.20
2	8.00	7.80
3	6.65	6.60
4	5.77	6.00
5	5.00	5.40
$g/h = 20, o_L = 6$		
0.5	18.38	18.20
1	17.39	17.12
2	16.10	15.68
3	15.45	14.96
4	15.10	14.60
5	14.74	14.24
6	14.56	14.00
7	14.43	14.00
8	14.32	14.00
9	14.24	14.00
10	14.18	14.00
$g/h = 20, o_L = 8$		
0.5	17.83	17.60
1	16.51	16.16
2	14.77	14.24
3	13.88	13.28
4	13.40	12.80
5	12.98	12.32
6	12.67	12.00
7	12.49	12.00
8	12.34	12.00
9	12.23	12.00
10	12.14	12.00

TABLE 3 (*continued next column*)

TABLE 3 (*continued*)

Left Turns Per Cycle	Vehicles Per Cycle: Basic Quadratic Formula (Eq. 13)	Vehicles Per Cycle: Simplified Formula (Eq. 15)
$g/h = 20, o_L = 10$		
0.5	17.28	17.00
1	15.62	15.20
2	13.40	12.80
3	12.27	11.60
4	11.64	11.00
5	11.10	10.40
6	10.70	10.00
7	10.46	10.00
8	10.26	10.00
9	10.12	10.00
10	10.00	10.00

TABLE 4 INITIAL AND MODIFIED IMPEDANCE FACTORS

Left Turns Per Cycle	Initial *K* Factor	Modified *B* Factor[a]
0.5	0.25	0.300
1	0.40	0.480
2	0.60	0.720
3	0.70	0.840
4	0.75	0.900
5	0.80	0.960
6	0.84	1.000
7	0.86	1.000
8	0.88	1.000
9	0.89	1.000
10	0.90	1.000

[a]1.2*K* or 1, whichever is less.

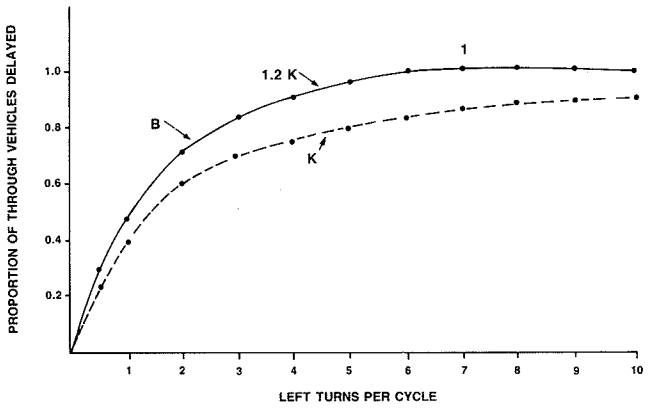


FIGURE 5 Modified impedance factor (*B*).

mulas show how the lane capacity per cycle can be computed on a lane-by-lane basis for four typical conditions:

1. Through lane with no opposing left turns or shared lane with no opposing traffic,
2. Through lane with opposing (conflicting) left turns,
3. Shared lanes with opposing traffic on two-lane streets, and
4. Shared lanes with opposing traffic on multilane streets.

TABLE 5 LANE CAPACITY FORMULAS

Case	Description	Lane Capacity Per Cycle	Constraints[a]
1.	Through lane with no opposing (conflicting) left turns or shared lane on one-way street	(A) $c = \frac{g}{h}$	
2.	Through lane with opposing (conflicting) left turns	(B) $c = \frac{g}{h} - (l_2 - s_2)$	$l_2 \geq s_2$
3.	Single shared lane (on two-lane street)	Minimum of	
	Conflict	(C) $c = \frac{g}{h} - (l_2 - l_1 - s_2)$	$(l_2 - l_1 - s_2) \geq 0$
	Blockage	(D) $c_s = \frac{g}{h} - B(o_2 - l_2)$	
4.	Shared lane on multi-lane street	Minimum of	
	Conflict	(E) $c = \frac{g}{h} - (l_2 - s_2)$	$l_2 \geq s_2$
	Blockage	(F) $c_s = \frac{g}{h} - Bo_2$	

NOTES:
g = effective green time (sec/cycle)
h = headway, adjusted for factors other than left turns (sec/veh)
l_1 = left turns per cycle in given direction
l_2 = opposing left turns per cycle
s_2 = opposing sneakers per cycle (always $\leq l_2$)
o_2 = opposing traffic per lane per cycle
B = modified blockage (impedance factor)
c = capacity (veh/lane/cycle)
c_s = capacity, veh/lane/cycle (shared lane—blockage)
[a]When these conditions are not met, apply Formula A.

Cases 3 and 4 involve the application of a pair of formulas. The formulas are similar, but Case 3 takes into account the cancelling effect of opposing left turns on a two-lane street. In both of these cases, it is necessary to check for the capacity losses caused by opposing left turns. They may be the actual capacity constraints when the number of left turns per cycle is light and the opposing left turns are heavy.

Figure 6 illustrates how the dual blockage and conflict criteria would apply. In this example, the conflict criteria apply only when the opposing left turns are substantially heavier than the left turns in the given direction of travel. Further computational examples are contained in Figure 7.

The number of sneakers per cycle can be assumed to be equal to 1 when there is a shared lane and 2 when there is an exclusive left-turn lane. However, the sneakers can never be more than the actual number of left turns.

The formulas should be applied on a lane-by-lane basis. The headways used for each lane may vary, depending on the effects of right turns and other impedances.

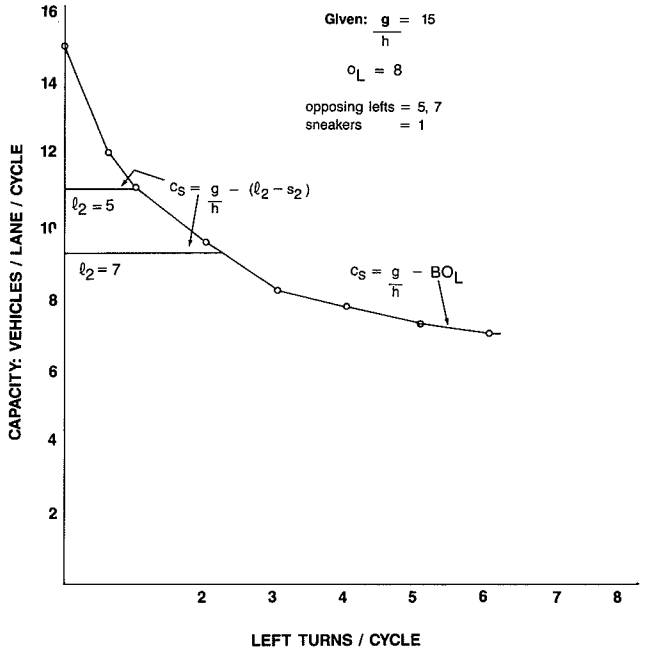


FIGURE 6 Sample computation for multilane approach showing application of dual criteria.

Leading or Lagging Green

The formulas shown in Table 5 can be used to compute the capacities of a shared lane with a leading or lagging green. The effective leading (or lagging) green time should be treated as unimpeded (see Formula A). The remaining green time should be based on Formulas C and D for a single-lane approach or Formulas E and F for a multilane approach. However, the number of left turns per cycle used for computing Formula B should be reduced proportionately to reflect the left turns moving in the leading (or lagging) period. This adjustment is computed as follows:

$$l_1' = \frac{g_2}{g_1 + g_2} l_1 \quad (16)$$

where

g_1 = effective green, normal part of phase;
g_2 = effective green, leading or lagging part of phase;
l_1 = actual left turns per cycle; and
l_1' = left turns per cycle, normal part of phase.

Hourly Capacities

The formulas can also be expressed in vehicles per hour. This is accomplished by substituting $(g/C)\,S$ for g/h, where C equals the cycle length and S equals the adjusted saturation flow.

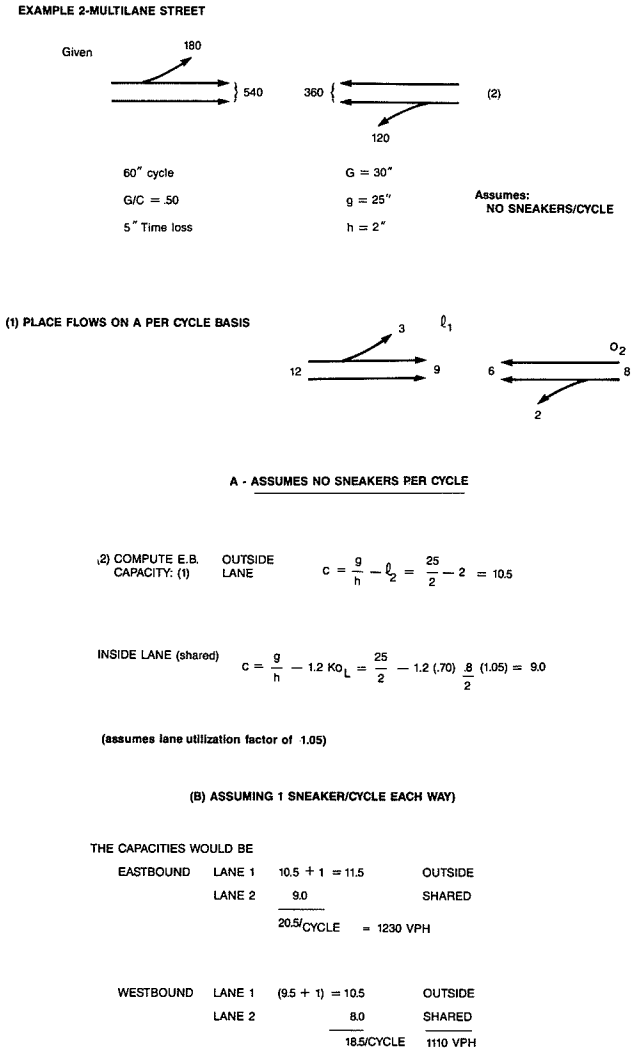


FIGURE 7 Illustrative application.

Thus, for the multilane approaches, the shared-lane capacities (Case 4) would become

$$C_s\ (\text{conflict}) = \frac{g}{C}\,S - (L_2 - S_2)$$
$$C_s\ (\text{blockage}) = \frac{g}{C}\,S - B\,O_L \qquad (17)$$

where

L_2 = opposing left turns (vph),
O_L = opposing traffic/lane/hour,
S_2 = sneakers per hour, and
C_s = shared lane capacity (vph).

COMPARISONS

Pilot comparisons were made with the capacities obtained in Example 1 of the HCM. The two analytical methods produced generally consistent results. The total intersection capacity obtained by the conflict/blockage method was about 7 percent less than that computed by the HCM (see Table 6). The multilane capacities were comparable, while the single-lane capacities produced by the suggested method were about 15 percent lower.

IMPLICATIONS

This paper has developed practical analytical approaches for dealing with left turns at intersections, particularly shared left turns. The formulas are simple and straightforward to use. They include the relevant variables, and they produce reasonable limits and values. They are user friendly and can be solved manually or by computerized methods. They apply on a lane-by-lane basis, and they produce results on the same order of magnitude as the HCM.

The formulas are dynamic in that they consider the effects of the opposing traffic flow on intersection capacity. They reflect the blockage effect of opposing vehicles in a systematic and predictable way, and they take cycle length into consideration.

A suggested blockage factor has been derived from probability and simulation analysis. However, the formulas can use other blockage factors that may emerge from further research or field tests.

The significant findings are as follows:

- The capacity of a shared left-turn lane is reduced by both the number of left turns using the lane and the traffic volume

TABLE 6 COMPARATIVE CAPACITIES (*1*)

Approach	This Paper	Example 9-1 in HCM	Percent Difference[a]
Northbound (one lane)	648	753	−14
Southbound (one lane)	669	807	−17
Eastbound (two lanes)	878	843	4
Westbound (two lanes)	951	995	−4
Total	3,146	3,398	−7.4

[a]From HCM example.

in the opposing direction. Where opposing left turns are heavy, they, too, will limit the capacity of a shared lane.

- When there are five or more left turns per cycle in a shared lane, they preempt that lane for all practical purposes.
- Short signal cycles are preferable to maximize the capacity of shared left-turn lanes since they reduce the left turns per cycle and their blockage effects. They also allow more sneakers per hour on the yellow signal.
- Capacity of a nonshared lane (through vehicles only) is reduced by the opposing left turns.

Additional field studies are desirable to verify the analysis and to adjust the suggested formulas. Their ease of application warrants such efforts. The formulas and factors provide, at minimum, a viable method for use in planning applications.

ACKNOWLEDGMENTS

Many individuals offered valuable suggestions on the analytical procedures. Special thanks are extended to Paul Menaker of Urbitran, Inc., K. Zografos of the University of Miami, and Elena Shenk of the Polytechnic University of New York.

REFERENCE

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, 1985.

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

Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations

JAMES H. BANKS

In this paper, recent Canadian work challenging long-held theories about speed-flow-concentration relationships in freeway traffic is verified and extended using data from San Diego. Major conclusions of these studies regarding near-constant free-flow speeds and flow-concentration relationships resembling an inverted V are confirmed; other past findings related to the effect of queuing on downstream free-flow speeds and the effect of secondary bottlenecks on flow-concentration relationships are not confirmed. When data are averaged across all lanes (both upstream and downstream of a bottleneck), speed-flow and flow-concentration relationships are found to be consistent with those predicted by queuing and shock wave theory. The functioning of the bottleneck studied is more complicated than had been assumed, however, and this creates further problems in the interpretation of the data. Finally, the inverted-V model of the flow-concentration relationship is shown to imply a simple and plausible model of driver behavior in which speeds and spacings are adjusted to keep the average front-to-back time gaps approximately constant until some desired maximum speed is reached.

Several recent reports dealing with freeway traffic flow have challenged long-held views about relationships among speed, flow, and measures of traffic concentration such as density or occupancy. Although there has never been complete agreement about the nature of these relationships, the representations found in the 1985 *Highway Capacity Manual* (*1*) (for instance, the speed-flow relationships depicted in Figures 3 and 4) may be said to represent the current conventional wisdom. Recent empirical work that challenges these views includes a series of reports by Hurdle and various associates (*2–4*) and a series of papers by Hall et al. (*5–8*). Both of these efforts, which were based on data from Toronto, are indebted to some degree to previous work done by Koshi et al. (*9*) in Japan.

A common point of departure for this work is concern that data may have been misinterpreted in the past due to lack of sensitivity to the effects of location on freeway flow phenomena. Specifically, both groups of Canadian researchers are concerned with the ways in which neglect of queuing phenomena could lead to misinterpretation of speed-flow-concentration data. Particular concerns include

- The effect of queues in limiting flows upstream of bottlenecks, thus creating gaps in the data;
- The implications of data-gathering and data-reduction procedures that may involve averaging data from dissimilar flow conditions; and
- Possible misinterpretation of data taken in the zone of acceleration downstream from a queue.

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

The primary concern of Hurdle and Datta (*2*), Persaud (*3*), and Persaud and Hurdle (*4*) is speed-flow relationships under free-flow conditions. Their major finding is that there is no precipitous drop in free-flow speeds as flow approaches capacity, although there may be a gradual drop. This is contrary to the popular belief that speeds decline rapidly once the volume/capacity ratio exceeds about 0.80. Other results of this work include the following:

- A suggestion that free-flow speeds of drivers who have been in queues may be noticeably less than those of drivers who have not, even at considerable distances downstream of the bottleneck (*2*);
- A finding that acceleration back to free-flow speeds was still incomplete at a point about 0.8 km (0.5 mi) downstream from the bottleneck (*3*); and
- Speculation that precipitous drops in free-flow speeds reported by previous researchers may have resulted from lumping together data from different locations in the acceleration zone downstream of the bottleneck, where speed would be expected to increase at nearly constant volumes as one moves downstream (*3,4*).

The series of papers by Hall et al. deals with the overall nature of speed-flow-concentration relationships. Major findings are stated in terms of the flow-occupancy relationship. They are that this relationship varies according to both lane and location (*7*) and that the basic relationship is best described (except for shoulder lanes) as a continuous but not continuously differentiable relationship resembling an inverted V, as opposed to the conventional inverted-U shape, the reverse lambda shape proposed by Koshi, and various discontinuous models (*6,7*) (see Figure 1). This finding is consistent with the idea that there is little or no drop in speed with increasing flow under uncongested conditions. The left branch of the flow-occupancy relationship is interpreted as representing free flow. If this branch is linear (or nearly linear), speeds in the uncongested regime are constant (or nearly constant) as flow increases. In addition, Hall and Gunter (*7*) suggest that flow-occupancy relationships may be shifted upward (that is, higher flow at a given occupancy) in secondary bottlenecks.

Hall et al. are also concerned with the conditions under which transitions between congested and uncongested flow take place. They approach this question by studying time

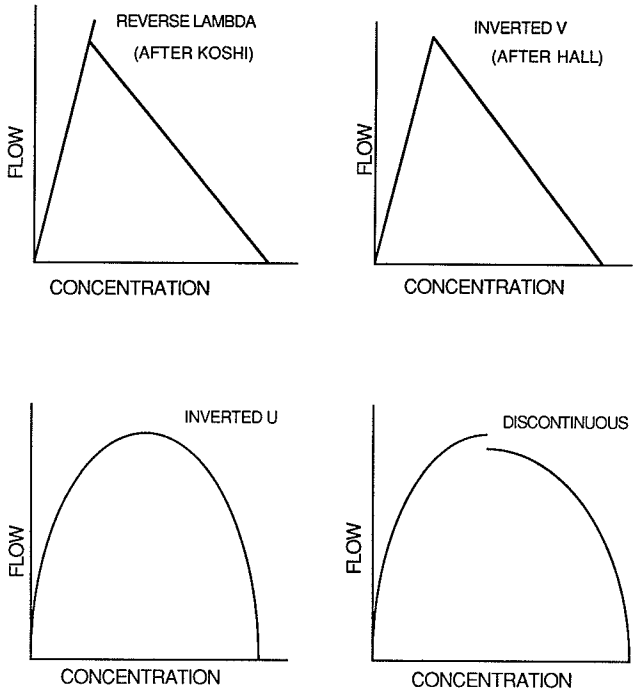


FIGURE 1 Flow-concentration models.

sequences of speed-flow and flow-occupancy states. They report some general patterns (*5,8*) and also use such sequences in the screening and interpretation of data (*5–7*).

RESEARCH OBJECTIVES

The research described in this paper was intended to verify and extend certain aspects of the recent Canadian work. Specific objectives were

- To compare conclusions drawn from the Toronto and Tokyo data with similar data from San Diego,
- To take a more systematic look at the effect of queuing on speed-flow-concentration data, and
- To extend the interpretation of Hall's flow-concentration model by determining its implications as a model of driver behavior.

A major issue considered in the research but not dealt with in this paper is the interpretation of time sequences of speed-flow and flow-concentration states. Interpretation of such data needs to be clearly related to theories of shock wave movement. The discussion of this subject by Hall et al. (*6*) is by no means complete; unfortunately, the matter is too complex to address here.

The first objective of this paper is to verify the recent Canadian findings using data from the San Diego ramp metering system. The San Diego data are quite similar to those used by Hall because they were produced by the metering system's detectors during their routine operation. In this respect they differ from those used by Hurdle, which were produced from time-lapse films. Consequently, the San Diego data set shares the advantages and disadvantages of that used by Hall: the data are abundant, but their exact interpretation is less certain than that of data reduced from a visual record. In general, less screening of data was employed in this study than in the Canadian ones. No effort was made to screen out data resulting from incidents or transitions between congested and uncongested flow. The justifications for this are (a) that incident data should resemble other congested flow data and (b) that allowances can be made for the effects of transitions and the averaging of data from dissimilar flow states without eliminating the data. The major question to be addressed by this part of the work is the extent to which the relationships implied by the Toronto and Tokyo data are confirmed by the San Diego data.

The second objective, to take a slightly different look at the way the data are affected by the relationship between the data-gathering location and the bottleneck, requires a bit of explanation. Although both groups of Canadian researchers are concerned about the effect of queuing on the proper interpretation of speed-flow-concentration data, the only comparison of data gathered both upstream and downstream of the same bottleneck is a discussion in Persaud (*3*) of the process of flow breakdown, which includes data from immediately upstream of the bottleneck as well as from the bottleneck section itself.

A more serious problem arises because Hall's analysis of the effect of queuing on relationships observed upstream of the bottleneck is based on data analyzed on a lane-by-lane basis. At first glance, this appears to be an advantage; it provides more detail and, as it turned out, somewhat different relationships were found in different lanes. However, the whole basis for the concern about queuing is the observation that, once the queue backs up into the upstream section, flow at that point is limited to the capacity of the bottleneck. Thus, maximum flows depend on conditions downstream rather than at the point of observation. This means speed-flow-concentration relationships for near-capacity flows may never be observed; rather, discontinuous data and sudden transitions in speed and concentration can be expected, even if the underlying relationship is continuous (see Hall et al. (*6*), especially Figure 1, for the full argument).

The problem with this reasoning is that it applies only to the facility as a whole, not to individual lanes. In fact, the discontinuous data were found only in the shoulder lane (*6,7*), which suggests that flows up to capacity were occurring in the other lanes (although the average flow across all lanes was less than capacity) during periods of uncongested flow. Consequently, hypotheses about the ways in which queuing and shock wave movement affect speed-flow-concentration data can be verified only if data are averaged across all lanes.

This paper addresses these issues using data collected both upstream and downstream of a bottleneck and data averaged across all lanes. Specifically, the hypothesis to be tested is that, for such data, speed-flow and flow-concentration relationships will appear to be "truncated" in terms of flow (possibly to the point of appearing discontinuous) upstream of the bottleneck but not downstream.

In the process of testing this hypothesis, yet another problem arose: the operation of the bottleneck was more complicated than anticipated. In particular, it does not appear that any one location is always the critical point, so data representing both upstream and downstream conditions may have existed at two of the three locations, and many of the data may actually represent transitional states.

The third objective of this paper is to extend the interpretation of the inverted-V flow-concentration relationship pro-

posed by Hall to determine its implications as a model of driver behavior. Since several models might represent the data equally well, it is interesting to note that this one is consistent with a plausible hypothesis about how drivers determine speeds and spacings.

DATA AND DATA ANALYSIS

The data used in this study were morning-peak (6:00–9:00 a.m.) data produced by loop detectors in the San Diego ramp metering system. Raw data consisted of volumes and occupancies reported at 30-sec intervals for each lane. These had been used in a previous study (*10,11*) and had already been aggregated across all lanes and over 6-min intervals, with the volumes converted to flow rates expressed in vehicles per hour per lane. In addition, average speeds had been calculated from the volume and occupancy data, assuming an average effective vehicle length (including the detector) of 25.75 ft. While there may be some errors in the speed observations due to variations in vehicle length, the speeds appear to be consistent and reasonably accurate.

For purposes of this study, three locations near a major bottleneck on Interstate 8 were selected. Data were collected over a total of 30 days in the months of September and November 1987. Most data represented normal weekday morning commute conditions; however, two holidays were included to provide data related to low-volume free-flow conditions.

Figure 2 is a schematic diagram showing the detector locations and lane configurations. When these locations were selected, it was believed that the section just downstream from the College Avenue onramp, which experiences extremely high flows per lane, is the bottleneck under normal conditions. Further observation and careful examination of the data suggest that no single bottleneck location can be identified, however, and that the detectors upstream of this onramp are often in the zone of acceleration downstream of the queue. At other times, minor queues of short duration form at or downstream from the detectors at Waring Road.

In selecting these locations, the intention was to choose one normally downstream from the bottleneck, one immediately upstream, and one far enough upstream that most queuing would result from traffic backing up from the primary bottleneck. Data were also available from other bottleneck locations in the San Diego area, but this one was used for analysis because of the extremely high volumes just downstream from College Avenue and because there is very little flow on the offramp upstream of the Waring Road detectors. Other bottlenecks for which data were available terminate in freeway branch connectors and involve large drops in flow on the freeway main line upstream of the first set of detectors that are downstream from the bottleneck. In addition, the College Avenue and Waring Road detectors are separated by a comparatively long distance and steep downgrade, which means Waring Road should be beyond the acceleration zone from the College Avenue bottleneck.

As in most recent studies of speed-flow-concentration relationships, the data analysis concentrated on qualitative relationships. A microcomputer with color graphics capability was used to combine and display data. Displays that proved useful in analyzing the data included

- Graphs of speeds and flows versus time of day, for different days at the same location and different locations on the same day;
- Scatter plots of speed versus flow and flow versus occupancy for each location separately and for the three locations combined; and
- Scatter plots of speeds at the same time at two different locations.

In addition, regression analysis was used in the study of speed-flow relationships under free-flow conditions.

Graphs of speeds and flows versus time of day were used to identify linkages between speed and flow conditions at

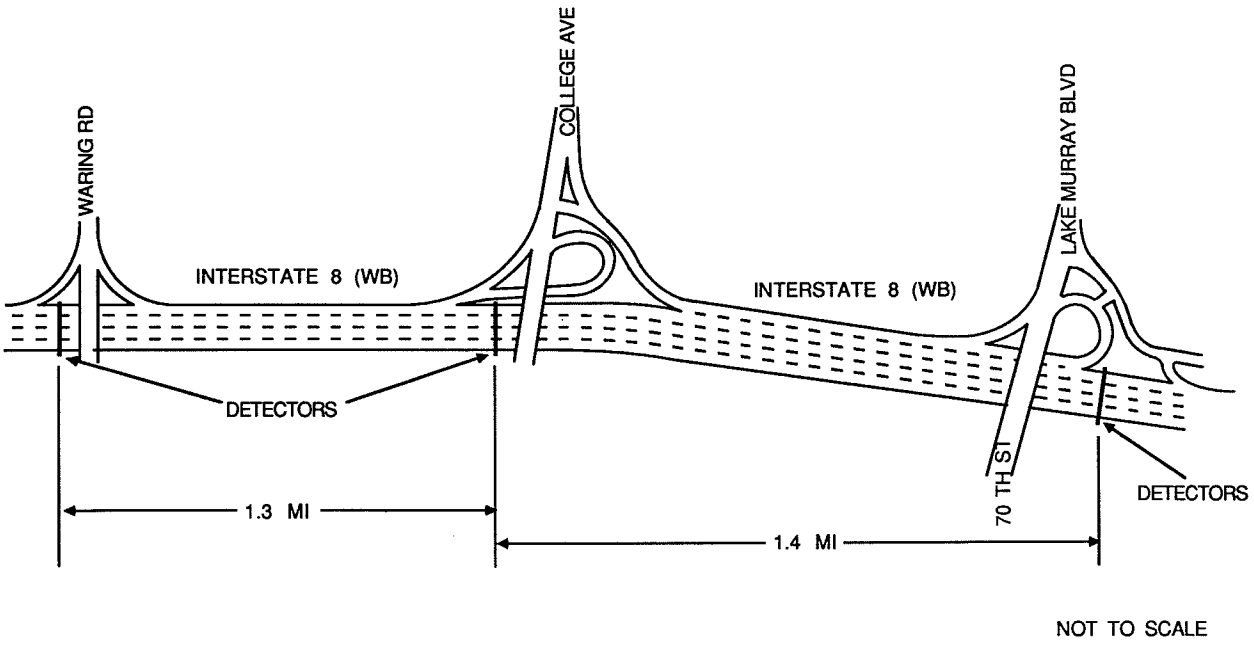


FIGURE 2 Schematic diagram of study site.

different locations and to identify the time of day at which various phenomena occur for purposes of further observation. Scatter plots of speed versus flow and flow versus occupancy were used to investigate these relationships at different locations and under different conditions. Scatter plots of speeds occurring simultaneously at different locations were used to determine whether any correlations between speeds upstream and downstream of the bottleneck might indicate that speed variations downstream of the bottleneck were due to acceleration effects.

VERIFICATION OF PAST WORK

Figure 3 shows scatter plots of speed versus flow at the three locations; similar plots for flow versus occupancy are provided in Figure 4. These plots confirm the major findings of the recent Canadian work:

- There is only a slight drop in speed as flow increases under free-flow conditions, with speeds in excess of 50 mph persisting up to the highest flow levels recorded; and
- The overall relationship between flows and occupancies could well be described as an inverted V, allowing for the effects of queuing and data that might result from averaging over dissimilar flow conditions.

As in Persaud and Hurdle (*4*), there is a slight decline in free-flow speeds as flows increase beyond 1,500 veh/lane/hr. In this study, all data involving flows less than this were gathered on holidays and may not be comparable to the rest of the data. As can be seen in Figure 3, regression analysis confirmed that the slope of the speed-flow relationship varies slightly with location, with the steepest slope occurring at College Avenue. Comparison with the Toronto data indicates that the decline in speed is far more gradual in San Diego, with drops of from 0.3 to 0.5 mph per 100 vph in San Diego as opposed to a rough estimate of 2 to 3 mph per 100 vph for the data in Figures 8 and 9 of Persaud and Hurdle (*4*).

Figure 4 shows the same basic relationships as those discussed by Hall et al. (*6,7*) and Koshi et al. (*9*). The overall flow-occupancy relationship might be described by either the inverted V of Hall or the reverse lambda of Koshi. As noted by Hall, data points are tightly distributed about the left branch of the relationship; somewhat more scatter appears in the right branch. In addition, some points fall "inside" the V or lambda. Time sequences of the data reveal that these often occur during rapid transitions from one branch to the other; hence, they appear to represent averages of data from dissimilar flow conditions. As might be expected, the greatest difficulty in interpreting the data occurs at high volumes in the region where the two branches of the relationship appear to meet.

Other findings by the Canadian researchers concerning the effects of upstream queues and secondary bottlenecks were not confirmed by the San Diego data. Hall and Gunter (*7*) found that flow-occupancy relationships in a "secondary bottleneck" were shifted upward; in other words, higher flows occurred for given occupancies than at other locations. Hall and Gunter's definition of a secondary bottleneck is somewhat unclear. However, the situation appears to have been similar to that at 70th Street and Lake Murray Boulevard in that

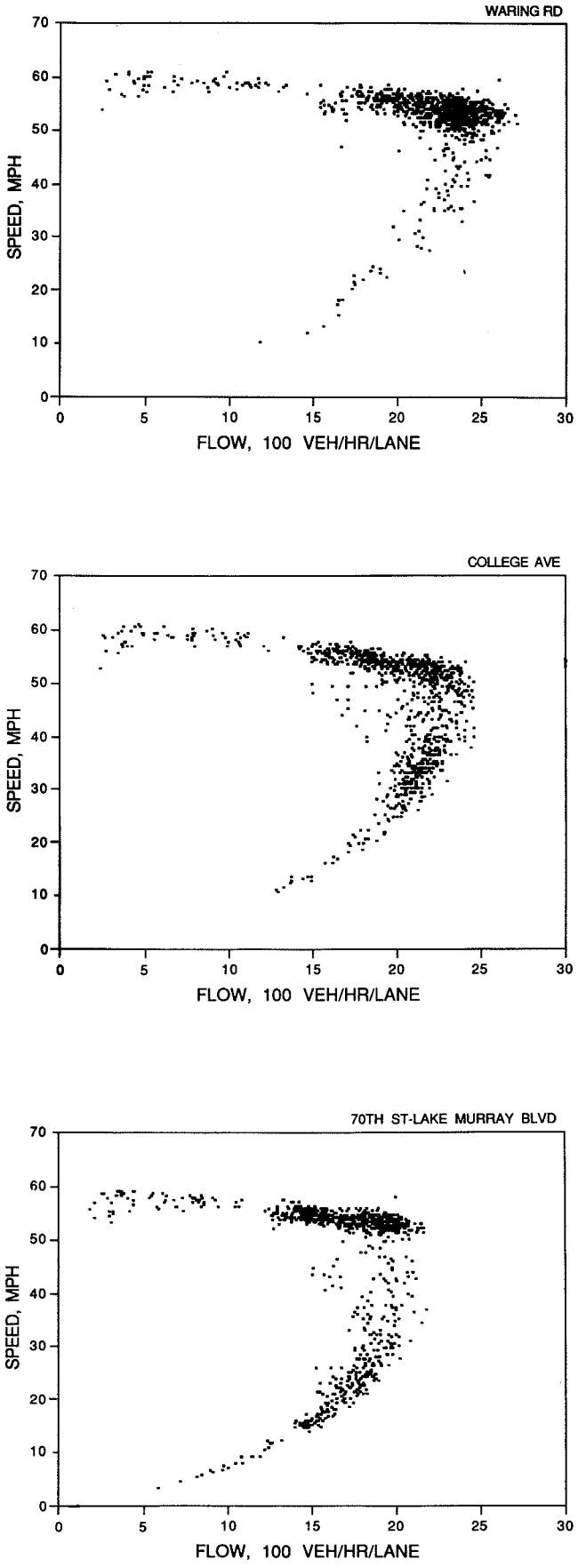


FIGURE 3 Speed-flow scatter plots.

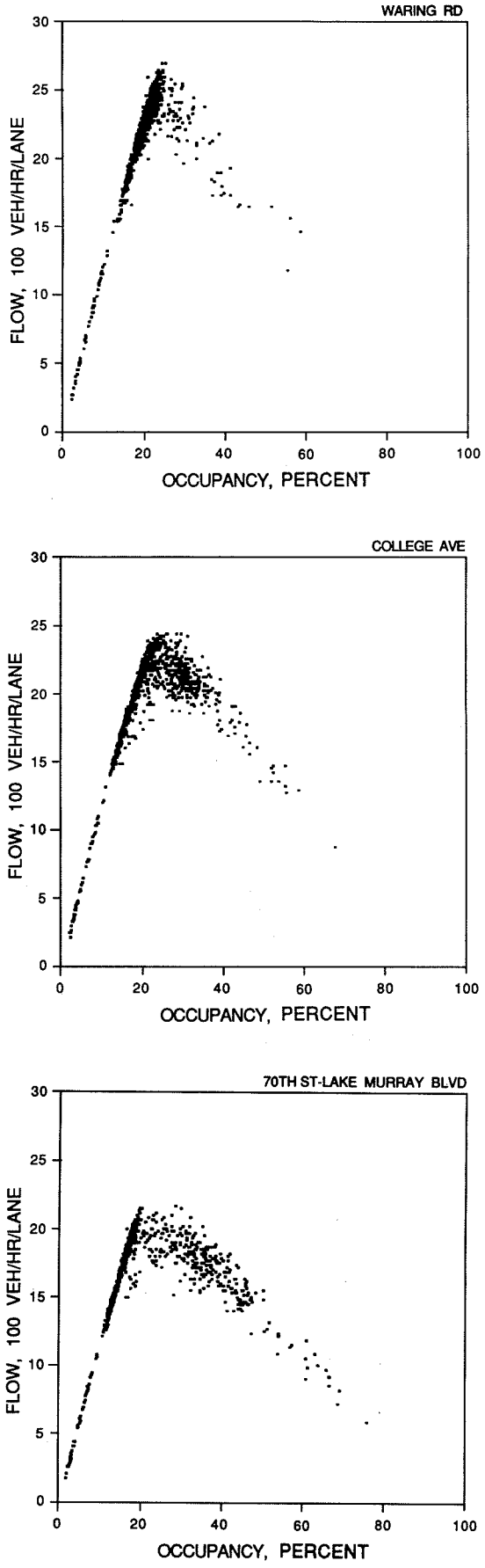


FIGURE 4 Flow-occupancy scatter plots.

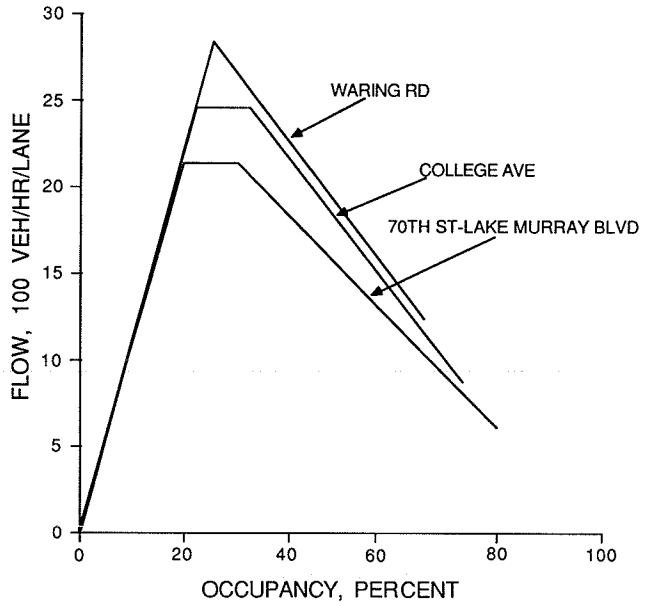


FIGURE 5 Approximate upper bounds of flow-occupancy data.

queues normally backed up into the section from downstream, and heavy onramp flows resulted in smaller mainline flows than occurred in the queue immediately downstream.

If the situations were similar, the San Diego data reveal the opposite result: the flow-occupancy relationship at 70th Street and Lake Murray Boulevard is shifted downward relative to the others. Figure 5 compares the approximate upper bounds of the data at the three San Diego locations. It would appear that, whatever the reasons for the variations in the flow-occupancy relationships at different locations, the presence of "secondary bottlenecks" in the sense of Hall and Gunter is not the cause.

The suggestion by Hurdle and Datta (*2*) that free-flow speeds are reduced when drivers have passed through upstream queues was also not confirmed. Comparisons of speeds taken simultaneously at Waring Road (downstream of the bottleneck) and College Avenue (presumably just upstream) showed that normally there was no relationship between them, and there was no consistent tendency for speeds at Waring Road to drop as soon as the queue formed upstream of College Avenue. When a drop in speed occurred at both locations, it appears most likely that the cause was a minor queue backing upstream from Waring Road or some point downstream. Consequently, these data provide no support for the idea that drivers who have experienced queues upstream alter their free-flow speeds.

EFFECTS OF QUEUING

Figures 3 and 4 show that, when data from all lanes are combined, the speed-flow and flow-occupancy relationships predicted by queuing and shock wave theory do appear: the relationships are progressively more truncated in terms of flow as one moves from locations downstream of the bottleneck to those upstream.

Although the San Diego data appear to confirm expectations about the effects of queuing and shock wave movement

on observed speed-flow-concentration relationships, the functioning of the bottleneck was more complex than had been anticipated. As previously mentioned, at the time these locations were selected, the bottleneck was assumed to be the section just downstream from the College Avenue onramp. Flows in vehicles per hour per lane are maximum in this section, and the lower speeds associated with queuing usually begin at College Avenue and move upstream from there. At the beginning, it was assumed that queuing would most often involve slowing in the right-hand lane due to merges into inadequate gaps at the College Avenue onramp; the major question was whether the entire section downstream of the ramp should be viewed as the bottleneck or only the merge point itself.

Subsequent observation showed that this picture was false in two ways. The first, which may not be important, was that there appeared to be quite a bit of sporadic queuing at the Waring Road detectors, which were assumed to be downstream of the bottleneck. In this case, an issue of interpretation exists: do the brief episodes of low speed actually represent queuing, an acceleration effect, or sporadic occurrences of the precipitous drops in free-flow speed that Persaud and Hurdle (*4*) failed to find? Acceleration effects seem to be ruled out by the lack of correlation of speeds at this location with speeds upstream, as discussed above. The most likely explanation is that they represent queuing. In some cases, this can be substantiated by the fact that the speed drop at Waring Road appeared to propagate upstream. To be sure, however, visual observations need to be correlated with the data. This was not undertaken because the phenomenon is rare and a great deal of observation might have been required. The lesson here is probably that, if "capacity" is a random variable, it will sometimes be exceeded at places other than the normal bottleneck. This is especially likely where flows are extremely high throughout the vicinity of the bottleneck. An alternative explanation is that all these queues are the result of potentially identifiable minor incidents and that such incidents are fairly common.

The more important way in which the functioning of the section violated the initial assumptions was that, during queue discharge, the College Avenue detectors (which are upstream of the onramp junction) appeared most often to be in the zone of acceleration downstream of the queue. It was hard to tell whether the location of the downstream end of the queue stabilized or not; if it did, the most frequent location would appear to be at a point of somewhat restricted sight distance where horizontal and vertical curves coincide just downstream of the College Avenue offramp junction. Meanwhile, occasional dense queues, which tended to discharge off the downstream end (so that the downstream end of the queue moved upstream), were observed in the right-hand lane. These were presumably the result of braking at the merge point, although it was difficult to actually see this happening. While it is conceivable that some such incident initiates the queuing at this location, dense queuing in the shoulder lane occurs only a small percentage of the time after the queue is established.

In terms of data interpretation, much of the College Avenue data do not represent what can be considered unambiguously to be congested flow; rather, they represent sporadic dense queuing in one lane combined with longer periods in which the location is in a zone of transition downstream of the queue. Despite this, data for periods in which this behavior typically occurs appear indistinguishable from the rest of the "congested" flow data.

This circumstance illustrates some of the difficulty encountered in interpreting high-volume speed-flow-concentration data. First, the speed-flow and flow-concentration relationships can be expected to show truncation in terms of flow both upstream and immediately downstream of a bottleneck as long as no traffic enters the freeway at the bottleneck. Downstream of the bottleneck, flow is limited to the capacity of the bottleneck, which is presumably less than that of the point of observation, just as it is when the point of observation is upstream of the bottleneck. On the other hand, data taken in the zone of acceleration should be distinguishable from congested flow data. Particular levels of flow should be associated with lower speeds and higher concentrations in congested flow than in the zone of acceleration. The most likely reason that the two conditions are indistinguishable in practice is that the scatter of the congested flow data obscures the slight tendency of the transitional data to be shifted toward higher speeds and lower occupancies.

BEHAVIORAL IMPLICATIONS OF INVERTED-V MODEL

The final objective of this research was to extend the interpretation of Hall's inverted-V flow-concentration model to determine its implications as a model of driver behavior. The inverted-V model implies that drivers maintain a roughly constant average time gap between their front bumper and the back bumper of the vehicle in front of them, provided their speed is less than some critical value. Once their speed reaches this critical value (which is as fast as they want to go), they cease to be sensitive to vehicle spacing, and speeds remain more or less constant with respect to volume or concentration. The average time gaps may vary with location, geometrics, driver population, and the like, and may be subject to some random variation, but they should be reasonably constant at any location under similar conditions.

This can be stated mathematically as follows. For the sake of simplicity, it is assumed that the central tendency of the right-hand branch of the flow-concentration relationship is linear. Concentration is defined in terms of density rather than occupancy to take advantage of the definition of flow as the product of speed and density. The following also apply:

u = average speed;
u_m = average speed at capacity, assumed to be roughly equal to free-flow speed;
q = average flow;
q_m = maximum flow;
k = average density;
k_j = jam density;
k_m = critical density (i.e., the density at which q_m occurs);
L = average effective vehicle length, which includes the physical length of the vehicle plus any buffer established by the drivers to provide a margin of safety;
h = average time headway between common points on two successive vehicles; and

g = average acceptable time gap between the front of a vehicle and the rear of the buffer space behind the preceding vehicle.

The equation of the line representing the congested flow regime is

$$q = \frac{q_m(k_j - k)}{k_j - k_m} \qquad \text{for } k_m < k < k_j \tag{1}$$

The assumption that drivers adjust speeds to maintain average time gap g results in

$$h = g + L/u \qquad \text{for } 0 < u < u_m \tag{2}$$

which is equivalent to

$$q = 1/g - Lk/g \qquad \text{for } 0 < u < u_m \tag{3}$$

If it is further assumed that the average spatial gap between common points on vehicles is equal to L at jam density, and that the average space and time gaps at maximum flow are equal to $gu_m + L$ and $g + L/u_m$ respectively, then

$$q_m = \frac{1}{g + L/u_m} \tag{4}$$

$$k_m = \frac{1}{gu_m + L} \tag{5}$$

and

$$k_j = 1/L \tag{6}$$

By substituting these relationships into Equation 1, it can be shown that Equation 1 is identical to Equation 3. Consequently, a linear flow-density relationship in the congested flow regime is identical to a model of driver behavior that assumes that, on average, drivers maintain constant time gaps between vehicles until they reach the desired maximum speed.

The branches of the flow-concentration relationship do not have to be exactly linear for the behavioral model to be approximately true. For instance, most recent work shows a gradual decline in free-flow speeds with increasing flow, which implies some nonlinearity in the left branch of the flow-density relationship. The explanation of this in the model of driver behavior is that different drivers desire different maximum speeds, and, because maneuverability is not complete in high density flows, not all of them achieve this speed at the same flow.

Similarly, some nonlinearity in the congested-flow branch would indicate a slight change in the acceptable time gap (g) with increasing speed. Both the Koshi and Hall models represented the right branch of the flow-concentration relationship as slightly convex to the origin. Whether this is supported by their data (as opposed to a linear relationship) is questionable. The strongest support for the nonlinear relationship comes from Figures 3 and 4 of Koshi (*9*); however, it should be noted that the "density" data in these figures were actually transformed occupancy data, in which the relationship between density and occupancy was assumed to be nonlinear. This nonlinearity is apparent when their occupancy and density data are compared [(*9*), Figure 1], but no specific information is given as to how the density data were related to the occupancy data and no explanation for the nonlinearity (in terms of detector operation, for example) is provided. The normal assumption would be that the relationship between density and occupancy is linear; if it is, the nonlinearity in Koshi's flow-density relationships might disappear.

It has been shown that the inverted-V model of the flow-concentration relationship yields a simple, plausible model of driver behavior; however, the reverse lambda model is in some ways a better representation of the actual data. The contradiction between the models can be resolved if it is understood that the inverted V is primarily valuable as a behavioral model, whereas the reverse lambda is a good representation of the actual central tendency of the data. The key is that both the maximum flow rate for any given time period and the density at which it occurs are random variables with considerable variance. If the line representing the free-flow branch is extended to the highest flow levels recorded (in situations in which the relationship is not truncated by queuing), it will naturally overlap the line representing the central tendency of the congested-flow branch.

CONCLUSION

The research described in this paper involved the use of data from the San Diego ramp metering system to verify and extend recent Canadian research related to freeway speed-flow-concentration relationships. Specific findings include the following:

- The major results of the recent Canadian work were verified by the San Diego data. In particular, there appear to be only slight declines in speed with increasing flow under free-flow conditions, and the overall speed-concentration relationship is well described by the inverted-V model proposed by Hall et al.
- Certain other results of the Canadian work related to the effects of queuing on downstream free-flow speeds and the effects of secondary bottlenecks in shifting flow-concentration relationships were not confirmed.
- When data are averaged across all lanes and include locations both downstream and upstream of the bottleneck, observable speed-flow and flow-concentration relationships are consistent with those predicted by queuing and shock wave theory. This proved to be true even though the functioning of the bottleneck was more complicated than had been assumed.
- The inverted-V flow-concentration model implies a model of driver behavior in which speeds and spacings are adjusted to keep the average front-to-back time gaps between vehicles approximately constant until some desired maximum speed is reached.

The research described in this paper needs to be extended in a variety of ways. First, data from other bottlenecks in the San Diego area, as well as other geographical areas, need to be examined to further confirm the main conclusions. Second, a complete theory of the effect of shock wave movement on time sequences of speed-flow and flow-concentration states needs to

be developed and compared with the data. Finally, extensive film or video records need to be taken at detector locations and compared with automatically collected speed-flow-concentration data to better establish their interpretation.

ACKNOWLEDGMENT

This paper is based on data provided by the California Department of Transportation. The data were reduced to their present form as part of a previous study funded jointly by the California Department of Transportation and the U.S. Federal Highway Administration.

REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
2. 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.
3. B. N. Persaud. *Study of a Freeway Bottleneck To Explore Some Unresolved Traffic Flow Issues*. Dissertation. University of Toronto, Toronto, Canada, 1986.
4. 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.
5. 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.
6. F. L. Hall, B. L. Allen, and M. A. Gunter. Empirical Analysis of Freeway Flow-Density Relationships. *Transportation Research*, Vol. 20A, Elmsford, N.Y., 1986, pp. 197–210.
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. M. A. Gunter and F. L. Hall. Transitions in the Speed-Flow Relationship. In *Transportation Research Record 1091*, TRB, National Research Council, Washington D.C., 1986, pp. 18–21.
9. M. Koshi, M. Iwaski, and I. Okhura. Some Findings and an Overview on Vehicular Flow Characteristics. *Proc., 8th International Symposium on Transportation and Traffic Theory, 1981*. University of Toronto Press, Toronto, Ontario, Canada, 1983, pp. 403–426.
10. J. H. Banks. *Performance Measurement for Centrally-Controlled, Traffic Responsive Ramp Metering Systems*. Civil Engineering Report Series 87141. San Diego State University, San Diego, Calif., 1987.
11. J. H. Banks. Performance Measurement for a Metered Freeway System. In *Transportation Research Record 1173*, TRB, National Research Council, Washington, D.C., 1988.

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

DISCUSSION

PETER G. FURTH
Department of Civil Engineering, Northeastern University, Boston, Mass. 02115

Dr. Banks has given us an interesting interpretation of a linear right branch of the flow-density relationship, namely, that in congested flow a driver maintains a constant time gap between the rear of the preceding car and the front of his own car. A parallel interpretation is that the spacing between the front of a car and the rear of the preceding car is proportional to speed. One special case of the model implied by this interpretation is the well-known safety rule that drivers should keep one car length between their car and the car ahead of them for every 10 mph.

If $s = 1/k$ = average spacing between common points on two successive vehicles, then Banks's interpretation leads (by manipulating Equation 2) to

$$s = L + gu \tag{2a}$$

while the proportional distance interpretation, with distance expressed in terms of car lengths, leads to

$$s = L + cLu \tag{4}$$

where c is a proportionality constant. The equivalence between these models is apparent. As long as both models use the same vehicle length L, the models' equivalence implies that

$$c = g/L \tag{5}$$

The old safety rule of one car length per 10 mph specifies $c = 1/(10 \text{ mph})$. With a 20-ft car length, this is equivalent, based on Equation 5, to $g = 1.36$ sec. While Banks does not explicitly estimate g from his data, an examination of Figure 3 suggests maximum flows and corresponding speeds of $q_m = 2{,}000$ to 2,400 veh/hr and $u_m = 52$ mph. The corresponding range for g, using the unnumbered equation following Equation 3 and $L = 20$ ft, is $g = 1.54$ to 1.24 sec, showing very good agreement with the old safety rule.

AUTHOR'S CLOSURE

I want to thank Furth for pointing out that the assumption that the right branch of the flow-density relationship is linear is closely related to the well-known safety rule that drivers should maintain a one-car-length separation for every 10 mph. The close numerical agreement between the safety rule and the data is especially surprising when one considers that the safety rule is stated in very round numbers. It should be remembered, of course, that the safety rule was intended to prescribe minimum separations, whereas the flow-density relationship is based on average separations. In fact, many drivers are following much more closely than the average separation, so it cannot be said that the driver behavior reflected by the data is "safe" in terms of the rule.

Operation of Major Freeway Weaving Sections: Recent Empirical Evidence

MICHAEL CASSIDY, ALEX SKABARDONIS, AND ADOLF D. MAY

Weaving areas are critical elements in the operation of the freeway network, involving complex vehicle interactions. Many of the existing procedures for the design and analysis of weaving sections are based on data collected in the late 1960s and mid-1970s and may not reflect current driver/vehicle characteristics, especially in California. This paper describes what the analysis of more recently collected data has revealed. As part of a 2-year project to develop improved weaving analysis procedures, a large amount of data were collected using a video camera at eight major freeway weaving locations throughout California. Information on geometrics, volumes, and speeds of weaving and nonweaving vehicles were then extracted from the tapes, thoroughly checked, and verified with field observations. Six existing methods for the design and analysis of freeway weaving sections were then applied to all data sets. Results indicate that significant discrepancies exist between the predicted and measured average speeds of weaving and nonweaving vehicles. Several statistical analyses were performed using regression analysis and classification and regression trees, to identify basic relationships between weaving section design and traffic characteristics. Regression techniques were also employed in an effort to improve existing analysis and design procedures and to develop new performance prediction models. One important finding is that speed was insensitive to changes in geometric and traffic factors over the range of values in the data set. Overall, analyses suggest that average travel speed may not be an ideal measure of effectiveness. More research needs to be carried out to develop a more accurate procedure for designing and analyzing freeway weaving sections. Such procedures could consider measures of effectiveness other than speed. This paper also discusses the direction of possible future research.

In November 1987, research on freeway weaving sections was undertaken at the Institute of Transportation Studies (ITS), University of California at Berkeley. The research project, entitled "Evaluation of Existing Methods for the Design and Analysis of Freeway Weaving Sections," was sponsored by the California Department of Transportation (Caltrans). The work has focused on major freeway weaving sections, such as those at or near freeway interchanges. Research was concurrently performed on ramp-weave freeway sections (*1*).

Preliminary findings of the research on major weaving sections are outlined in Skabardonis et al (*2*), which describes initial research tasks, including data collection, and findings from the application of existing weaving analysis methods to the collected data.

This paper describes efforts subsequent to Skabardonis et al. to (a) identify factors significantly influencing traffic operations on major freeway weaving sections, (b) assess the predictability of measures of effectiveness (i.e., speed and density), and (c) develop improved procedures for predicting weaving section performance.

Institute of Transportation Studies, University of California, Berkeley, Calif. 94720.

The purpose of this research project is to develop improved procedures for the design and analysis of major freeway weaving sections—focusing on weaving locations in the state of California. The work has had the following specific objectives: to evaluate all known existing methods used for the design and analysis of weaving sections, and to develop a new or modified methodology as needed.

BACKGROUND

The 1985 Highway Capacity Manual (HCM) (*3*) 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." Typically, a weaving section is formed by a merge area followed closely by a diverge area. Four types of traffic movements generally occur on a freeway weaving section: freeway to freeway (a nonweaving traffic stream), freeway to off-ramp (a weaving traffic stream), on-ramp to freeway (a weaving traffic stream), and on-ramp to off-ramp (a nonweaving traffic stream).

The intense lane-changing maneuvers that go on in weaving areas often present sharp operational problems.

The 1985 HCM has designated several classifications of weaving sections: a simple weaving area is formed by a single merge followed by a single diverge; a multiple weaving area is formed by one merge followed by two diverges or by two merges followed by a single diverge; a ramp-weave section is formed by a one-lane on-ramp followed closely by a one-lane off-ramp where the two are joined by a continuous auxiliary lane; and a major weaving section is formed when at least three entry and exit legs have two or more lanes.

The 1985 HCM also defines three weaving area configuration types—A, B, and C. These classifications are based on the minimum number of lane changes required by weaving vehicles as they travel through the section:

- Type A configurations require that each weaving vehicle perform one lane change in order to execute their desired movements. Ramp-weave freeway sections are typically of this configuration.

- Type B weaving areas require vehicles in one weaving traffic stream to execute one lane change, while vehicles in the other weaving traffic stream perform desired movements without changing lanes.
- Type C weaving sections require vehicles in one weaving traffic stream to perform two or more lane changes, while

vehicles in the other weaving traffic stream can perform desired maneuvers without changing lanes.

The research described in this paper focuses on simple, major freeway weaving sections but gives consideration to all three configuration types.

Initial Research Tasks

This section summarizes initial tasks performed in the early stages of the research project. A more complete description of these tasks is in Cassidy et al. (*4*) and Skabardonis et al. (*2*).

Literature Review

At the beginning of this study, a detailed literature search was performed using the ITS library and TRISNET computerized searches. Over 100 documents concerning weaving were identified. These published materials were classified into seven major categories: methods for design/analysis of freeway weaving sections, simulation models, theoretical studies, data collection, related capacity analysis, nonfreeway weaving sections, and California studies. Several researchers and practitioners were contacted to find out which methods are commonly employed and what application experience has been.

Seven existing procedures for the design and analysis of freeway weaving sections were identified through the literature search. In chronological order of development, these methods are as follows:

1. California Method (Level D Method) (*5*),
2. HCM-65 (*6*),
3. Leisch (*7,8,9*),
4. PINY (*10*),
5. JHK (*11*),
6. HCM-85 (*3*), and
7. Fazio (*12*).

The Level D Method (*5*) is used to predict the spatial distribution of vehicles operating on ramp-weave freeway sections under heavy volume conditions. The HCM-65 technique (*6*) predicts a "Quality of Flow" (and corresponding approximate travel speeds) for vehicles traveling on a subject weaving section. The remaining five procedures predict average vehicle travel speeds and typically establish a level of service designation based on those predicted speeds. With the exception of the Level D Method, all existing procedures can be applied to major freeway weaves as well as to ramp-weaves. All existing procedures can be used to evaluate proposed geometric designs for weaving sections (*2*).

Data Collection

As part of this research project, data from major freeway weaving sections in California were collected and analyzed. These data were used to evaluate the reliability of existing procedures and to calibrate new prediction models as needed.

Eight test sites were chosen for data collection. These sites were all major weaving sections, representing a wide variation of design characteristics and section configurations. Table 1 of this report describes the site locations and geometric configurations.

Following the selection of the study locations, data were collected using video recording. Videotaping traffic provides for a permanent record of the data, which can later be analyzed at various levels of detail or rechecked as necessary.

A Panasonic 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 sections so that the operation of the entire weaving section could be videotaped. Several tests proved that accurate measurements could be made using the video camera for distances up to 3,000 ft.

Six hours of video recording were made on each site to obtain a range of traffic conditions. The tapes were analyzed to obtain input data for the procedures—volumes and traffic composition—and performance measures—average speeds of weaving and nonweaving vehicles.

Information on geometrics was obtained from field measurements and checked through aerial photos and Caltrans photologs.

The data were extensively checked, and a number of floating car runs were made to verify the speeds at the weaving sections. Statistical tests were performed to check sample size requirements for speed measurements. A total of 143 data points were obtained, each representing 15-min observations at the study locations. Approximately 11 hr of additional data were excluded from further analysis because of congestion, incidents, and inclement weather occurring during the videotaping.

Evaluation of Methods

Six of the seven existing methods were applied to all test site data. The California procedure (LOS D method) was not applied because it was developed specifically for the analysis of ramp-weave freeway sections.

Table 2 presents a comparison of measured and predicted performance measures (i.e., average weaving and nonweaving speeds) for a sample of data points. Each data point listed represents 60-min observations collected at the eight test sites. As the table shows, the methods typically predict operating speeds that are lower than actual speeds. The differences between predicted and measured speeds range from −17 percent to + 64 percent for weaving traffic, and from −16 percent to +44 percent for nonweaving traffic.

Table 2 also lists mean values of the differences between field-measured and predicted average travel speeds; these average differences were computed using absolute values of the differences between measured and predicted speeds. The mean squared errors are also listed on this table. Figure 1 illustrates the differences between field-measured and average weaving speeds predicted using the 1985 HCM weaving model. Predicted average weaving speeds equal to field-measured values would be positioned directly on the diagonal line bisecting Figure 1, but these weaving speeds are all lower than their corresponding field-measured values. [See Skabardonis et al.

TABLE 1 WEAVING SECTION TEST SITES

Caltrans District	Location	Geometric Characteristics L	N	N_b	N_r	N_f	Comments
4	WB 92 Ralston onramp and 280 I/C	1400	3	2	1	2	Lane Balance Type C
4	SB 280 at 88 I/C and and Bascom offramp	1347	5	3	2	1	Lane Balance Type C
7	SB 101 at 110 I/C and Broadway offramp	792	5	3	2	2	Lane Imbalance Type A
7	NB 101 at Los Angeles onramp and 110 I/C	787	5	4	1	2	Lane Imbalance Type C
7	EB 10 605 I/C and Frazier St. offramp	1437	6	4	2	1	Lane Imbalance Type B
7	WB 10 at Garvey St. onramp and 605 I/C	1690	5	4	1	2	Lane Balance Type B
8	WB 10 15 I/C and Etiwanda onramp	1989	5	4	1	2	Lane Balance Type B
11	NB 805 at University onramp and El Cajon offramp	1371	5	4	1	2	Lane Balance Type B

Definitions:

L: length of weaving section (ft)
N: number of lanes in weaving section
Nb: number of freeway lanes approaching the weaving area
Nr: number of lanes on entrance ramp
Nf: number of lanes on exit ramp

TABLE 2 MEASURED VERSUS PREDICTED SPEEDS

CALTRANS DISTRICT	TEST SITE	FIELD MEAS: Sw	FIELD MEAS: Snw	HCM-85: Sw	HCM-85: Snw	LEISCH: Sw	LEISCH: Snw	JHK: Sw	JHK: Snw	FAZIO: Sw	FAZIO: Snw	PINY: Sw	PINY: Snw	HCM-65: Sw
4	WB92 AM	51	55	*	*	41	59	42	38	55	31	50	50	> 40
4	WB92 NOON	53	56	29	41	44	65	48	50	60	45	55	55	> 40
4	SB280 AM	67	67	33	47	41	54	51	56	44	48	50	50	> 40
4	SB280 NOON	62	63	34	48	42	55	51	57	42	49	50	50	> 40
4	SB280 PM	60	62	31	43	36	39	49	54	34	41	*	*	40-45
7	SB101 AM	50	55	*	*	25	48	37	42	31	41	50	50	40-45
7	SB101 NOON	55	59	*	*	20	46	36	42	28	40	*	*	40-45
7	NB101 AM	44	52	*	*	24	35	34	39	38	33	*	*	30-35
7	NB101 PM	56	62	29	42	25	59	37	42	39	42	44	46	40-45
7	EB10 AM	48	56	36	50	34	59	53	58	47	51	56	56	> 40
7	EB10 NOON	54	59	34	48	32	54	52	57	45	48	56	56	> 40
7	EB10 PM	41	42	*	*	19	34	48	52	32	38	*	*	30-35
7	WB10 AM	56	59	40	40	38	40	52	56	53	47	*	*	40-45
7	WB10 NOON	63	64	43	45	42	50	54	58	56	51	45	47	> 40
7	WB10 PM	58	62	42	43	40	47	53	58	53	50	42	44	> 40
8	WB10 NOON	58	61	48	54	53	56	58	61	58	60	55	55	> 40
8	WB10 PM	64	66	47	52	50	54	57	60	58	57	54	54	> 40
11	NB805 AM	56	63	42	42	37	40	50	56	44	51	*	*	> 40
11	NB805 NOON	56	64	41	41	38	41	50	55	42	51	*	*	> 40
11	NB805 PM	52	59	41	41	38	41	50	56	45	51	*	*	> 40
avg. difference+				19	16	19	12	8	8	11	13	8	9	
mean sqrd error				432	293	420	190	96	96	188	205	112	131	

* METHOD COULD NOT BE APPLIED BECAUSE TRAFFIC AND/OR GEOMETRIC CHARACTERISTICS EXCEED LIMITS OF THE MODEL

\+ MEAN DIFFERENCE BETWEEN FIELD-MEASURED AND PREDICTED SPEEDS FOR EACH SAMPLE DATA POINT (ABSOLUTE VALUES USED TO REFLECT EACH DIFFERENCE)

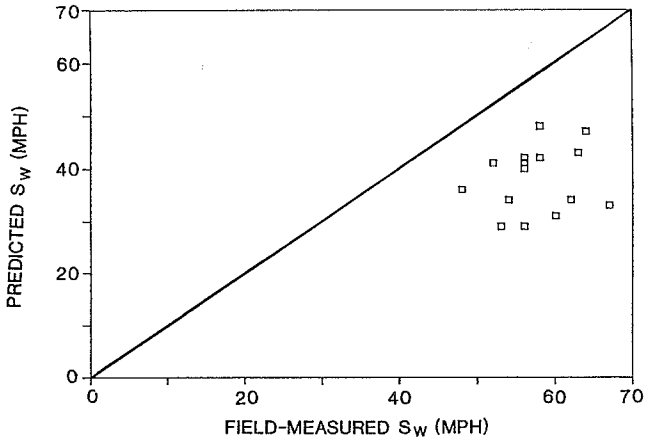


FIGURE 1 Measured versus HCM predicted weaving speeds.

(*2*) for further discussion of the existing methods' predictive value.] Overall, the existing analysis and design procedures do not appear to have strong predictive ability. Based on these findings, it was decided that new or modified procedures should be developed.

The following two sections of this paper describe efforts to calibrate stronger predictive models.

FUNDAMENTAL RELATIONSHIPS

Before attempting to calibrate predictive models, speed, flow, and density relationships were examined to lend a better understanding of operational phenomena occurring at freeway weaving sections.

Speed-Flow Relationships

Relationships between speed and volume-capacity (*v/c*) ratios were first studied. A value of 2,000 passenger cars/hr/lane (pcphpl) was used for capacity. This value represents only a rough estimate since the actual capacity of weaving sections is uncertain. However, the reasonableness of a value such as 2,000 is not important. A capacity value was used here simply to normalize volumes operating on the weaving section. The number of lanes in the weaving area, rather than an estimated capacity, could just as easily have been used.

Figure 2 illustrates speed versus *v/c* scatter plots constructed using 5-min observation data. Speeds used to construct these scatter plots were average weaving speed and average nonweaving speed. The scatter plots were developed using aggregated data from all eight test sites.

The speed versus flow diagram, constructed using nonweaving speed, does to some degree resemble a conventional speed-flow curve. However, speed appears to be insensitive to flow up to *v/c* values of about 0.8. This nondifferentiable behavior between speed and flow agrees with findings by Persaud and Hurdle (*13*) and others concerning "straight-pipe" freeway sections. Unpublished research on multilane rural highways also indicates that speed varies only slightly with increasing traffic volumes under low and moderate flow conditions.

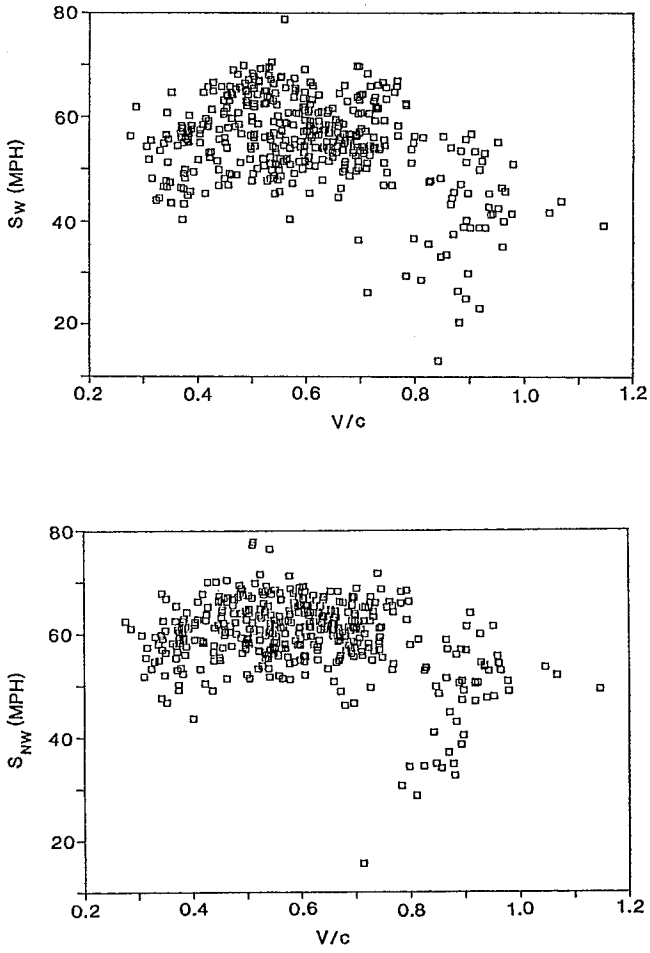


FIGURE 2 Speed versus *v/c* scatter plot.

Figure 2 illustrates the high degree of scatter among the data, which is particularly prevalent in the weaving speed versus *v/c* scatter plot. Separately grouping data by individual location or by configuration type (as defined in the 1985 HCM) did little to improve the high variance between speed and flow.

The data collected at the California test sites therefore do not indicate that obvious relationships exist between speed and flow. Indeed, the apparent insensitivity of speed under low and moderate flow conditions and the high degree of scatter among the data initially indicate that speed may not be an ideal measure of performance for weaving sections.

Density-Flow Relationships

Relationships between density and *v/c* were also examined. Average density/lane, within each weaving section, was computed by dividing total volume by a weighted average of weaving and nonweaving average speeds. Figure 3 illustrates the density-flow relationships for all eight test sites using 5-min observation data. The scatter between density-versus-flow data is significantly tighter than the speed-flow data points. This is because volume is contained in both the *x*- and *y*-axes of the density-flow scatter plot. Nonetheless, relationships between density and flow do appear to be desirable in that density is

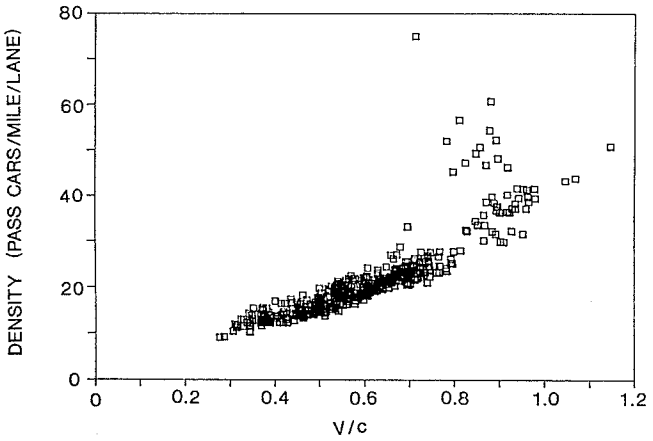


FIGURE 3 Density versus *v/c* scatter plot.

sensitive to flow, and scatter is great only under heavy flow conditions.

EMPIRICAL MODELS TO PREDICT PERFORMANCE

Analyses were performed to assess the predictability of speed and density, and to better understand factors affecting performance at weaving sections. Two types of analyses were done on the California test site data, regression analysis and classification and regression trees (CART) analysis.

Regression Analysis

Consideration was first given to recalibrating existing prediction models. For these recalibrations, the structures of the existing models were left unchanged. Regression coefficients and exponents were modified based on the data collected in this research. The models chosen to be recalibrated were the JHK model and the 1985 HCM model.

These models were selected because they were recently developed speed-based models originally calibrated using nonlinear regression techniques.

The two models differ in that the 1985 HCM model considers geometric configuration, while the JHK model does not. Also, the 1985 HCM model considers the type of operation (i.e., constrained or unconstrained) occurring within the weaving section. In constrained operation, nonweaving vehicles operate at significantly higher speeds than do weaving vehicles. In unconstrained operation, the average speeds of weaving and nonweaving vehicles typically differ by less than 5 mph.

The JHK Model

The basic structure of the JHK models for predicting average weaving and nonweaving speeds is as follows:

$$S_W = 15 + 50/[1 + (1 + V_W/L)^{B1}(V/N^{B2}/A_W] \quad (1)$$
$$S_{NW} = 15 + 50/[1 + (1 + V_W/L)^{B3}(V/N)^{B4}/A_{NW}]$$

where

$B1$, $B2$, $B3$, and $B4$ = regression exponents,
A_W and A_{NW} = regression coefficients,

and all other variables are defined in the glossary.

New exponents and coefficients were developed using the collected data, representing 5-min observations. The original JHK models were calibrated with an upper limit for weaving and nonweaving speeds of 65 mph. Thus, in recalibrating the models, data points having speeds in excess of 65 mph had to be excluded from the calibration data set.

The resulting recalibrated models were quite poor: for the weaving speed prediction model, the r^2 was 0.09 and only 0.06 for the nonweaving speed prediction model. The values for the recalibrated exponents and coefficients do not resemble those of the original JHK model. Table 3 of this report lists the original and recalibrated values for the coefficients and exponents.

The rather poor results would suggest that the models' structure does not fit the data collected at the eight California test sites.

The 1985 HCM Model

The structure of the 1985 HCM model is as follows:

$$S_W \text{ or } S_{NW} = 15 + 50/[1 + a(1 + VR)^b(V/N)^c/L^d] \quad (2)$$

where

a = regression coefficient,
b, c, and d = regression exponents,

and all other terms are defined in the glossary.

Recalibrating the HCM models was not as straightforward as recalibrating the JHK equations. The HCM equations use a greater variety of coefficients and exponents for weaving and nonweaving speeds than the JHK models.

Predicting travel speeds using the HCM equations is an iterative process. The procedure was first performed assuming unconstrained operation, and the resulting predicted speeds were then used to predict the type of operation (constrained or unconstrained) occurring on the section. If constrained operation was predicted, the equations were recalculated using constrained coefficients and exponents.

Since field-measured speeds had been determined for the test sites, these measured speeds were used to determine operation (*3*). Data were then grouped according to configuration and operation type.

Limitations in this project's database prohibited developing new coefficients and exponents for all configurations and types

TABLE 3 REGRESSION COEFFICIENTS AND EXPONENTS (JHK MODEL)

Coefficient/ Exponent	Original JHK Model	Re-Calibrated Model
Aw	7500	9.44
Anw	9000	17.32
B1	1.33	1.21
B2	1.1	-0.05
B3	1.1	0.89
B4	1.1	0.001

of operation. The 1985 HCM equations were recalibrated only for Type B weaving sections operating under unconstrained conditions and for Type C sections operating under constrained conditions. As with the JHK models, the 1985 HCM models required the exclusion of all calibration data having measured speeds greater than 65 mph.

Results from these recalibrations were not promising. For Type B sections, r^2 values were 0.25 and 0.13 for unconstrained weaving and nonweaving speed predictions, respectively. For Type C sections, r^2 values were 0.44 and 0.37 for constrained weaving and nonweaving speed models, respectively.

Recalibrated constants were not usually of the same magnitude as those of the 1985 HCM models. Table 4 lists constants for the original and for the recalibrated HCM models. Referring to Table 4, the recalibrated regression coefficient, a, approaches zero for weaving and nonweaving speed prediction models. Thus, the recalibrated HCM models invariably predict these speeds to be 65 mph. Such prediction models are not satisfactory.

The structure of the HCM models was derived from the JHK models; neither seems to fit the data gathered from the test sites.

New Speed Prediction Models

In an effort to assess the predictability of weaving and nonweaving speeds, a variety of linear regression analyses were performed using test site data. Data representing 15-min observations were used initially (15-min rates of flow were used simply as a matter of convention).

In calibrating the models, the data were aggregated in a variety of ways. Virtually all variables listed in the glossary were considered in developing proposed models. The following section highlights some of the more significant findings resulting from efforts to calibrate improved prediction methods.

Aggregating All Eight Sites

Models were first calibrated using all the data from the eight sites aggregated. The resulting model for predicting average weaving speed is as follows:

$$S_W = 47.38 - 0.01(V) + 10.38(N) - 13.46(N_b) + 0.01(V_1) + 0.01(L) + 0.01(V_2) \quad (3)$$

The r^2 value is 0.45.

The model for predicting nonweaving speed calibrated by aggregating all eight test sites is as follows:

$$S_{NW} = 97.22 - 64.13(v/c) + 0.03\ (V_{W2}) - 40.09(WR) - 6.41(N_b) + 0.01(V_1) - 0.01(V_3) \quad (4)$$

The r^2 value is 0.44.

The r^2 values for these equations were better than those developed from recalibrating the two existing methods, but the coefficients of determination are still rather low.

Aggregating by Configuration Type In calibrating Equations 3 and 4, efforts were made to account for weaving section configuration type; however, categorical and dummy variables proved to be statistically insignificant. Because it was thought that unspecified geometric factors might have been affecting operations, models were calibrated according to configuration type, yet aggregating data in this manner did not result in improved prediction models.

Aggregating by Number of Lanes Improved results occurred when models were calibrated by using the data from test sites with five lanes (i.e., N = 5). The two sites that did not have five lanes were excluded from the data set.

The resulting model for predicting average weaving speed is as follows:

$$S_W = 26.76 + 0.01(L) - 33.98(WR) + 67.38(VR) - 0.01(V_{W1}) + 0.002\ (V_N) + 3.58(C) + 0.005(V_3) \quad (5)$$

where

C = categorical variable representing configuration,
1.0 = configuration Type A,
2.0 = configuration Type B, and
3.0 = configuration Type C.

The r^2 value for this model is 0.82.

The manner in which the effects of geometric configuration are accounted for in the above model is somewhat unusual. Using dummy variables would intuitively seem more appropriate; however, the linear form of the categorical variable proved to be a better model.

The model for predicting average nonweaving speed where $N = 5$ is as follows:

TABLE 4 REGRESSION COEFFICIENTS AND EXPONENTS (1985 HCM MODEL)

	HCM MODEL				RE-CALIBRATED MODEL			
TYPE B UNCONSTRAINED:	a	b	c	d	a	b	c	d
WEAVING	0.1	1.2	0.77	0.5	2E-10	-3.65	2.57	0.41
NONWEAVING	0.2	2	1.42	0.95	3E-24	-7.7	2.93	4.42
TYPE C UNCONSTRAINED:								
WEAVING	0.1	1.8	0.8	0.5	9E-11	65.26	1.74	2.16
NONWEAVING	0.07	1.8	1.1	0.5	1E-14	42.12	3.72	4.05

$$S_{NW} = 15.70 + 0.02(V_1) + 120.91(VR) - 0.01(V_3) + 1.84C + 0.01\ (L) - 0.01(V_{UP}) \quad (6)$$

The r^2 value for this model is 0.58.

Aggregating by Individual Site In an effort to account for site-specific factors significantly influencing the operation of each test site, regression models were calibrated for each individual site. However, no factor or group of factors significantly influenced all eight sites. Each model was unique, leading researchers to conclude that developing a model to account for all geometric and traffic factors will be difficult.

Cross-Validation Analysis All previous regression analyses were performed using resubstitution. In an attempt to obtain more reliable prediction models, a cross-validation regression with two partitions was performed.

In performing this analysis, data from the test sites were divided into two groups. Alternating data points from each of the sites were assigned to each of the two data sets. Only one of these data sets was used to calibrate weaving and nonweaving speed prediction models. The unused data set was then used for validation. Residuals were computed for each data point, and the absolute values of all residuals were totaled. The process was then alternated so that the data set previously used for validation was used to calibrate new prediction models. Again, the unused data set was used for validation, and the absolute values of all residuals were totaled.

A number of models were constructed using this method. The models ultimately selected were those yielding the smallest sum of residuals.

The resulting model for predicting average weaving speed is as follows:

$$S_W = 116 - .004(V_3)J - 125.67(VR) - 5.82(N_b) + 3.26C - 22.49(WR) \quad (7)$$

The r^2 for this model is 0.52.

The resulting model for predicting average nonweaving speed is as follows:

$$S_{NW} = 80.43 - 0.01(V) - 8.02(N_b) + 6.10(N) - 52.57(VR) + 0.01(V_{UP}) \quad (8)$$

The r^2 for this model is 0.42.

Cross-validation typically yields more reliable models than does resubstitution regression. For this research, models yielded by cross-validation analysis appear more rational than do the previously calibrated models. None of the speed-prediction models developed in this project using linear regression is ideal, however.

Density Prediction

Efforts were also made to predict average density/lane using regression techniques. Figure 3 shows that the density-versus-flow diagram behaves quite predictably. Only where flow conditions apparently approach the congested regime does high scatter among data points occur.

Prediction models were calibrated using data representing 15-min observations. A linear model was first constructed. The resulting equation is as follows:

$$d = -3.71 + 40.90\ (v/c) \quad (9)$$

The r^2 value for this model is 0.88.

Referring to Figure 3, the density-flow scatter plot forms a somewhat exponential relationship. Thus, a nonlinear model was calibrated, and the resulting model is as follows:

$$d = 35.35\ (v/c)^{1.06} \quad (10)$$

The r^2 value for this model is 0.89.

Equations 9 and 10 should be used only to predict densities in uncongested flow regions. Outlying points occurring at v/c values greater than 0.8 can best be predicted by a second nonlinear model. In this way, the density-flow curve would resemble the conventional parabolic model.

Sufficient data points have not yet been collected to calibrate a reliable density model for flow conditions in the congested region. Future efforts in this research project will involve obtaining additional data for traffic flow operating under heavy flow or near congested conditions. This data may be obtained from field studies and/or simulation.

Equations 9 and 10 were developed using ordinary least-squares regression. A weighted least-squares regression might better reflect density-flow relationships for v/c values between 0.8 and 1.0. Had weighted least-squares regression been used, regression coefficients for these equations would have been slightly larger. Nonetheless, they do represent acceptable approximations.

Analyzing 5-Minute Observations Attempts were made to develop speed and density prediction models using 5-min observation data. The obvious advantages of using 5-min flow rates are the increased number of data points and the wider range of traffic flow volumes. Statistical tests indicated that contiguous 5-min data points were independent from one another. However, efforts to calibrate models using 5-min observations did not yield improvements over 15-min models. It would seem that the higher variations occurring in the 5-min flow rates resulted in prediction difficulties.

Classification and Regression Tree Analysis

In an effort to gain deeper insights into factors influencing the operation of freeway weaving sections, the eight test sites were evaluated using CART analysis. CART is a technique and computer program developed by Breiman et al. (*14*). The methodology classifies and groups entities based on a set of measurements or characteristics by tree methodology.

The concept is to partition response variables (e.g., average weaving and nonweaving speeds) by a sequence of binary splits into terminal nodes (see Figure 4). The tree structure output provides easily understood and interpreted information regarding the main factors and interaction of factors that are significant in speed prediction.

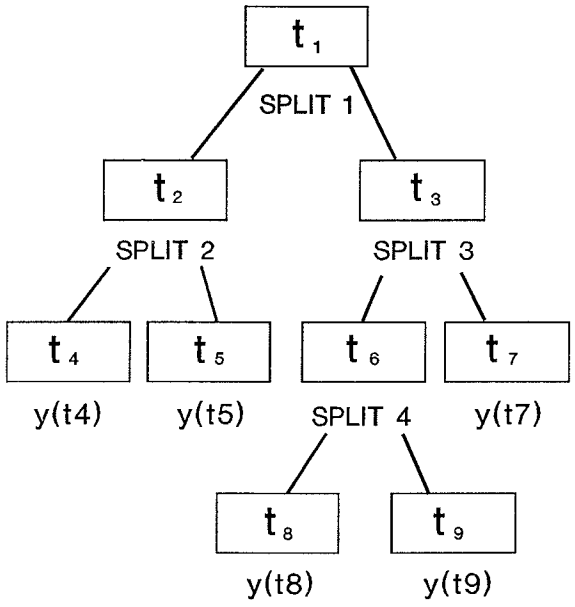


FIGURE 4 Sample CART.

Using the California test site data to develop predictive models, three issues for determining the tree predictor needed to be addressed:

1. **A way of selecting a split (or factor) at every intermediate node.** A regression tree is formed by iteratively splitting nodes. Splits are selected at every intermediate node in such a way as to maximize the decrease in resubstitution error measure. The best split of a node is the split that minimizes the weighted variance;
2. **A rule for determining when a node is terminal.** This decision is related to issues of tree pruning. It can be influenced by user-specified options in the CART program; and
3. **A rule for assigning a value $y(t)$ to every terminal node.** This is simply the average of responses of the terminal group t.

The CART program is written in standard FORTRAN and can be run on most computers in both interactive and batch modes. The program offers three options for estimation method, namely, resubstitution, test sample, and cross validation.

CART analyses were performed on data aggregated in several ways. Where sufficient numbers of data points existed, the cross-validation estimation method with two partitions was the preferred method as it tends to provide the most reliable trees. Where data were aggregated in such a way that an insufficient number of cases were available, resubstitution was used.

Speed Prediction

Attempts were first made to directly predict weaving and nonweaving speed using the CART program. Essentially all variables defined in the glossary were incorporated into the CART learning sample (i.e., input data set). Fifteen-min flow rate data from the California weaving sections were aggregated several ways.

Aggregating All Eight Sites Weaving and nonweaving speed prediction CART models were first developed by aggregating the data from all eight California test sites. The resulting CART for weaving speed prediction is illustrated in Figure 5.

Numbers in each node represent the sample size, the average value of speed, and the standard deviation (reading from left to right). In the first node, the learning sample consisted of 143 data points, with an average weaving speed of 56 mph and a standard deviation of 6.6. The first split is based on volume ratio, *VR*. Data points having *VR* values less than 0.36 are split to the left intermediate node, while data points with *VR* values greater than 0.36 are split to the right.

For the most part, the binary tree illustrated appears to follow logical splitting. There is, however, an exception. Two data points have a lower total volume which leads to lower operating speeds. This causal relationship does not seem logical. This binary tree was constructed without incorporating site-specific categorical variables. When such categorical variables were incorporated, they tended to significantly influence the splitting in the resulting CARTs. However, individual sites were split in such a way that additional insights could not be obtained.

Aggregating by Number of Lanes Somewhat improved results did occur when CARTs were constructed using data only from five-lane weaving sections in the learning sample. The two sites not having five lanes were excluded from the analysis. The resulting weaving speed binary tree is illustrated in Figure 6. The causal relationships contained in this model appear to be logical.

Aggregating by Individual Site CARTs were constructed for each individual test site in an effort to identify factors significantly influencing the operation of each weaving sec-

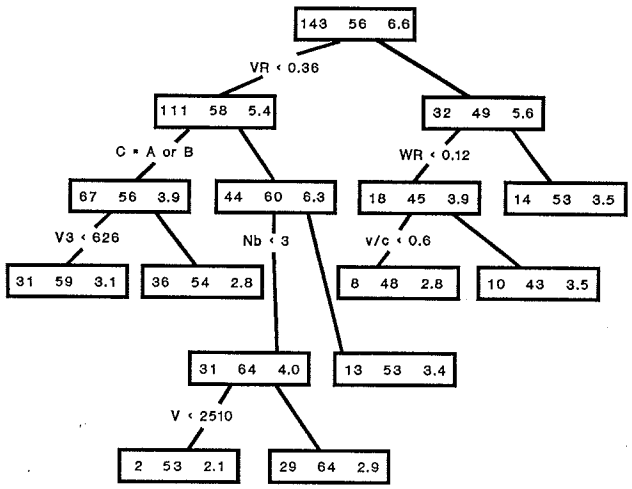


FIGURE 5 CART for weaving speed (all eight sites).

tion. Unfortunately, no single factor or group of factors seemed to dominate tree splitting.

Residual Prediction

CART analysis was also used in this research to predict speed-flow residuals. It was thought that, in the absence of flow impedance caused by weaving vehicles, traffic operating on major freeway weaving sections would exhibit conventional speed-flow relationships. Curves were constructed for speed versus *v/c* scatter plots, like the ones illustrated in Figure 2. The curves were drawn to be an envelope of the observed data points (see Figure 7).

Vertical distances between the envelope curve and each data point were considered to be the residuals. The CART program was then used to predict each residual.

It was assumed that freeway traffic on a straight-pipe section would have data points grouping closely to the envelope curve, and that perhaps factors related to the weaving phenomena (e.g., conflicts between weaving vehicles) were causing points to fall below the curve. By predicting the residuals, researchers thought, factors significantly affecting operation on the test sites might be identified.

Residual prediction was performed on the data set aggregating all eight test sites. Envelope curves and residual measurements were made for both the weaving speed versus *v/c* and the nonweaving speed versus *v/c* scatter plots. Data representing 15-min observations were used.

The resulting CART for weaving speed residual prediction is illustrated in Figure 8. The learning sample used in this analysis incorporated categorical variables representing configuration type and contains 143 data points with an average residual of 0 and a standard deviation of 5.0.

Developing residual prediction CARTs using data aggregated over numerous sites (see Figure 8) did not yield significant insights into factors influencing weaving section operation. Further efforts were also directed toward residual prediction for individual weaving sites. However, results did not appear to be extremely promising.

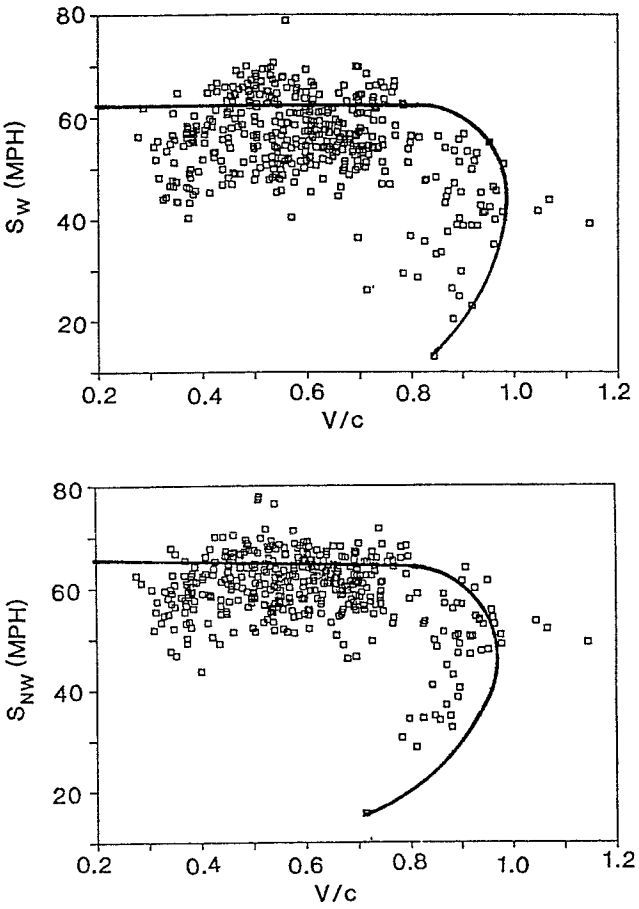


FIGURE 7 Sample envelope curves.

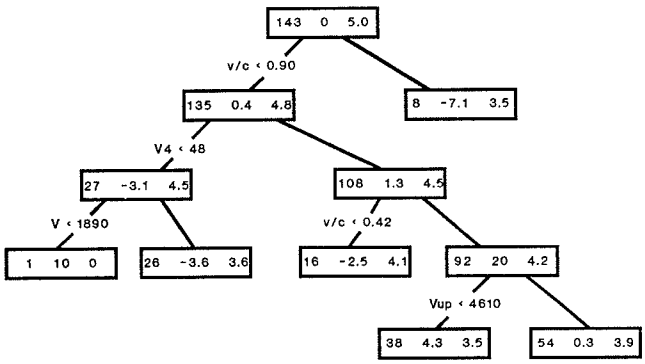


FIGURE 8 CART for weaving speed residuals (all eight sites).

CONCLUSIONS

Summary of Project Findings

The objectives of the research were to evaluate existing methods for the design and analysis of major freeway weaving sections and to develop new or modified procedures as deemed appropriate. Research results determined that existing speed prediction techniques are generally not very reliable, so efforts were made to calibrate improved prediction models.

Attempts to develop new speed prediction tecdhniques have not been completely successful. Analysis indicates that travel speed is not easily predicted, given aggregate flow information and geometric characteristics. Table 5 presents a comparison between average field-measured weaving speeds and average weaving speeds predicted using newly calibrated models. It also contains a sample of the data points collected as part of

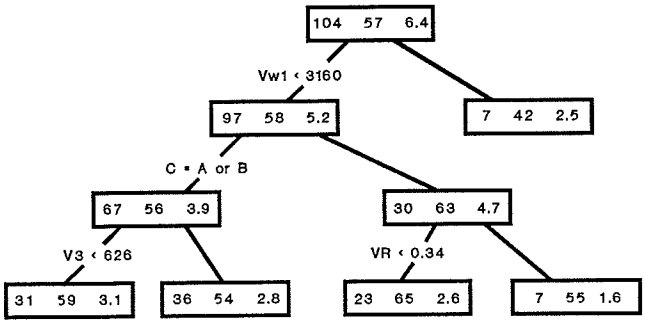


FIGURE 6 CART for weaving speed (N = 5).

the research. These data points are identical to those contained in Table 2. Each data point represents 60-min observations. Mean squared errors between field-measured and predicted weaving speeds are also listed in Table 5.

Comparing mean squared error values in Tables 2 and 5, recalibrated JHK and HCM-85 models represent improvements over their corresponding original models. This, of course, is to be expected since data used to validate recalibrated models are essentially the same data used to develop the models. As previously stated, small values for the recalibrated HCM-85 regression coefficient forced all predicted speeds to be 65 mph.

Newly developed models calibrated using regression techniques and CART may not be satisfactory. The traffic and geometric factors that most significantly affect the operation on test sites vary from location to location. Factors having the greatest overall influence on the operation of major freeway weaving sections therefore could not be identified.

Referring again to Table 5, the regression-based weaving speed prediction model calibrated with data from five-lane weaving sections (Equation 5) performed better than the regression equation calibrated with data from all eight test sites (Equation 3). As expected, the regression model calibrated using the cross-validation technique (Equation 7) seems to have the best predictive ability of all regression-based weaving speed prediction models.

It is rather surprising that the CART weaving speed model developed with data from all eight test sites (Figure 5) predicts speed better than the CART model calibrated with only from five-lane weaving sections (Figure 6). These results are opposite to those found for the regression models.

In general, none of the newly developed speed prediction models (weaving and nonweaving) seems to perform in a satisfactory manner. It should be noted that speed predictions in Table 5 were the results of input data used to calibrate the models. Had unused validation data been input to test the newly developed models, even greater prediction errors would likely occur. In addition, none of the newly developed speed prediction models readily accommodates design considerations.

Table 5 also lists average speeds predicted using the HCM's Chapter 3 method for analyzing straight-pipe freeway sections. Comparing these predicted speeds to weighted average values of observed weaving and nonweaving speeds yields relatively small prediction errors. It would seem that the HCM's Chapter 3 procedures more reliably predict average speeds on weaving sections than do the HCM's weaving analysis methods (*3*).

Analyses do indicate that average density can readily be predicted when *v/c* values are less than 0.8. These predicted densities can perhaps be used as a measure of effectiveness or incorporated as an input variable to predict speed. Preliminary evaluations do indicate that using density as an independent variable enhances speed-prediction regression models. Overall, it may be that average travel speed is not an ideal measure of effectiveness for weaving sections. The poor predictive abilities of speed-prediction models calibrated in this

TABLE 5 MEASURED VERSUS PREDICTED SPEEDS (NEWLY CALIBRATED MODELS)

CALTRANS DISTRICT	Test Site	Field Meas. Sw	Recal JHK Sw	Recal HCM Sw	EQ3 Sw	EQ5 Sw	EQ7 Sw	CART Fig 5 Sw	CART Fig 6 Sw	Field Meas. Speed	HCM Ch3 Calc. Speed
4	WB92 AM	51	58	65	52	70	55	53	65	53	58
4	WB92 NOON	53	59	65	55	58	57	53	55	55	58
4	SB280 AM	67	57	*	57	65	60	64	65	67	58
4	SB280 NOON	62	57	*	58	63	61	64	65	63	58
4	SB280 PM	60	56	*	56	62	61	64	65	61	54
7	SB101 AM	50	53	*	51	51	55	54	54	54	57
7	SB101 PM	55	52	*	49	53	51	54	54	58	56
7	NB101 AM	44	47	*	25	44	53	43	42	49	46
7	NB101 PM	56	51	*	33	54	48	53	55	60	57
7	EB10 AM	48	57	65	52	69	47	48	55	53	*
7	EB10 NOON	54	56	65	51	67	50	53	65	64	*
7	EB10 PM	41	52	65	37	64	40	43	42	42	*
7	WB10 AM	56	56	65	41	55	58	59	59	58	53
7	WN10 NOON	63	57	65	44	58	58	59	59	64	57
7	WB10 PM	58	57	65	44	59	54	59	59	61	56
8	WB10 AM	58	59	65	52	58	59	59	59	60	59
8	WB10 PM	64	59	65	51	57	57	59	59	65	59
11	NB805 AM	56	57	65	49	52	58	54	54	61	55
11	NB805 NOON	56	57	65	46	53	55	54	54	62	56
11	NB805 PM	52	57	65	48	51	56	54	54	57	56
	avg. difference+	4.4	10.4	8.5	5.6	3.9	2.2	3.6			4.5
	mean sqrd error	29	140	114	81	20	6	25			26

* METHOD COULD NOT BE APPLIED BECAUSE TRAFFIC AND/OR GEOMETRIC CHARACTERISTICS EXCEED LIMITS OF THE MODEL

\+ MEAN DIFFERENCE BETWEEN FIELD-MEASURED AND PREDICTED SPEEDS FOR EACH SAMPLE DATA POINT (ABSOLUTE VALUES USED TO REFLECT EACH DIFFERENCE)

project, as well as the speed-versus-v/c scatter plots in Figure 2, suggest that other performance measures should perhaps be investigated.

Future Research

Additional research on major freeway weaving sections is required to develop reliable prediction methodologies.

Perhaps one reason for the inability to reliably predict operating speeds lies in the nature of the independent variables utilized. In this research, traffic characteristics were analyzed in their aggregate form; total volumes for each traffic movement (e.g., freeway to freeway, freeway to ramp, and so on) were used to predict speed. Existing design and analysis methodologies typically take this form.

However, the operation of freeway weaving sections may be largely influenced by what is occurring in individual lanes. Congestion at freeway weaving areas often occurs as a result of breakdown in a single lane. In addition, the effects of conflicting weaving vehicles might best be modeled on a lane-by-lane basis.

Preliminary findings of some research indicate that the Level D Method (*5*) can reliably predict the lane utilization rate of vehicles at ramp-weave freeway sections. Perhaps here lies the solution to reliably predicting operation on major weaving sections. If the presegregation of vehicles entering a weaving section and the lane-changing maneuvers within the section can be modeled, improved prediction methods might be developed. Fazio (*12*) developed such a model and reported improved prediction results. What is needed is additional research to refine and validate a lane-changing model that can be used to perform design tasks.

Previously collected videotaped data will be used to calibrate the proposed model. Additional calibration and validation data will be used to develop the model so that more reliable inferences can be drawn about the influences of traffic and geometric factors on traffic flow behavior in major weaving sections. Simulations may also be used to augment empirical data. In this project, simulations conducted using the INTRAS model (*15*,*16*) have been performed and seem to hold promise. A detailed discussion of simulation experiments performed in this research project is described in Skabardonis et al. (*17*).

GLOSSARY

The terms listed below are parameters affecting weaving area operation.

- C = categorical variable representing configuration.
- d = average density, in pcph/mi/lane.
- L = length of weaving area (ft).
- N = total number of lanes in the weaving area.
- N_b = number of freeway lanes approaching the weaving area.
- S_{NW} = average running speed of nonweaving vehicles in the weaving area, in mph.
- S_W = average running speed of weaving vehicles in the weaving area, in mph.
- V = total flow rate in weaving area, in pcph.
- V_1 = total flow rate of the freeway to freeway traffic stream, in pcph.
- V_2 = total flow rate of the freeway to off-ramp traffic stream, in pcph.
- V_3 = total flow rate of the on-ramp to freeway traffic stream, in pcph.
- V_4 = total flow rate of the on-ramp to off-ramp traffic stream, in pcph.
- v/c = volume-to-capacity ratio (assume c = 2000 pcphpl).
- V_{NW} = total non-weaving flow rate in the weaving area, in pcph ($V_{NW} = V_1 + V_4$).
- VR = volume ratio, V_W/V.
- V_W = total weaving flow rate in the weaving area, in pcph ($V_W = V_{W1} + V_{W2}$).
- V_{W1} = weaving flow rate for the larger of the two weaving flows, in pcph.
- V_{W2} = weaving flow rate for the smaller of the two weaving flows, in pcph.
- V_{UP} = total upstream flow rate, in pcph ($V_{UP} = V_1 + V_2$).
- WR = weaving ratio, V_{W2}/V_W.

ACKNOWLEDGMENTS

The research described in this paper was made possible through funding from the California Department of Transportation and the Federal Highway Administration. The project staff wish to thank the many knowledgeable professionals at the California Department of Transportation who assisted us in selecting appropriate test sites for data collection. We wish to specifically acknowledge Howard Fong for his assistance and instruction on the use of video cameras. The authors would especially like to thank Fred Rooney of Caltrans Traffic Operations for his invaluable advice and insights concerning freeway traffic operations.

REFERENCES

1. H. K. Fong and F. D. Rooney. *Weaving Data*. State of California, Department of Transportation, 1987.
2. A. Skabardonis, M. Cassidy, and A. D. May. Evaluation of Existing Methods for the Design and Analysis of Freeway Weaving Sections. Presented at 67th Annual Meeting of the TRB, Washington, D.C., 1988.
3. *Special Report No. 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
4. M. Cassidy, A. Skabardonis, and A. D. May. *Operation of Major Freeway Weaving Areas: Recent Empirical Evidence*. Institute of Transportation Studies, Univeersity of California, Berkeley, 1988.
5. K. Moskowitz and L. Newman. Notes on Freeway Capacity. In *Highway Research Record 27*, HRB, National Research Council, Washington, D.C., 1963, pp. 44–68.
6. *Special Report 87: Highway Capacity Manual*, HRB, National Research Council, Washington, D.C. 1965, pp. 160–186.
7. J. E. Leisch et al. *Procedure for Analysis and Design of Weaving Sections*. Report FHWA/RD-85/083, FHWA, U.S. Department of Transportation, 1985.
8. J. E. Leisch. Completion of Procedures for Analysis and Design of Traffic Weaving Sections. Final Draft Report, Vol. 1, *Research Findings and Development of Techniques for Application*, Sept. 1983.

9. J. E. Leisch. A New Technique for Design and Analysis of Weaving Sections on Freeways. *ITE Journal*, Vol. 49, No. 3, Mar. 1979.
10. L. J. Pignataro et al. *NCHRP Report 159: Weaving Areas—Design and Analysis*. HRB, National Research Council, Washington, D.C. 1975.
11. W. Reilly, H. Kell, and P. J. Johnson. *Weaving Analysis Procedures for the New Highway Capacity Manual*. JHK and Associates, 1984.
12. J. Fazio and N. M. Raiphail. Freeway Weaving Section: Comparison and Refinement of Design Analysis Procedures. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986.
13. B. Persaud and V. 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.
14. L. Breiman et al. *Classification and Regression Trees (CART)*. Wadsworth Publisher, 1984.
15. D. A. Wicks and E. B. Lieberman. Development and Testing INTRAS, A Microscopic Freeway Simulation Model. In Vol. 1, *Program Design, Parameter Calibration and Freeway Dynamics Component Development*, Final Report FHWA/RD-80/106, Oct. 1980.
16. D. A. Wicks and B. J. Andrews. Development and Testing of INTRAS: A Microscopic Freeway Simulation Model. In Vol 2, *User's Manual*, Final Report FHWA/RD-80/L07, Oct. 1980.
17. 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*. Institute of Transportation Studies, University of California, Berkeley, 1988.

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

Estimating Capacity and Delay at a Single-Lane Approach, All-Way Stop-Controlled Intersection

MICHAEL KYTE AND JOSEPH MAREK

This paper presents the results of a study of single-lane approach, all-way stop-controlled intersections. Data have been collected for nearly 25 hr of operation for eight sites in Idaho, Oregon, and Washington. Estimates of the capacity and nature of the delay/flow rate relationship have been made for these sites. A methodology for analyzing operational performance is proposed. This study is part of a larger effort where data have been collected for 23 all-way stop-controlled intersections in Idaho, Oregon, Washington, Iowa, Colorado, and Texas. In the next phase of this study, the effects of nonstandard conditions (number of approach lanes, pedestrians, heavy vehicles, etc.) will be considered.

The 1985 *Highway Capacity Manual* (HCM) (*1*) establishes vehicle delay as the primary measure of effectiveness (MOE) for evaluating the performance of signalized intersections. Delay is the logical parameter because it satisfactorily describes intersection performance, can be easily measured by the transportation engineer, and can be clearly communicated to the layperson. Much literature exists on the operation and performance of signalized intersections. The relationship between vehicle delay and traffic flow and the procedures for capacity and level of service analysis are well documented.

The situation is different for unsignalized intersections. Reserve capacity is used to define the level of service for two-way stop-controlled (TWSC) intersections. This is a difficult notion for the layperson and not as satisfactory a concept as delay for the transportation engineer. Further, the method described in Chapter 10 of the HCM for evaluating TWSC intersections is based on procedures developed in Europe that have yet to be calibrated with data from the United States. The situation is even worse for all-way stop-controlled (AWSC) intersections. The HCM cites only one study as the basis for rather meager capacity guidelines for AWSC intersections. The Transportation Research Board (*2*) recognizes this deficiency and has identified the development of capacity and level-of-service analysis procedures for AWSC intersections as a high priority for future research. Clearly, there is a need for an improved methodology that can be used in the analysis of AWSC intersections. Further, there is a need for a data base to investigate the relationships between vehicle delay, intersection flow rates, and other key variables that affect intersection performance.

In response to this need, in 1987 the University of Idaho and the Idaho Transportation Department jointly initiated a research project to study the traffic flow characteristics of AWSC intersections. This paper describes the results of this study: the accumulation of a data base, the identification of the relationship between delay and traffic flow rates, and the estimation of intersection capacity. The focus, in general, is on the four-way, single-lane approach, AWSC intersection and specifically on the eight sites for which data have been collected and analyzed.

In this paper, previous studies in this area are described, the data collection method and the site characteristics are presented, and the flow rate/delay relationship is analyzed. Procedures for estimating intersection capacity are given, an evaluation method is proposed, and findings and conclusions are summarized.

Department of Civil Engineering, University of Idaho, Moscow, Idaho 83843.

PREVIOUS WORK

A literature search was undertaken to determine relevant previous research on stop-controlled intersections (*3*). Two of the more important studies, one empirical and the other theoretical, are summarized below.

Empirical Study by Hebert

In 1963, Hebert (*4*) presented the results of a study of three AWSC intersections in Chicago. The objective of his research was to determine the capacities of AWSC intersections under a variety of traffic and operating conditions. He investigated the average departure headway for an approach for three cases:

1. When both the major and minor streets are loaded,
2. When the subject approach is loaded but no vehicles are on the cross street, and
3. When the study approach is loaded and is affected by cross-street vehicles.

Data were gathered using a movie camera operating at a rate of 100 frames/min, or one frame every 0.6 sec. Three intersections were filmed (two single-lane approach intersections and one multilane approach intersection), and the vehicle headways were calculated. It was found that the ratio of major street volume to total intersection volume (defined as "volume split") affects departure headway and, thus, capacity. Hebert's estimates of departure headway and capacity are the basis for Table 10-5 in the HCM.

Theoretical Model by Richardson

Besides a lack of data, there has also been a paucity of theory to explain the operation of AWSC intersections. Richardson (*5*) addressed this problem by proposing a model based on queuing theory for a single-lane approach, AWSC intersection with no turning movements. Richardson used departure headway data from Hebert's work as a basis for his model.

Richardson proposed an M/G/1 queuing model that assumes Poisson-distributed vehicle arrivals and service rates described by a general distribution. The queue discipline is single server, first come, first served, since a single-lane approach is assumed. Richardson noted that the service time for a vehicle on any approach is bimodal: either there is a vehicle waiting on one of the conflicting approaches or there is not. For the first case, the service time includes the time for both the conflicting vehicle and the waiting vehicle to clear the intersection, or 7.6 sec from Hebert's study for a single-lane approach case. For the second case, the service time, also from Hebert's study, is 4.0 sec. In calculating the actual service time for each approach, there is an interaction between the approaches that relates to the probability of vehicles waiting for service on each of the approaches (which is a function of the demand on each approach). The model calculates the service times iteratively, and convergence to stable values is rapid.

Richardson used his model to forecast intersection capacity and delay under a variety of volume split conditions. The results of his capacity analysis correlate well with the forecasts of Hebert's intersection capacity equation. Further, Richardson postulated that delay on a given approach is a function of, in decreasing order of importance, flow rate on the subject approach, flow rate on the cross streets (conflicting flow rate), and flow rate on the opposing approach. The Richardson model can be programmed on an electronic spreadsheet (e.g., Lotus 123 or Quattro), and delay calculations can be easily performed.

DATA COLLECTION AND SITE CHARACTERISTICS

The review of the literature revealed that there is little data on capacity or delay for AWSC intersections. Thus, one of the first tasks in this study was to assemble a data base of traffic flow rate and delay data.

Delay can be defined in several ways. "Stopped delay," used here, is defined as that time beginning when a vehicle enters a queue and ending when it crosses the stop line at the intersection and leaves the queue. The use of stopped delay is consistent with the signalized intersection methodology of Chapter 9 of the HCM.

One common method to collect delay data is to monitor the length of the vehicle queue at periodic intervals (*1*). This method yields good results for signalized intersections where at least some of the vehicles are stopped for a portion of the signal cycle. For AWSC intersections, delays may often be less than the sampling interval. Thus, to successfully collect delay data at an AWSC intersection, the progress of individual vehicles in the queue, from entrance to exit, must be traced.

A video camera was used for data collection in this study, which required only one person in the field to collect the data. With data in real-time videotape format, intersection operations can be reviewed as often as needed to record additional data or to observe traffic dynamics. The camera is placed to provide an unobstructed view of all intersection approaches and the resultant vehicle turning movements. In addition, the view must include all queue formation and dissipation activity for at least one of the approaches.

Software was developed that allows delay and flow rate data to be entered into the computer while the videotape of the intersection operation is being observed. The demands placed on software used for real-time data entry are severe and require a certain degree of robustness in the program operation (*6*).

While the eight study sites have a common geometry (four approaches with a single lane on each approach), traffic conditions vary widely among the sites. A comparison of the flow rate and delay data for the sites is given in Table 1.

METHODS OF ESTIMATING DELAY

Forecasting Vehicle Delay

One of the most important objectives of this study is to develop a method for forecasting vehicle delay at AWSC intersections. To achieve this objective, the factors contributing to vehicle delay must first be identified and then quantified. Two methods for estimating vehicle delay are described here. The first is empirically based and uses regression analysis to determine the relationship between subject approach delay and subject approach flow rate, conflicting approach flow rates, and opposing flow rate. The second method is a validation of Richardson's theoretical queuing model to determine how accurately the model replicates "real world" intersection operations.

Empirical Estimation of Vehicle Delay

Plots of delay versus subject flow rate for each site show, in general, an exponential relationship between these variables. Figure 1 shows an example of this plot for Site 4.

A basic model form is suggested using Richardson's hypothesis that subject delay is a function of subject, conflicting, and opposing flow rate. Delay/flow rate equations were developed for each of the eight sites individually and for all 297 data points together using this hypothesized relationship. Equation parameters for the best model fit for each site are summarized in Table 2.

Several conclusions can be drawn from an analysis of these models:

- Subject flow rate is the most important contributor to vehicle delay, and the subject flow rate coefficient is numerically the highest coefficient for each model.
- Conflicting flow rate is also a significant factor in forecasting vehicle delay, but somewhat less important than subject flow rate.
- Opposing flow rate is only significant in those models where opposing flow rate exceeds 300 vph. For those sites where opposing flow rate never exceeds 300 vph for any 5-min period, it is not a significant factor in estimating delay.

TABLE 1 FLOW RATE AND DELAY DATA

Mean 5-Minute Data

Site	Location	Intersection Flow Rates	Approach Flow Rates (vph) Subject	Conflicting	Opposing	Delay (sec)
1	Moscow,ID	851	417	309	126	13.5
2	Moscow,ID	960	340	469	151	15.5
3	Portland,OR	922	476	348	98	6.0
4	Aloha,OR	1453	391	659	403	22.2
5	Boise,ID	1140	461	233	233	19.8
6	Boise,ID	865	388	388	185	11.2
7	Boise,ID	975	312	336	327	10.5
8	Spokane,WA	548	202	279	67	4.7

5-Minute Data Ranges

Site	Location	Intersection Flow Rates	Approach Flow Rates (vph) Subject	Conflicting	Opposing	Delay (sec)
1	Moscow,ID	624-1164	288-636	252- 420	72-168	5-40
2	Moscow,ID	700-1140	180-480	360- 600	60-216	7-34
3	Portland,OR	612-1104	288-648	252- 480	12-180	3- 8
4	Aloha,OR	984-2016	204-564	360-1032	192-600	5-73
5	Boise,ID	852-1488	252-660	144- 720	144-324	5-48
6	Boise,ID	420-1404	144-732	36- 660	24-684	5-37
7	Boise,ID	708-1164	156-492	180- 480	204-528	5-20
8	Spokane,WA	276- 804	108-396	96- 468	0-156	3- 7

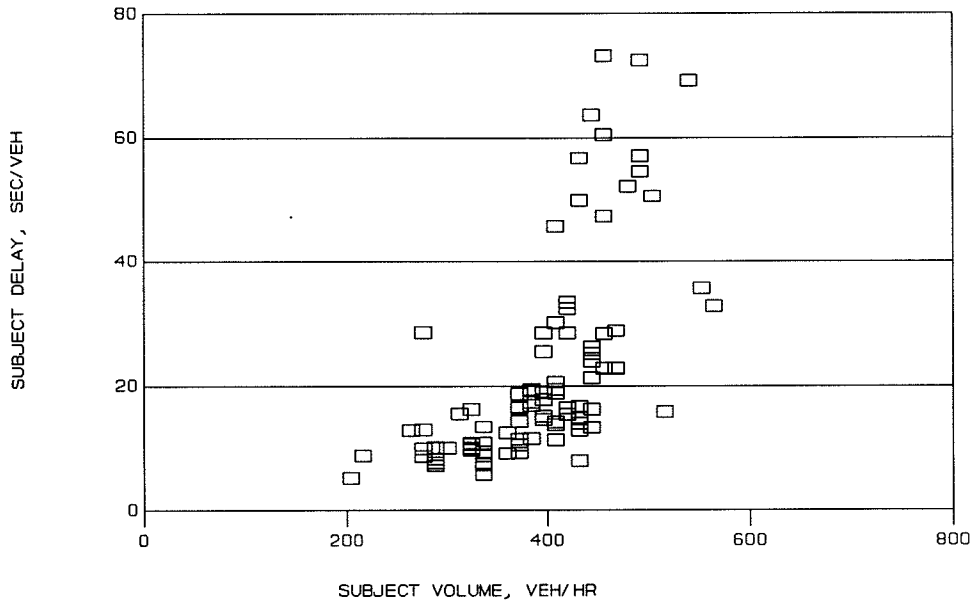


FIGURE 1 Delay versus subject volume, Site 4.

- Contrary to Richardson's hypothesis, the effect of opposing flow rate is approximately equal to that of conflicting flow rate.
- The exponential model form best fits the data for seven of the eight sites and for the overall model. For the other site, flow rates are relatively low and delay remains nearly constant at 5 to 10 sec/veh. The linear form is most appropriate for this case.
- For the sites where high delays were measured (greater than 30 sec/veh for any 5-min period), the models tend to underpredict delays consistently. This leads to the conclusion that there are probably two regions of delay. The first is for low subject approach flow rates (less than 300 to 400 vph) in which delays are constant and average about 5 sec/veh. The second region is for higher flow rates (greater than 400 vph) in which delays increase exponentially.

TABLE 2 DELAY/FLOW RATE MODELS

Site	SubVol	ConVol	OppVol	Constant	R-Sq	Functional Form
1	.00401 ±.00052	.00231 ±.00072	n/s		.99	Exponential
2	.00518 ±.00111	.00181 ±.00082	n/s		.98	Exponential
3	.00243 ±.00080	.00161 ±.00108	n/s		.98	Exponential
4	.00490 ±.00058	.00068 ±.00026	.00131 ±.00055		.98	Exponential
5	.00399 ±.00050	.00216 ±.00051	n/s		.99	Exponential
6	.00240 ±.00029	.00052 ±.00026	.00068 ±.00028	2.99029 ±.25207	.73	Exponential
7	.00377 ±.00064	.00219 ±.00062	.00108 ±.00055		.99	Exponential
8	.00480 ±.00169	.00316 ±.00170	n/s	2.86427 ±.54244	.35	Linear
All	.00375 ±.00019	.00132 ±.00016	.00153 ±.00023		.98	Exponential

Coefficient estimates are shown with their standard errors.
R-Sq is the coefficient of determination, a measure of goodness of fit of the model to the data.
n/s = the variable was not statistically significant and was not included in the final model estimation.

The forms for the linear and exponential equations are shown below:

$$D = a_1\ SUBVOL + a_2\ CONVOL + a_3\ OPPVOL + Constant$$

$$D = Constant \times \exp(a_1\ SUBVOL + a_2\ CONVOL + a_3\ OPPVOL)$$

Validation of Richardson Model: Theoretical Estimation of Vehicle Delay

The data collected at the eight sites were used to test the validity of the Richardson model. The validation was performed by using the model to calculate vehicle delay based on the measured flow rates and comparing this estimate with measured delays.

As seen from Figure 2, the Richardson model yields good forecasts of subject delay for ranges of subject flow rates up to 400 vph. For this range, residuals (i.e., the difference between model forecasts and actual delay values) are usually 2 sec or less. As subject flow rates increase beyond 400 vph, the model forecasts deteriorate, which is probably the result of two factors. First, the model is based on capacity departure headways estimated by Hebert for three cases of volume splits. These headways have not been verified by other researchers. If they are in error, they will yield incorrect results in the Richardson model. Second, the model does not consider the effects of opposing left turns on subject delay.

METHODS OF ESTIMATING CAPACITY

This section develops and analyzes alternative methods of estimating the capacity of AWSC intersections.

Measuring Capacity

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 specified time period under prevailing roadway, traffic, and control conditions." There is a fundamental difference in the notion of capacity for uninterrupted flow and interrupted flow facilities. The HCM specifies that 15-min periods are to be used when measuring capacity on an uninterrupted flow facility. Since traffic flows are subject to instability, primarily because of the interactions between vehicles, there must be some degree of sustainability if a high flow rate is indeed a capacity flow rate and not just a transient effect.

By contrast, the capacity of interrupted flow facilities is controlled not by vehicle interaction but by the nature of the intersection control. Because of the cyclic nature of its operation, intersection capacity is measured on a time scale much less than 15 min. In fact, capacity data for both signalized and unsignalized intersections are based on saturation departure headways. Another issue is the importance of intersection capacity versus approach capacity. The HCM suggests that, for signalized intersections, approach capacity, rather than intersection capacity, is of paramount importance. For unsignalized intersections, however, there is an important interaction between flows on each approach. Hebert showed that

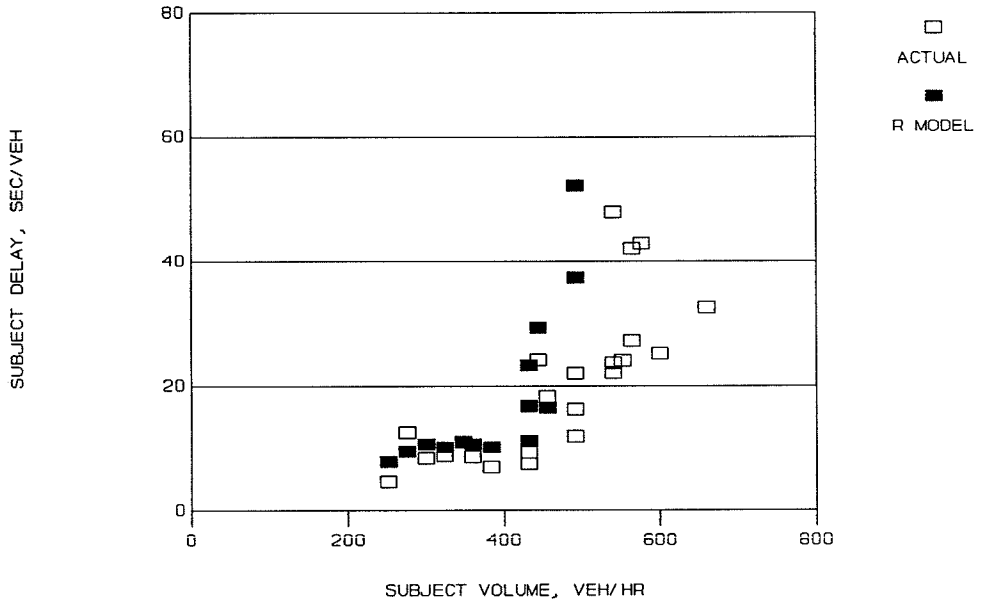


FIGURE 2 Validation of Richardson model, Site 5.

the "capacity" departure headways on one approach of an intersection increase as the volumes on the conflicting and opposing approaches increase.

Finally, when measuring capacity, it is important to understand the basic patterns of traffic flow at AWSC intersections. At low volumes, a vehicle is served as it arrives; while the rule of priority to the vehicle on the right may apply, there is no other apparent pattern to the intersection operation. As demand increases and approaches capacity, a pattern does begin to emerge. For the single-lane approach case, opposite approaches flow simultaneously, and a two-phase operation develops. For multilane approaches, a four-phase operational pattern emerges; that is, each approach travels as a group.

Methods of Estimating Capacity

The method used to estimate the capacity of an AWSC intersection must consider the importance of the interaction of traffic on the various approaches. If the volume on one approach increases, the volumes that can potentially be served on the other approaches are reduced.

On a microscopic basis, the fastest rate at which vehicles can depart is the capacity of that approach. Hebert estimated the departure headway to be 4.05 sec when there is no conflicting flow. This implied a capacity of 890 vph per approach. At the other extreme, he estimated a departure headway of 7.65 sec for a fully loaded intersection, or an approach capacity of 470 vph. Another method for estimating capacity is to examine the empirically based models using the delay/flow rate data for the intersections studied here. When an approach is at capacity, delay approaches very high values. Delay can continue to increase indefinitely as queues continue to increase in length, but there is some maximum acceptable level of delay that drivers will tolerate. Plots of delay versus flow rate show the point where delay begins to increase rapidly; this is the point at which the intersection has reached unstable conditions. The asymptotic value of flow rate might be called the theoretical capacity.

Two methods are explored here: an analysis of the highest flow rate observations and an analysis of departure headway data.

High Volume Observations

One method of estimating capacity is to record the highest measured flow rates at an intersection and to study the conditions present when these high flow rates are observed. A total of 297 5-min observations were made during this study. Of this number, 10 observed intersection flow rates exceeded 1,700 vph, and 33 observations were above 1,600 vph. The highest single observation was 2,016 vph at Site 4. The maximum observed flow rate on an individual intersection approach was 732 vph, recorded at Site 6.

According to Hebert's hypothesis, the highest approach flow rates will be observed on those approaches where volume splits are the highest; this occurs when traffic is concentrated on only one approach. This finding is confirmed in the present study in an analysis of high approach flow rates. In Figure 3, approach flow rates are plotted against the proportion of flow on the subject approach. As expected from Hebert's hypothesis, maximum observed flow rates increase as that approach becomes the dominant flow of the intersection approaches. A regression equation relating departure headway (the reciprocal of flow rate) on the subject approach to the proportion of flow on the subject approach is

$$H = -3.894(\%SUBVOL) + 8.2099 \quad (1)$$

A similiar pattern emerges between intersection flow rate and the distribution of traffic on the intersection approaches. Figure 4 shows a plot of high intersection flow rates versus

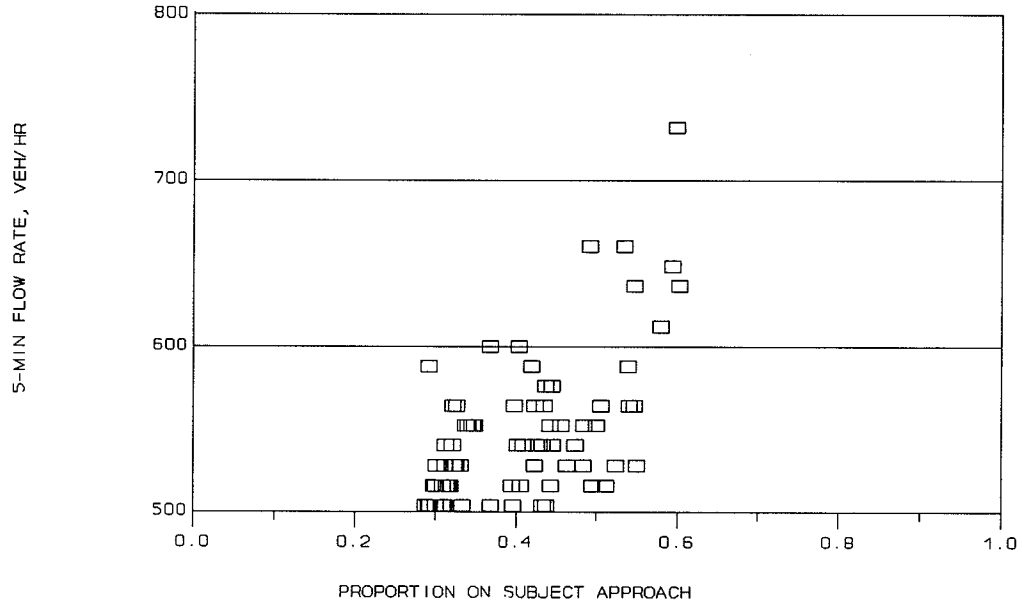


FIGURE 3 Subject approach volumes versus proportion of subject volume, high volume cases.

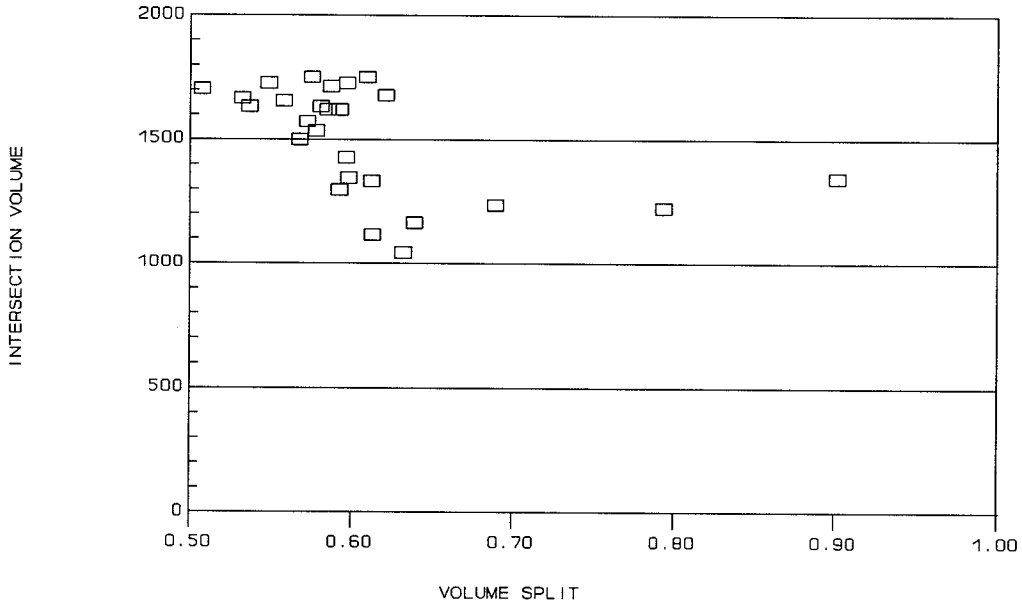


FIGURE 4 Intersection flow rate versus volume split for subject delays above 30 sec.

volume split for cases when delay exceeds 30 sec/veh. These high delay cases approximate capacity conditions.

Headway Analysis

Headway data recorded to one-tenth of a second were collected as part of this study and calculated based on the time between two vehicles departing a given approach.

The conditions of traffic flow on the subject, opposing, and conflicting approaches were noted for each departing vehicle. The subject approach is considered to be loaded or "at capacity" if there is at least one vehicle in queue or waiting to be served. The capacity departure headway for the subject approach can be calculated for several basic situations:

1. When no opposing or conflicting vehicles are waiting, and two subject approach vehicles can travel through the intersection consecutively with no other intervening vehicles;
2. When one opposing vehicle crosses through the intersection between consecutive subject approach vehicles;
3. When one or more conflicting vehicles cross the intersection between two subject approach vehicles; and
4. When both conflicting and opposing vehicles cross the intersection between subject approach vehicles.

So far, headway data have been summarized only for Site 1 (see Table 3).

Even with this limited sample, several important facts are revealed:

1. Departure headways are lowest when there are no vehicles on the conflicting or opposing approaches. The mean value for this case is 4.1 sec, approximately the same value measured by Hebert.
2. Departure headways are somewhat higher when there is a vehicle on the opposing approach. The mean headway is 5.1 sec for this case, but this distribution may be bimodal, indicating that there may be a different effect from straight-through versus left-turning opposing vehicles.
3. Departure headways for Cases 3 and 4 (conflicting vehicles present) average 7.0 and 8.3 sec, respectively.

PROPOSED EVALUATION METHODOLOGY

Elements of an Operational Analysis

One of the basic tools of the HCM is the operational analysis, a technique for determining the capacity and level of service of a highway facility. The operational analysis method has been developed for nearly all types of highway facilities including freeways, multilane highways, rural highways, signalized intersections, arterial systems, and two-way stop-controlled unsignalized intersections. However, there is currently no operational analysis method established for AWSC intersections.

The common features of the operational analysis method include

- Definition of the MOE, a parameter that can be used to assess the performance of the facility over a wide range of operating conditions;
- A method of forecasting the MOE for a given set of traffic and geometric conditions; and
- A method for estimating the capacity of the facility.

Based on the results of this study, an operational analysis method for AWSC intersections should include the following three attributes:

1. Vehicle delay should be used as the MOE. It can be measured easily and adequately describes how well the intersection is performing from the driver's perspective. Use of delay as the MOE will provide for a consistency of analysis and performance evaluation between signalized and nonsignalized intersections.
2. Vehicle delay should be estimated as a function of the forecasted volumes for each intersection approach. Delay has been shown to be a function of the flow rates on the subject, conflicting, and opposing approaches. This relationship appears sound from both theoretical and empirical perspectives.
3. Capacity should be estimated as a function of the distribution of flows on each intersection approach.

While it may be premature to suggest an operational analysis methodology based on a study of only eight sites, it is useful to at least describe possible methods for estimating two of the important parameters needed to perform an operational analysis, namely capacity and delay. A set of alternative methods is presented below for estimating delay and capacity for AWSC intersections. It is hoped that these methods can be field tested by other researchers and practicing engineers.

Suggested Methods for Estimating Vehicle Delay

Two methods are proposed for estimating vehicle delay, one analytical and one graphical.

TABLE 3 DEPARTURE HEADWAY DATA, SITE 1

Departure Headways

Case	Number of Observations	Maximum Headway (sec)	Minimum Headway (sec)	Mean Headway (sec)	Estimated Capacity (vph)
1	155	10.8	1.3	4.1	878
2	32	15.2	2.5	5.1	706
3	98	12.4	2.5	7.0	514
4	29	16.1	5.1	8.3	434

Capacities

Case	Subject Headway (sec)	Subject Flow (vph)	Resultant Conflicting Flow (vph)	Resultant Opposing Flow (vph)	Estimated Intersection Capacity (vph)	Percent Subject Volume
1	4.1	878	0	0	878	1.00
2	5.1	706	0	706	1412	.50
3	7.0	514	868	0	1382	.33
4	8.3	434	868	434	1736	.25

Estimating Vehicle Delay: Method 1

The graphical method is based on the plots of subject approach delay versus subject approach, conflicting, and opposing approach flow rates. The graph is based on an idealized description of the delay/flow rate relationship:

- Delay is constant at approximately 5 sec/veh up to a subject flow rate of 300 vph.
- At subject flow rates above 300 vph, the curve branches into several segments, based on the conflicting and opposing flow rates, as delay increases at an exponential rate.

At least with respect to order of magnitude, conflicting and opposing flow rates have been found to affect delay equally. Thus, it is proposed that the sum of these two variables determine the proper curve to be used in estimating delay. The proposed delay estimation curves are shown in Figure 5.

Estimating Vehicle Delay: Method 2

The analytical method is based on an equation relating vehicle delay on the subject approach to the flow rates on the subject, conflicting, and opposing approaches. The equation was developed using data for all eight study sites:

$$D = e^{.00375\ SUBVOL\ +\ .00132\ CONVOL\ +\ .00153\ OPPVOL} \quad (2)$$

Suggested Methods for Estimating Intersection Capacity

Two methods are suggested for estimating intersection capacity: one based on measured departure headways at capacity operations, and the other based on the characteristics of the high flow rates measured as part of this study.

Estimating Intersection Capacity: Method 1

The departure headways calculated for Site 1 under various flow configurations can be used to estimate intersection capacity. Capacity values for four flow cases are given in Table 3. These represent the range of intersection loading conditions described earlier. The subject approach capacity for each case is estimated by the following equation:

$$\text{Subject approach capacity} = \frac{3{,}600}{\text{Mean departure headway}} \quad (3)$$

Using this equation, the subject approach capacity for Case 1 is 878 vph. The intersection capacity for Case 1 is simply the subject approach capacity, since no vehicles are present on any of the other approaches.

The subject approach capacity for Case 2 is estimated to be 706 vph. Because this case assumes that one opposing approach vehicle enters the intersection between two consecutive subject approach vehicles, by symmetry, the opposing approach capacity is also equal to 706 vph. Thus, the intersection capacity is 1,412 vph.

The subject approach capacity for Case 3 is 514 vph, but the estimation of intersection capacity is somewhat more complex because of the method used to estimate departure headways. Since this case includes vehicles from either one or both of the conflicting approaches traveling between consecutive subject approach vehicles, conflicting volumes can range from 514 to 1,028 vph. If the maximum conflicting volume is assumed to occur at capacity conditions, the intersection capacity is estimated to be 1,542 vph.

For Case 4, the subject approach capacity is 434 vph. This case includes one opposing vehicle and either one or two conflicting vehicles (one from each approach) traveling between consecutive subject approaches. Thus, the conflicting and

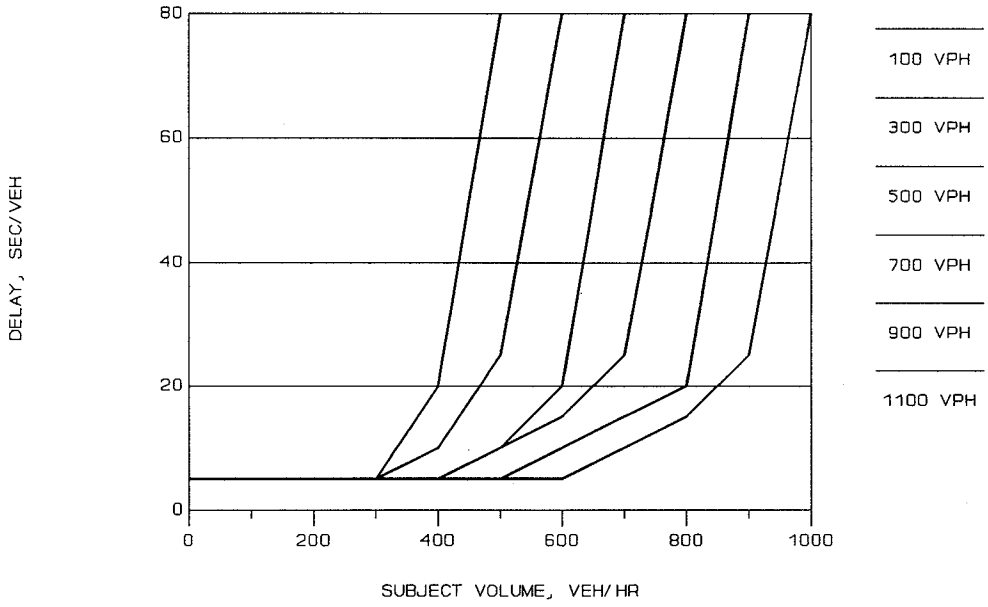


FIGURE 5 Graphical delay estimation method (conflicting and opposing volume range of 100 to 1,100 vph).

opposing approach flow rates can range from 868 to 1,302 vph. If the maximum conflicting volume is assumed to occur at capacity conditions, the intersection capacity is estimated to be 1,736 vph.

This data is used to develop the relationships between subject approach capacity and intersection capacity as a function of the proportion of traffic on the subject approach. A plot of this relationship, shown in Figure 6, can then be used to estimate intersection capacity as a function of conditions on the subject approach.

Estimating Intersection Capacity: Method 2

The proportion of traffic on the subject approach for high flow rate observations was analyzed and an equation relating departure headway (the reciprocal of subject approach flow rates) to this variable was estimated:

$$H = -3.894(\%SUBVOL) + 8.2099 \quad (4)$$

Table 4 gives estimates of subject approach and intersection capacity using Equation 4 for a range of intersection flow conditions (see Figure 7).

FINDINGS AND CONCLUSIONS

The results presented in this paper are the first step in the development of a methodology to determine the level of service of unsignalized intersections and to provide a consistent measure of effectiveness for all types of intersections. The data represent a wide range of delay and flow rate conditions.

Several important findings and conclusions result from this study:

- Delay on a given approach (called the subject approach) is a function primarily of the flow rate on that approach. It is secondarily a function of flow rates on the conflicting and opposing approaches, which contribute to delay in an approximately equal manner.
- Delay remains constant (at approximately 5 sec/veh) for low flow rates, up to subject flow rates of 300 to 400 vph. At this point, delay begins to increase exponentially, and asymptotically approaches the subject approach capacity point.
- The minimum or capacity departure headways on a given approach are a direct function of the traffic patterns on the conflicting and opposing approaches. The lowest departure headways occur when there is no traffic on either the conflicting or opposing approaches. Departure headways increase with the following conditions:
  1. No traffic on either the conflicting or opposing approaches,
  2. Traffic on the opposing approach only,
  3. Traffic on the conflicting approach only, and
  4. Traffic on both conflicting and opposing approaches.
- Methods for estimating delay and capacity are proposed. Two methods for estimating delay are described, one using a graphical approach and the other an analytical approach. Two methods for estimating capacity are proposed, one based on departure headways and the other on conditions observed for high flow rate situations. It is hoped that other researchers will test these methods and suggest improvements where warranted.
- Hebert's hypothesis that the proportion of traffic on the major and minor streets affects intersection capacity has been verified. However, a more accurate indicator is the distribution of traffic on each of the four approaches. Hebert's estimates of capacity are probably low, however, and his assertion that capacity is not affected by the proportion of left-turning vehicles is probably incorrect.
- Richardson's queuing model provides good estimates of vehicle delay for subject flow rates up to 400 to 450 vph. Above this point, the model gives poor results. Two likely reasons are incorrect departure headways as estimated by

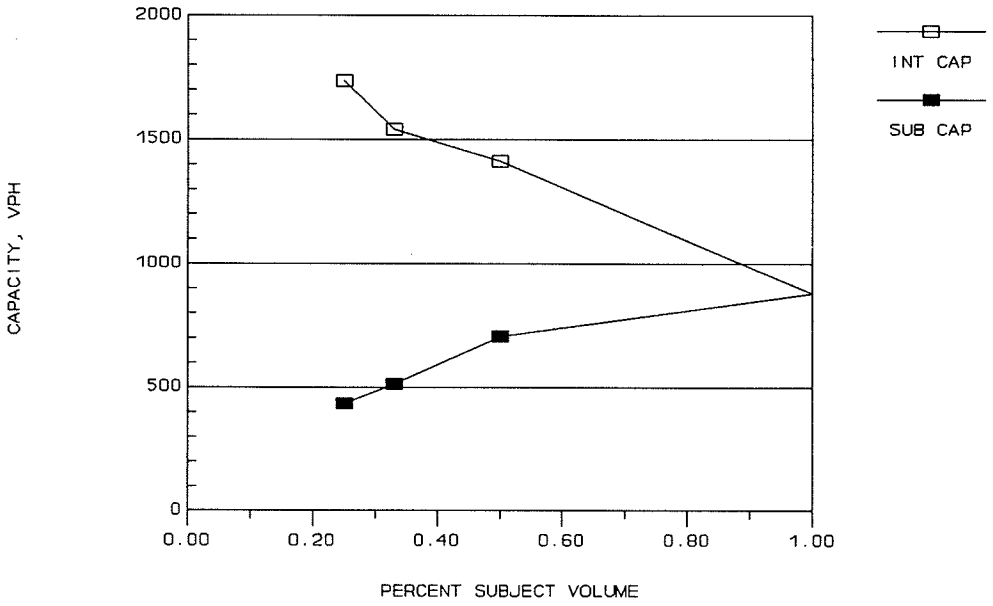


FIGURE 6 Capacity estimation, Method 1.

TABLE 4 CAPACITY AS A FUNCTION OF PERCENT SUBJECT VOLUME

%SubVol	Headway	Approach Capacity	Intersection Capacity	Hebert's Estimate
.25	7.236	497	1990	1882
.30	7.042	511	1704	1678
.35	6.847	526	1502	1547
.40	6.652	541	1353	1463
.45	6.458	557	1239	1416
.50	6.263	575	1150	1398
.55	6.068	593	1079	
.60	5.873	613	1022	

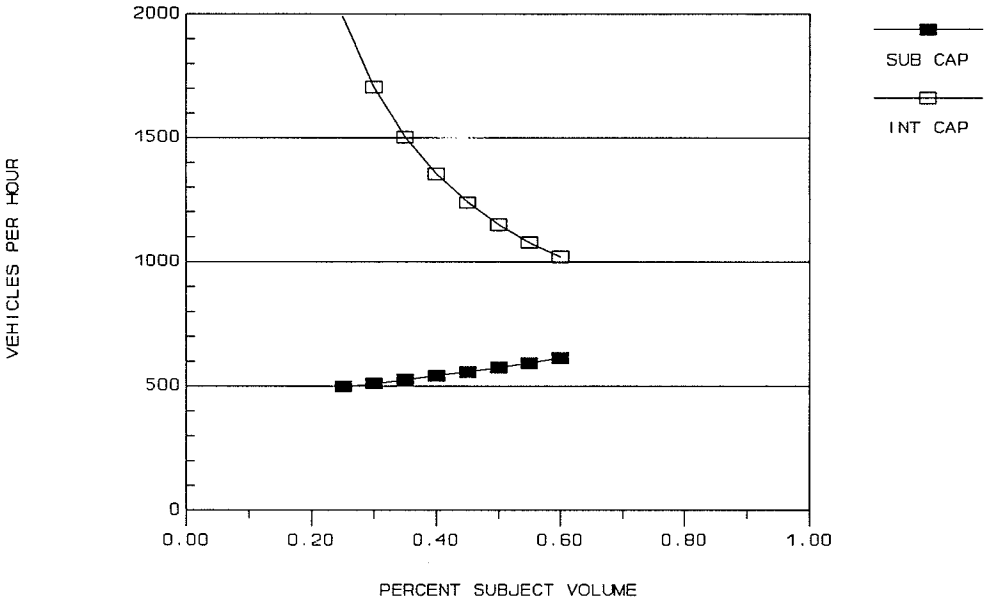


FIGURE 7 Capacity estimation, Method 2.

Hebert and the omission of the effects of opposing left-turning vehicles in the model.

More work must still be done to provide a complete understanding of the operations of AWSC intersections. It is hoped that the work summarized in this paper provides both motivation and a good foundation for future research.

REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
2. *Transportation Research Circular 319: Research Problem Statements: Highway Capacity*. TRB, National Research Council, Washington, D.C., 1987.
3. M. Kyte and J. Marek. *Stop-Controlled Intersections: An Annotated Bibliography*. Department of Civil Engineering, University of Idaho, rev. April 1988.
4. J. Hebert. A Study of Four-Way Stop Intersection Capacities. In *Highway Research Record 27*, HRB, National Research Council, Washington, D.C., 1963, pp. 130–147.
5. A. J. Richardson. A Delay Model For Multiway Stop-Sign Intersections. *Transportation Research Record 1112*. TRB, National Research Council, Washington, D.C., 1987, pp. 107–114.
6. M. Kyte and J. Marek. Collecting Traffic Data at Stop-Controlled Intersection. *ITE Journal*, April 1989.

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

Testing Delay Models with Field Data for Four-Way, Stop Sign-Controlled Intersections

MARK ZION, GEORGE F. LIST, AND CHARLES MANNING

Four-way, stop sign-controlled intersections are a relatively common phenomenon, especially in urban networks, yet little analysis has been devoted to determining their capacity and delay characteristics. This paper presents the results of new field studies and compares data collected from two recent delay models. Generally, findings are that delay increases as the intersecting volumes increase; intersections with balanced volumes have lower delays than those without; and the percentage of left turns has a noticeable effect on delay. Statistical analyses suggest that one of the two models considered in this research may provide satisfactory delay estimates.

Four-way, stop sign controlled intersections (FWSC) are a relatively common phenomenon, especially in urban networks, yet little analysis has been done to determine their capacity and delay characteristics (*1–6*). This paper reports the results of recent field studies on delay and compares data collected with predicted delays for two proposed models.

One of the first analyses of capacity and delay for FWSCs was conducted by Hebert (*1*), who collected data from three intersections in the Chicago area. He determined that, among other things, delay increased as volume increased and that the capacity of FWSC intersections increased as the volume split between the two streets became even (i.e., approached a 50/50 split). As the data in Table 1 show, capacity ranged from 1,900 veh/hr for an intersection with a 50/50 split to 1,500 veh/hr for one with a 70/30 split.

More recently, Richardson (*5*) developed a model, based on Hebert's work, that predicts the stopped delay for FWSC intersections. Richardson's model makes certain simplifying assumptions, but it does produce capacity estimates nearly identical to those found by Hebert (see Table 2). The model assumes that no turning movements occur; that a certain prescribed pattern (arguably plausible) operates at the intersection; and that the volumes on the opposing approaches are equal.

Assuming the two streets are labeled a and b (these labels can also be applied to the approaches on those streets), Richardson's model asserts that the average length of queue (vehicles), L_a, can be obtained by applying the Pollaczek-Kyintchine formula, as follows:

$$L_a = \frac{2u_a - (u_a)^2 + (x_a)^2 V(s_a)}{2(1 - u_a)} \quad (1)$$

where

x_a = average arrival rate for approach a (veh/sec),
s_a = average service time for approach a (sec/veh),
$V(s_a)$ = variance of service time for approach a, and
u_a = utilization ratio (unitless) for approach a = (arrival rate)*(service time).

Substituting subscript b for a throughout yields the average length of queue for approach b.

The average "time in system" (in seconds) for vehicles on approach a, w_a, (in this instance, the stopped delay) is given by

$$w_a = L_a/x_a \quad (2)$$

Again, substituting b for a yields the stopped delay for approach b.

TABLE 1 CAPACITY OF A TWO-LANE BY TWO-LANE, FOUR-WAY STOP SIGN INTERSECTION (FROM *HIGHWAY CAPACITY MANUAL*)

Demand Split	Capacity (vph)
50/50	1900
55/45	1800
60/40	1700
65/35	1600
70/30	1500

TABLE 2 CAPACITY OF A TWO-LANE BY TWO-LANE, FOUR-WAY STOP SIGN INTERSECTION AS PREDICTED BY RICHARDSON

Demand Split	Capacity (vph)
50/50	1900
55/45	1760
60/40	1650
65/35	1600
70/30	1560
80/20	1520
90/10	1570
100/0	1800

M. Zion and G. F. List, Department of Civil Engineering, Rensselaer Polytechnic Institute, Troy, N.Y. 12180–3590. C. Manning, Roger Creighton Associates, Inc., Delmar, N.Y. 12054.

Moreover, the service time (in seconds, and in a queuing sense), s_a, for approach a is given by the expected value of the length of time the vehicle will have to wait before crossing the intersection:

$$s_a = t_m P_b(0) + T_c P_b(1) \quad (3)$$

where

t_m = minimum possible headway—the service time on approach a when there is no vehicle waiting to cross the intersection at the stopline on approach b,
T_c = impeded service time—the service time for a vehicle on approach a when a vehicle is already waiting to cross the intersection at the stopline on approach b,
$P_b(0)$ = probability of no vehicle being at the stopline on approach b, and
$P_b(1)$ = probability that there is a vehicle at the stopline on approach b.

For Equation 3, Richardson assumes that t_m is 4.0 sec for all approaches and that T_c is given by $2 * t_c$ where t_c, the clearance time on a single conflicting approach, is given by 3.6 sec plus 0.1 sec times the number of crossflow lanes that must be crossed (two in the case of a two-lane by two-lane intersection, resulting in a value of 3.8 for t_c and 7.6 for T_c).

Probabilities $P_b(0)$ and $P_b(1)$ depend on the utilization ratio, u_b, for approach b:

$$P_b(0) = 1 - u_b$$

$$P_b(1) = u_b \quad (4)$$

Inserting Equations 4 into Equation 3 gives the average service time (sec/veh) for approach a:

$$s_a = t_m(1 - u_b) + T_c u_b \quad (5)$$

Similarly, the average service time (sec/veh) for approach b is given by

$$s_b = t_m(1 - u_a) + T_c u_a \quad (6)$$

These equations (with four unknowns, s_a, s_b, u_a, and u_b) can be solved by substituting $x_a s_a$ for u_a in Equation 6 (in general $u = xs$) and then substituting $x_b s_b$ for u_b in Equation 5—where the expression used for s_b is obtained from Equation 6. This new Equation 5, when solved for s_a, yields the following:

$$s_a = \frac{x_a t_m T_c + t_m - x_b (t_m)^2}{1 - x_a x_b[(T_c)^2 - 2t_m T_c + (t_m)^2]} \quad (7)$$

Replacing b for a, and vice versa, throughout, yields s_b instead.

Once s_a has been found, $V(s_a)$, the variance of s_a, can be determined as follows:

$$V(s_a) = (t_m)^2(T_c - s_a)/(T_c - t_m) + (T_c)^2(s_a - t_m)/(T_c - t_m) - (s_a)^2 \quad (8)$$

The value for $V(s_a)$ and the value for s_a from Equation 7 can then be used in Equations 1 and 2 to determine L_a (the average length of queue on approach a) and w_a (the average time in system for approach a—here, the average stopped delay. Similar logic applies for finding L_b and w_b. Figure 1 demonstrates a sample calculation.

The major phenomena predicted by this model are the following:

- Delay increases at a greater-than-linear rate as volume increases.
- For a given total volume, the level of delay decreases as the flows become balanced. That is, an intersection with a 50/50 split in volume has lower delays than one with an uneven split.
- The foregoing phenomenon, conversely put, implies that intersection capacity (the point at which stopped delay reaches infinity) increases as the volume split approaches 50/50, as Hebert found. (A comparison of Tables 1 and 2 show, however, that the capacities are slightly different.)

A second model, developed by Chan et al. (*6*), also predicts delays for FWSC intersections. In this instance, a multivariable log-linear regression equation was fitted to field data and augmented by computer simulation to determine the predictive equation. The field data were collected at five intersections located in Idaho and Washington State.

The average delay/veh (sec/veh), D, is given by

$$D = a * \exp[(bS + cH + dT + e)V] \quad (9)$$

where

V = total arrival volume (veh/hr),
S = volume split factor (percent),
H = street-width factor (percent),
T = turning movement factor (percent), and
a, b, c, d, and e = calibration constants.

The volume split factor, S, is found by

$$S = s_1 - s_2 \quad (10)$$

where s_1 is the percentage of traffic on the major flow (percent) and s_2, the minor flow (percent).

The street width factor H (nondimensional) is given by

$$H = \max[0, (h - 30)/h] \quad (11)$$

where h is the road width (ft) and, in effect, 30 ft is the width (for a two-lane street) and no width adjustment is necessary.

The turning factor T (nondimensional) is given by

$$T = LTV/TV \quad (12)$$

where LTV is the left-turn volume and TV is the total volume for all four approaches. Sample calculations using these formulas are shown in Figure 2.

Three predictions given by this model are as follows:

- The average delay/veh rises rapidly as volume approaches capacity.

Given: Approach Vol. = 876 veh/hr
Vol. Split = 52.1/47.9
Balanced opposing volumes on both streets

Find: The time in system for approach a using the Richardson model:

$t_m = 4.0$ sec.

$t_c = 3.6$ sec. $+ 0.1(2) = 3.8$ sec.

$T_c = 2t = 2(3.8 \text{ sec.}) = 7.6$ sec.

$x_a = .521\ (876 \text{ veh/hr})/2 = 228.1 \text{ veh/hr} = 0.0634 \text{ veh/sec}$

$x_b = .479\ (876 \text{ veh/hr})/2 = 209.8 \text{ veh/hr} = 0.0583 \text{ veh/sec}$

$$s_a = \frac{x_b t_m T_c + x_b t_m^2}{1 - x_a x_b (T_c^2 - 2t_m T_c + t_m^2)} \quad (7)$$

$$= \frac{(0.0583)\ (4.0)\ (7.6) + (4.0) - (0.0583)\ (4.0)^2}{1 - (0.0583)\ (0.0634)\ ((7.6)^2 - 2(4.0)(7.6) + (4.0)^2)}$$

$$= 5.083 \text{ sec.}$$

$$V(s_a) = t_m^2(T_c - s_a)/(T_c - t_m) + T_c^2(s_a - t_m)/(T_c - t_m) - s_a^2$$

$$= (4.0)^2(7.6 - 5.083)/(7.6 - 4.0) + (7.6)^2(5.083 - 4.0)/(7.6 - 4.0) - (5.083)^2$$

$$= 2.725 \text{ sec} \quad (8)$$

$$u_a = s_a x_a = (5.083)(0.0634) = 0.322$$

$$L_a = \frac{2u_a - u_a^2 + x_a^2 V(s_a)}{2(1 - u_a)} \quad (1)$$

$$= \frac{2(0.322) - (0.322)^2 + (0.0634)^2(2.725)}{2(1 - 0.322)}$$

$$= 0.407$$

$$w_a = L_a/x_a = 0.407/0.0634 = 6.418 \text{ sec.} \quad (2)$$

time in system for approach a = 6.418 sec.

FIGURE 1 Sample calculation using the Richardson model (note: equation numbers refer to text).

- The capacity of a FWSC intersection decreases as its volume split approaches 50/50.
- Intersection capacities for the sites Chan studied ranged between 600 and 1,400 veh/hr (see Table 3).

DATA COLLECTION AND ANALYSIS

The data collected for this study were obtained from three intersections near Albany, N.Y. Three criteria were applied in selecting these sites: First, there had to be two-lane streets on all four approaches to ensure compatibility with the intersections used in the studies and models cited above; second, the sites needed a variety of volume splits so that variations in that parameter could be considered; and, last, the sites' volumes had to be high enough that delays of significance would be observed. The intersections selected were—

- Site 1: Moe Road and Clifton Park Center Road, Clifton Park, New York (this intersection has since been signalized);
- Site 2: an intersection internal to the Colonie Shopping Center, Colonie, New York; and
- Site 3: Pine and Lodge Streets, Albany, New York.

Following a methodology suggested by Box and Oppenlander (*7*), data were collected at Site 1 for 115 min during

Given: Approach Vol. = 876 veh/hr
Vol. Split = 52.1/47.9
% Left Turns = 27.4
Street Width = 24 feet

Find: The delay time using the Chan model

$$S = (s_1 - s_2)/100 = (52.1 - 47.9)/100 = 0.042 \quad (11)$$

$$H = \max[0, (h-30)/h] = \max[0, (24-30)/24] = 0$$

$$T = (\%\ \text{left turns})/100 = .274 \quad (12)$$

$$V = 876\ \text{veh/hr}$$

$$d = .186 \exp((-.007455\ S + .01333\ T + .004037)\ V) \quad (10)$$
$$= .186 \exp((-.007455\ (.042) + .01333(.274) + .004037)(876))$$
$$= 119.1\ \text{sec}$$

delay time = 119.1 sec.

FIGURE 2 Sample calculation using the Chan model (note: equation numbers refer to text).

TABLE 3 CAPACITY OF A TWO-LANE BY TWO-LANE, FOUR-WAY STOP SIGN INTERSECTION AS PREDICTED BY CHAN

Demand Split	Capacity (vph)
50/50	1076
60/40	1489
70/30	2419

an afternoon peak; at Site 2 for 50 min during the heaviest shopping hour of an afternoon; and at Site 3 for 70 min during an afternoon rush hour.

At each site, and for each 5-min time period observed, one person counted traffic (turning movement counts by approach and total) and a second collected delay data. The observation time, 5 min, was chosen as it was long enough to provide sample sizes adequate for statistically reliable estimates of delay and it allowed an opportunity to see the effects of volume split changes as the traffic volumes fluctuated. The person monitoring delays used a stopwatch to determine how long it took vehicles to pass through the intersection. (Vehicles were selected so that approximately equal numbers of observations would be made on each approach.) The stopwatch was started when the vehicle began braking and it was stopped when the vehicle cleared the intersection. (These observations thus included more delay than predicted by either the Richardson or Chan models; these delay estimates therefore should be bounded from below by the model predictions of delay.)

The data collected were processed to determine, for each 5-min time period, the volume split, the percentage of left turns, and the average delay (based on the way the data were collected and on the vehicles observed). The resulting statistics are presented in Table 4.

ANALYSIS AND OBSERVATIONS

The first and simplest observation made is that measured delays clearly increase as volume increases, as shown in Figure 3. Moreover, Table 5, which summarizes the average delays by site, shows Site 2 with the highest volumes and greatest delays, and Site 3 with the lowest volumes and lowest delays.

Table 6 also shows delays increasing as the volume split becomes unbalanced, which again agrees with Hebert and Richardson.

A stepwise linear regression analysis of the data collected generates the following results:

$$D = 12.361V \text{ (no constant allowed)} \quad (13)$$
(adjusted $R^2 = 0.9168$)

$$D = 21.24V - 8464 \text{ (adjusted } R^2 = 0.5813) \quad (14)$$

$$D = 8.517V + 16.74L \text{ (no constant allowed)} \quad (15)$$
(adjusted $R^2 = 0.9281$)

$$D = 16.91V + 13.16L - 7212 \quad (16)$$
(adjusted $R^2 = 0.6210$)

where, for a given 5-min time period, D is an estimate of total vehicular delay (the average observed delay times total volume), V is the total volume passing through the intersection (vehicles), and L is the total number of vehicles making left turns.

The implications of these regression results are as follows:

- Including a constant in the predictive equation lowers the adjusted R^2 value considerably. Therefore, the constants should be omitted from the regression equations.
- Including the left-turn volume in the predictive relationship always increases the adjusted R^2, whether the constant

TABLE 4 FIELD DATA

Moe Road and Clifton Park Center Road:

Data entry*	Vol. (veh/hr)	Vol. Split	% Left Turns	Ave. Delay** (sec)
1	984	58.5/41.4	20.9	13.27
2	444	59.5/40.5	32.4	8.19
3	672	55.4/44.6	23.2	10.25
4	960	50.0/50.0	30.0	13.97
5	828	60.9/39.1	14.5	8.31
6	684	57.9/42.1	31.6	12.92
7	876	52.1/47.9	27.4	10.32
8	960	60.0/40.0	25.0	10.91
9	900	56.0/44.0	30.7	13.93
10	900	57.3/42.7	30.7	8.89
11	912	59.2/40.8	27.6	9.52
12	888	50.0/50.0	27.0	12.14
13	948	53.2/46.8	22.8	14.17
14	1140	60.0/40.0	31.6	15.50
15	1188	51.5/48.5	27.3	13.30
16	864	59.7/40.3	27.8	14.47
17	1128	60.6/39.4	24.5	12.17
18	1032	52.3/47.7	24.4	17.00
19	1188	50.5/49.5	25.3	15.37
20	984	51.2/48.8	28.0	8.81
21	996	56.6/43.4	28.9	11.32
22	804	49.3/50.7	31.3	9.55
23	708	52.5/47.5	25.4	14.22

Colonie Center:

Data entry*	Vol. (veh/hr)	Vol. Split	% Left Turns	Ave. Delay** (sec)
24	1104	62.0/38.0	23.9	10.86
25	828	70.0/30.0	24.6	8.38
26	1104	70.0/30.0	31.5	26.30
27	1068	67.4/32.6	25.8	12.59
28	792	62.1/37.9	31.8	14.76
29	1044	57.9/42.5	33.3	12.31
30	1008	71.4/28.6	27.4	9.62
31	1044	64.4/35.6	29.9	17.22
32	1152	58.3/41.7	35.4	13.64
33	972	67.9/32.1	30.9	10.43

Pine Street and Lodge Street:

Data entry*	Vol. (veh/hr)	Vol. Split	% Left Turns	Ave. Delay** (sec)
34	888	71.6/28.4	13.5	9.12
35	816	57.4/42.6	14.7	8.32
36	924	76.6/23.4	7.8	7.39
37	1020	62.4/37.6	10.6	8.29
38	1152	56.2/43.8	15.6	16.37
39	1116	53.8/46.2	8.6	14.51
40	660	50.9/49.1	9.1	7.04
41	696	75.9/24.1	5.2	5.83
42	732	82.0/18.0	4.9	6.72
43	864	56.9/43.1	8.3	11.84
44	1056	67.0/33.0	11.4	13.27
45	888	63.5/36.5	9.5	9.49
46	888	60.8/39.2	4.1	8.73
47	408	52.9/47.1	5.9	6.61

* Each entry contains many readings taken over a 5 min. period.

** Ave. of all delay readings during 5 min. period.

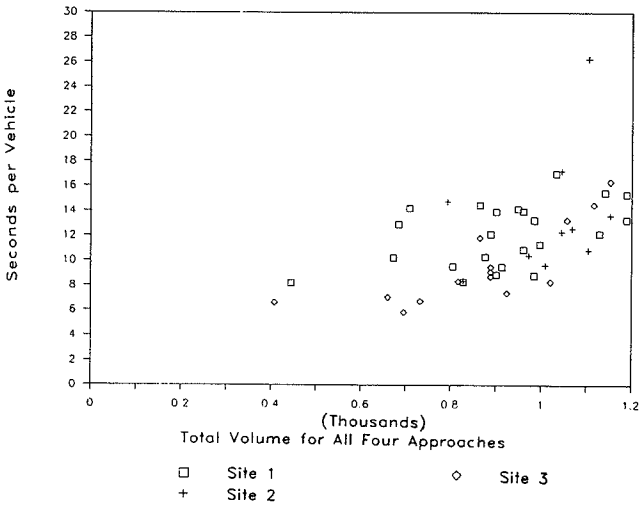


FIGURE 3 Delay versus volume for the three intersections studied.

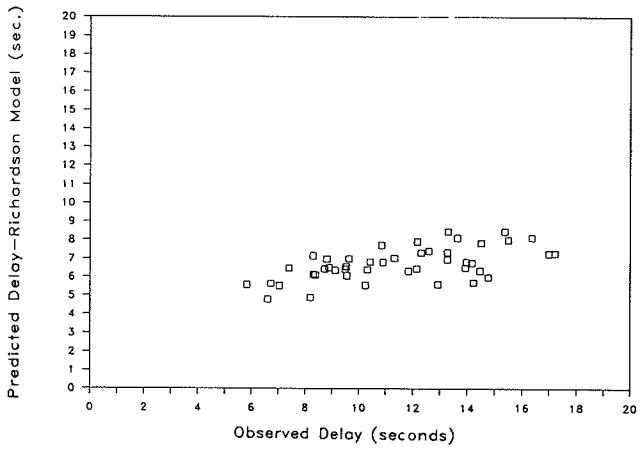


FIGURE 4 Correlation between the Richardson model and the delays observed.

TABLE 5 AVERAGE VOLUMES AND AVERAGE DELAYS

Site	Intersection	Average Volume	Percent Lefts	Average Delay
1	Moe and Clifton Center Roads	913	26.9%	12.11 sec
2	Colonie Center	1012	29.5%	13.61 sec
3	Pine and Lodge Streets	865	9.2%	9.54 sec

TABLE 6 DELAY TIMES FOR VARIOUS DEMAND SPLITS AS MEASURED IN FIELD (APPROACH VOL. = 900)

Demand Split	Delay Time
50/50	11.10
55/45	11.27
60/40	11.70
65/45	12.49
70/30	12.96

is included or not. This suggests that left turns do indeed have an effect on intersection delays.

- The best relationship among those tested involves the total volume and the left-turn volume.

The coefficients of Equation 15, in particular, imply that each vehicle passing through the intersection adds, on average, 8.517 sec to the total intersection delay (including deceleration time and the time required to cross the intersection) and each left turn makes the total intersection delay increase by an additional 16.74 sec.

The data can also be compared with the delay predictions given by the Richardson and Chan models, as shown in Figures 4 through 7. In Figure 4, the correlation between the observed delays and the Richardson delay estimates seems clear even though the correlation coefficient is low (adjusted $R^2 = 0.3232$), while in Figure 5 no such correlation is apparent (the adjusted R^2 is 0.0693). In Figures 6 and 7, the delay estimates and the observed delays from the models are plotted against volume. It is clear that the Richardson model does indeed fit the data observed in that it provides a lower bound on all of the observations. (This is because the field data include stopped delay plus deceleration time and the time required to cross the intersection.) In fact, it appears that the additional delay buried in the field observations is somewhere between 0 and 8 sec. The Chan model, as Figure 5 shows, does not provide such a bound, at least for the three intersections examined in this study.

Further analysis of the data, breaking the observations down into groups based on their volume splits, reveals additional insights. To conduct this analysis, each data point shown in Table 4 is placed in one of seven classes, based on the split category to which it most closely corresponds: 50/50, 55/45, 60/40, 65/35, 70/30, 75/25, and 80/20. For three of these classes, sufficient data are available to meaningfully plot the observed delays against the delay estimates from the Richardson and Chan models. As shown in Figures 8, 9, and 10, the Richardson model provides a consistent lower bound for the delays observed while the Chan model does not. In addition, the delays predicted by the Chan model increase more quickly than do those observed.

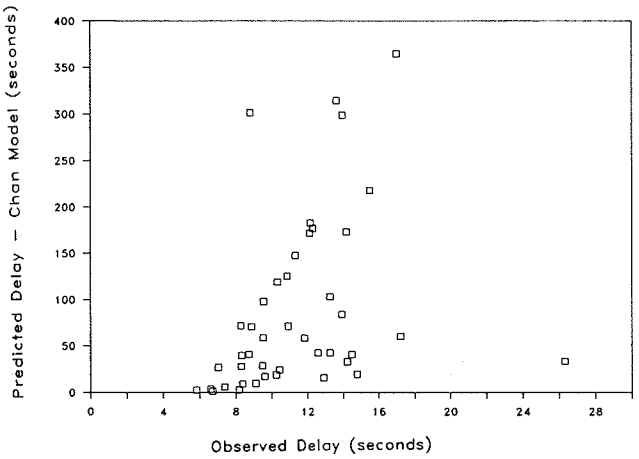


FIGURE 5 Correlation between the Chan model and the delays observed.

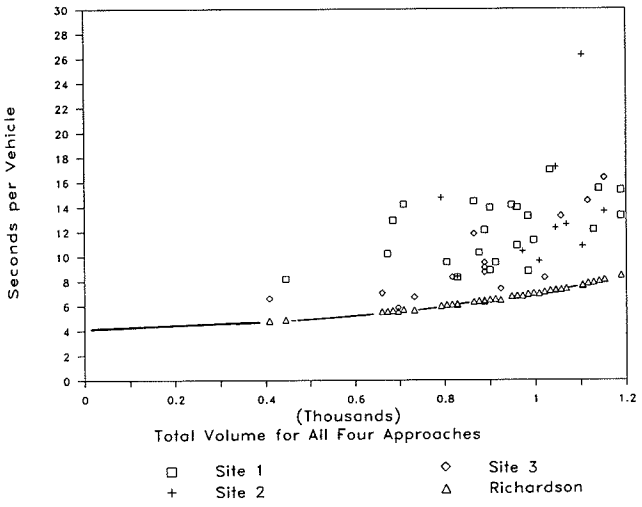


FIGURE 6 The Richardson model and the delays observed plotted against total intersection volume.

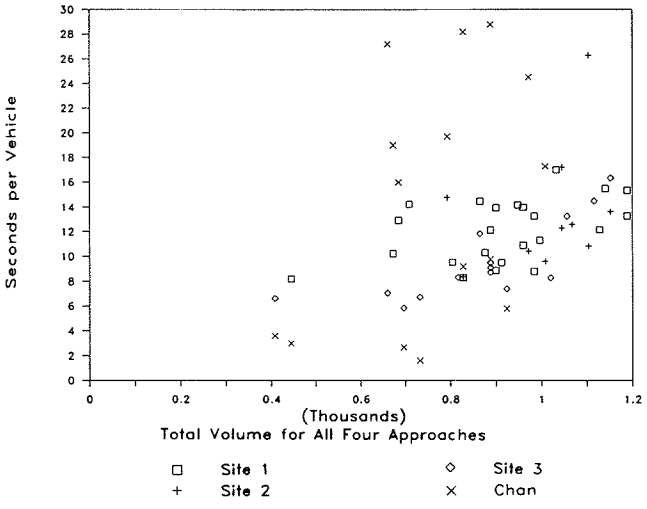


FIGURE 7 The Chan model and the delays observed plotted against total intersection volume.

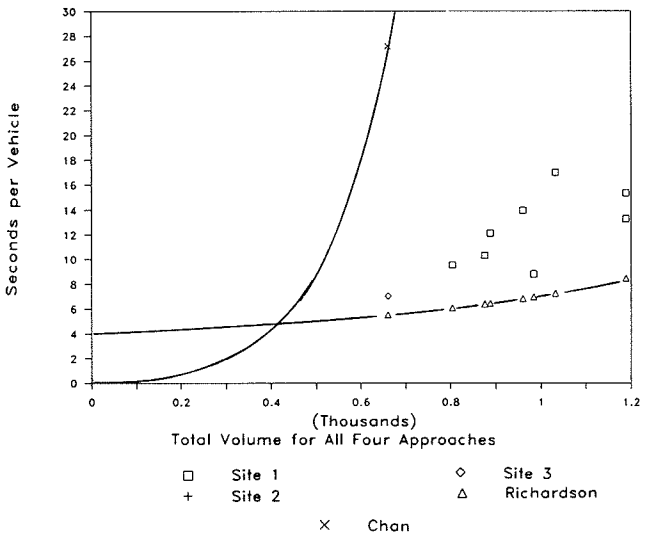


FIGURE 8 Delay versus volume for the observations with a 50/50 split in approach volumes.

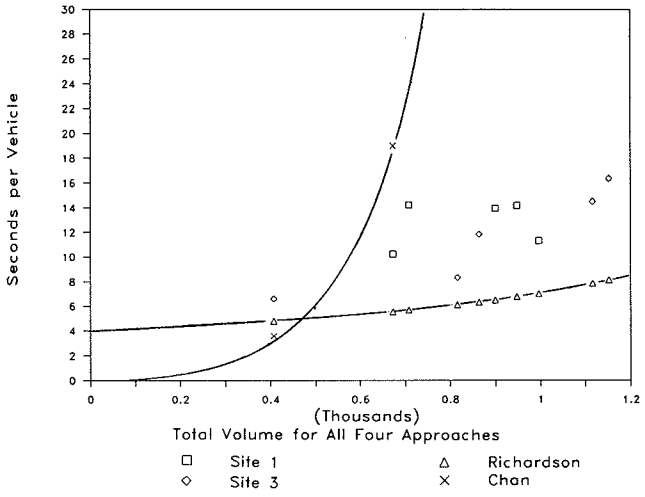


FIGURE 9 Delay versus volume for the observations with a 55/45 split in approach volumes.

CONCLUSION

Clearly, it should not be assumed that the data and analyses presented here can be used to defend definitive statements about the nature of delays at FWSC intersections. What the data do suggest, however, is that the following trends should hold in further analyses:

- Average delays should decrease as volume splits near a balanced 50/50 split.
- The percent of left-turning traffic should affect the delays observed.
- The Richardson model may provide a credible estimate of stopped delay.

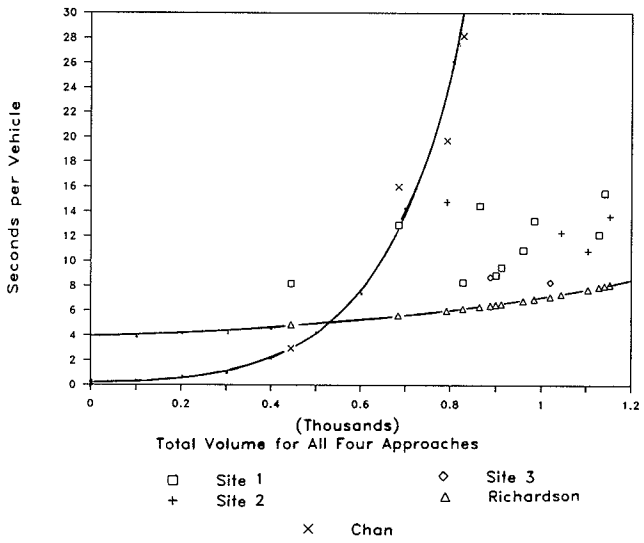


FIGURE 10 Delay versus volume for the observations with a 60/40 split in approach volumes.

The more extensive studies currently under way in support of revisions to Chapter 11 of the *Highway Capacity Manual* (*8*) will expand upon the work presented here and provide information sufficient to determine whether the hypotheses presented above are indeed true.

REFERENCES

1. J. Hebert. A Study of Four-Way Stop Intersection Capacities. In *Highway Research Record 27*, HRB, National Research Council, Washington, D.C., 1963, pp. 130–147.
2. Hannson et al. Capacity of Unsignalized Intersections—Swedish Capacity Manual. *Transportation Research Record 667*, TRB, National Research Council, Washington, D.C., 1978.
3. *Capacity of At-Grade Junctions*. Organization for Economic Cooperation and Development, Paris, France (1974).
4. A. Bakare and P. Jovanis. Analysis of Unsignalized Intersection Capacity. *Transportation Research Record 971*, TRB, National Research Council, Washington, D.C., 1984.
5. A. J. Richardson. A Delay Model for Multiway Stop-Sign Intersections. In *Transportation Research Record 1112*, TRB, National Research Council, Washington, D.C., 1987.
6. Y. Chan, L. J. Flynn, and K. J. Stocker. Volume-Delay Relationship at Four-Way-Stop Controlled Intersections: A Response-Surface Model. *ITE Journal*, Mar. 1989, pp. 27–34.
7. P. C. Box and J. C. Oppenlander. *Manual of Traffic Engineering Studies*. 4th ed. Institute of Transportation Engineers, Arlington, Va. 1976, pp. 106–112.
8. *Special Report 209: Highway Capacity Manual*, TRB, National Research Council, Washington, D.C., 1985.

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

Application of Simulation To Evaluate the Operation of Major Freeway Weaving Sections

A. SKABARDONIS, M. CASSIDY, A. D. MAY, AND S. COHEN

This paper describes the findings from the application of the INTRAS microscopic simulation model to evaluate the traffic performance at major freeway weaving sections. The work performed is part of an ongoing research project to develop improved weaving analysis procedures that are particularly applicable to California conditions. The INTRAS model was modified to predict the speeds of weaving and nonweaving vehicles and applied on eight major freeway weaving sections for a range of traffic conditions at each site. Good agreement was obtained between the measured and predicted values. Comparisons with speeds estimated from existing analytical procedures indicated that INTRAS predictions are considerably closer to the field measurements. The potential of the model to predict the capacity and level of service at weaving areas was also investigated. The model produced consistent results on the data sets tested, indicating that it may be used in conjunction with field measurements to develop improved methodologies for the design and analysis of freeway weaving sections. Future steps in this direction are discussed.

Weaving is defined as the crossing of two or more traffic streams traveling in the same general direction, along a significant length of the roadway, without the aid of traffic control devices (*1*). Weaving sections are common design elements on freeway facilities, such as near ramps and freeway-to-freeway connectors. The operation of freeway weaving areas is characterized by intense lane changing maneuvers and influenced by several geometric and traffic characteristics. Because of the complexity of vehicle interactions, operational problems may occur at weaving areas even when traffic volumes are less than capacity.

As a result of continuous traffic growth, major efforts are currently under way to design new interchanges or improve existing ones. Therefore, accurate procedures are needed to assess the operation of existing facilities, the effectiveness of alternative designs, and other operational improvements. However, a recent evaluation of the existing weaving analysis techniques (*2*) found significant discrepancies between the performance measures estimated from the existing methods and the field measurements, indicating that additional research is needed. Further research on this topic was also recognized as the second highest priority from 29 research problem statements recently published by the TRB Committee on Highway Capacity and Quality of Service (*3*).

A. Skabardonis, M. Cassidy, and A. D. May, Institute of Transportation Studies, University of California, Berkeley, Calif. 94720. S. Cohen, Federal Highway Administration, 6300 Georgetown Pike, McLean, Va. 22101.

All of the existing weaving design and analysis procedures are based on empirical data collected at a number of weaving sections nationwide. Simulation is an alternative approach for evaluating the operation of weaving areas, and a simulation model is most suitable to test the effectiveness of alternative designs and traffic management schemes before their field implementation. Simulation results could also be used in the development of design and analysis procedures, provided the model employed is capable of replicating known field conditions.

The purpose of this paper is to present the findings from the application of the INTRAS microscopic simulation model on eight major freeway weaving sections and the comparison of the model's results with field measurements. The objectives of the simulation experiments were

- To assess if simulation can predict the operation of weaving areas with reasonable accuracy, and
- To investigate the potential of simulation to augment field data in developing improved methods for the design and analysis of weaving sections.

The paper first gives an overview of the ongoing research on freeway weaving sections in California. The INTRAS model is briefly described, along with the modifications and enhancements performed to the model for this study. The application of the model on the selected test sites, and the analysis of the results, are presented. Additional model applications and results are discussed. The final section summarizes the major findings from the study and discusses future research directions.

OVERVIEW OF CURRENT RESEARCH

The research reported in this paper is part of an ongoing project to develop improved weaving design/analysis methods that are particularly applicable to California conditions (*2*; see paper by Cassidy et al. in this Record). The work focuses on large, complex weaving sections near or at freeway-to-freeway interchanges. The objectives of the project are

- To evaluate the existing methods for design and analysis of weaving sections considering their methodology, accuracy, and ease of application using field data from major weaving sites, and

- To develop an improved weaving design and analysis procedure as needed.

Major tasks performed include (a) a detailed literature search on the topic, (b) collection of a large amount of data at several major freeway weaving areas throughout the state, (c) evaluation of existing weaving analysis procedures, and (d) statistical analysis of the field data to better understand the operation of weaving areas and to develop empirical prediction models. Currently, an improved weaving design/analysis procedure is being developed based on field data supplemented by simulation modeling.

A parallel study on weaving operations is also under way in California (*4*). This study is concerned with "simple" or "ramp" weaving sections (one-lane onramp and one-lane right-side offramp with a continuous auxiliary lane). Field data have been collected at several sites using video recording. Current findings indicate that both the 1985 *Highway Capacity Manual* (HCM-85) method (*1*) and the Leisch method (*5*) underestimate the speeds of the weaving vehicles, and the Level D method (*6*) reasonably predicts the distribution of traffic flows in the weaving area.

SIMULATION OF WEAVING AREAS

Selection of the Simulation Model

A number of simulation models have been developed over the past 20 years to analyze traffic operations on freeways (*7*). These models generally fall into two major categories:

1. Macroscopic models, either static or dynamic, consider the average traffic stream characteristics (flow, speed, density), incorporate analytical procedures to evaluate existing conditions, and predict performance under different design and control scenarios. The FREQ and FREFLO models (*7*) are examples of such macroscopic models, which are commonly used for freeway corridors. These models handle the effects of weaving using analytical procedures, such as those in the 1965 *Highway Capacity Manual* (HCM-65) (*8*).
2. Microscopic models consider the characteristics of each individual vehicle, and its interactions with other vehicles in the traffic stream. Therefore, they can simulate traffic operations in much greater detail than the macroscopic models, but they usually require additional data, staff time, and computer resources for their application.

For this study, a microscopic simulation model was chosen because it can model the complexity of the vehicle interactions at weaving areas. Of the microscopic models identified, the INTRAS model (*9,10*) was selected because it is the most detailed and well-documented freeway simulation model publicly available.

Description of the INTRAS Model

INTRAS (Integrated Traffic Simulation) is a microscopic model that simulates the movement of each individual vehicle on the freeway and surface street network, based on car-following, lane-changing, and queue-discharge algorithms. The model was originally developed for the FHWA in the late 1970s to assess the effectiveness of freeway control and management strategies (e.g., ramp metering and incident detection). The model was successfully used recently to generate design alternatives for improving the operation of existing weaving sections (*11*).

The model is operational on mainframe computers. Computer run times for typical weaving sections range from 15 to 60 sec of CPU time on an IBM3091 mainframe for a 30-min duration of simulation, depending on the length of the weaving section, the number of vehicles being processed, and the output options.

INTRAS requires that the network first be coded into links and nodes. Links represent unidirectional traffic streams with homogeneous traffic and geometric characteristics, and nodes indicate the locations where these characteristics change. Figure 1 illustrates the INTRAS coding of a typical weaving section in this study. An important factor in the development of the link/node diagram is the correct designation of lanes between the different links (lane alignment); this ensures that the design configuration of the weaving section is correctly interpreted by the model. Numerous preliminary computer runs were performed on each site to verify that the data had been coded correctly.

Input to the model consists of data on design characteristics for each link (length, number of lanes, location, and length of the acceleration and deceleration lanes), free-flow speeds, vehicle composition, traffic volumes (total and lane distributions), and percent of trucks for the freeway and ramps. Origin-destination (O-D) data can be input or computed by the program.

The output from the standard version of the model provides the total travel (veh-mi), average and total travel time, volume, density, average speed, number of lane changes, and average and total delay. These of effectiveness (MOEs) are provided for each link and for the total network at user specified time intervals during the simulation.

Model Modifications and Enhancements

Because the model output did not provide the speeds of the weaving and nonweaving vehicles, which are the MOEs most used to determine the level of service (LOS) at weaving areas, the program was modified to estimate the average travel times and speeds of vehicles for each origin and destination. Here, the freeway origins are the upstream end of the freeway and the onramps (Nodes 1 and 2 in Figure 1), while the freeway destinations are the offramps and the downstream end of the freeway (Nodes 3 and 4). The average speeds of weaving and nonweaving vehicles could be then estimated using the O-D specific information and the speeds and travel times in the weaving section. The calculation of speeds consisted of the following steps:

1. The average travel time was calculated for each movement in the weaving section (Link 2,3 in Figure 1). For example, the average travel time of the freeway-to-freeway vehicles is equal to the O-D travel time between Nodes 1 and 2, minus

the average travel times on freeway Links 1,2 and 3,4. Similarly, the average travel time of the onramp-to-freeway vehicles is equal to the travel time between Nodes 2 and 4, minus the average travel time on freeway Link 3,4.

2. The average travel time for the weaving and nonweaving vehicles was calculated using the movement-specific predicted volumes and travel times. For example, the average travel time of weaving vehicles is equal to the volume-weighted average of the freeway-to-ramp and ramp-to-freeway travel times estimated in Step 1.

3. The speeds of the weaving and nonweaving vehicles were calculated from the weaving section length and the average travel times estimated in Step 2.

This process was automated using spreadsheet microcomputer packages (Lotus, Quattro). The output from the INTRAS model was transferred directly from the mainframe to the microcomputer. Procedures were written for the spreadsheet programs to process the output and calculate the performance measures.

Modifications were also made to the logic to improve the ability of the model to simulate the ramp merging situation. The process of lane changing in general, and ramp merging in particular, can be described as a gap acceptance phenomenon. Consider Vehicle A, which desires to merge into a gap between Vehicles B and C (B currently is the leader of C). Vehicle A will accept the gap if the time headway between it and Vehicle B is greater than some critical value $g(A,B)$ and the time headway between itself and Vehicle C is greater than some critical value $g(A,C)$. However, the critical time headway values are not constant but are dependent on the following considerations:

- The critical time headways depend on the speeds of the two vehicles. For instance, Vehicle A will accept a smaller value of the critical headway if it is going slower than Vehicle B than it will if it is going faster than Vehicle B.
- The lane change takes place over a finite period of time. During this time period, the lane changer can adjust position with respect to the new leader by braking.
- The new follower may cooperate with the lane changer by braking to increase the size of the gap.

The INTRAS model combines the three considerations listed above into a measure called "risk." The lane change phenomenon can then be described as the acceptance or rejection of a critical value of risk, with respect to both the leader and follower in the new lane. The existing INTRAS model uses a constant value for the value of critical risk. Although the model works well in that vehicles complete their merge from an acceleration lane, it leads to a rather asymmetric distribution of successful merges versus the distance from the end of the acceleration lane. Thus, the model was modified by adopting logic developed for the FRESIM model, currently under development, which will be the successor to INTRAS. The modification consists of replacing the constant value of critical risk with a function that starts with a low value at the beginning of the acceleration lane and increases to a maximum value at the end of the acceleration lane. The rate of increase is the square root of the ratio between the distance of the vehicle from the end of the acceleration lane and the length of the acceleration lane. Preliminary results indicate that this process provides a good description of the merge process.

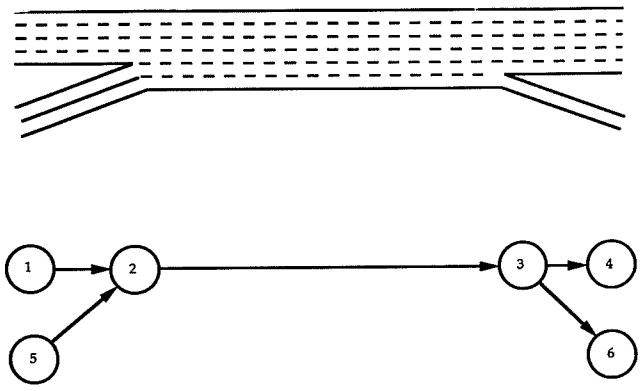


FIGURE 1 Sample coding for the INTRAS model.

MODEL APPLICATION

INTRAS was initially applied to several sample data sets to gain experience with the model and to test the sensitivity of input data and model parameters. The results from this initial application indicated that the model can simulate weaving operations with reasonable accuracy and can be used to predict traffic performance at existing freeway weaving sites.

The Data Base

Eight major freeway weaving sections were chosen for the application of the model. The selected sites represented a wide range of section configurations and design characteristics, such as length, number of lanes (N) in the weaving section, number of approaching freeway lanes, and number of lanes for the onramp and offramp) (see Figure 2). The configuration of each test site (A, B, or C) is given according to the definitions in the HCM. This classification is based on the minimum number of lane changes that must be made by weaving vehicles as they travel through the section. Type A weaving areas require that each weaving vehicle make one lane change to execute the weaving movement. Type B sections require vehicles in one weaving movement (onramp to freeway or freeway to offramp) to make one lane change while vehicles in the other weaving movement may accomplish their maneuver without changing lanes. Type C sections require vehicles in one weaving movement to make two or more lane changes while vehicles in the other weaving movement may accomplish their maneuver without changing lanes.

Information on traffic characteristics was collected using video recordings (see paper by Cassidy et al. in this Record). Six hours of operations were filmed on each site to obtain a range in traffic conditions. The data extracted from the tapes consisted of traffic volumes for each movement in the weaving area; the proportion of trucks, buses, and recreational vehicles; and the speeds of weaving and nonweaving vehicles. The data were extensively checked and verified for accuracy; data from 12 hr of operation were discarded due to congestion, incidents, and inclement weather during the videotaping. The

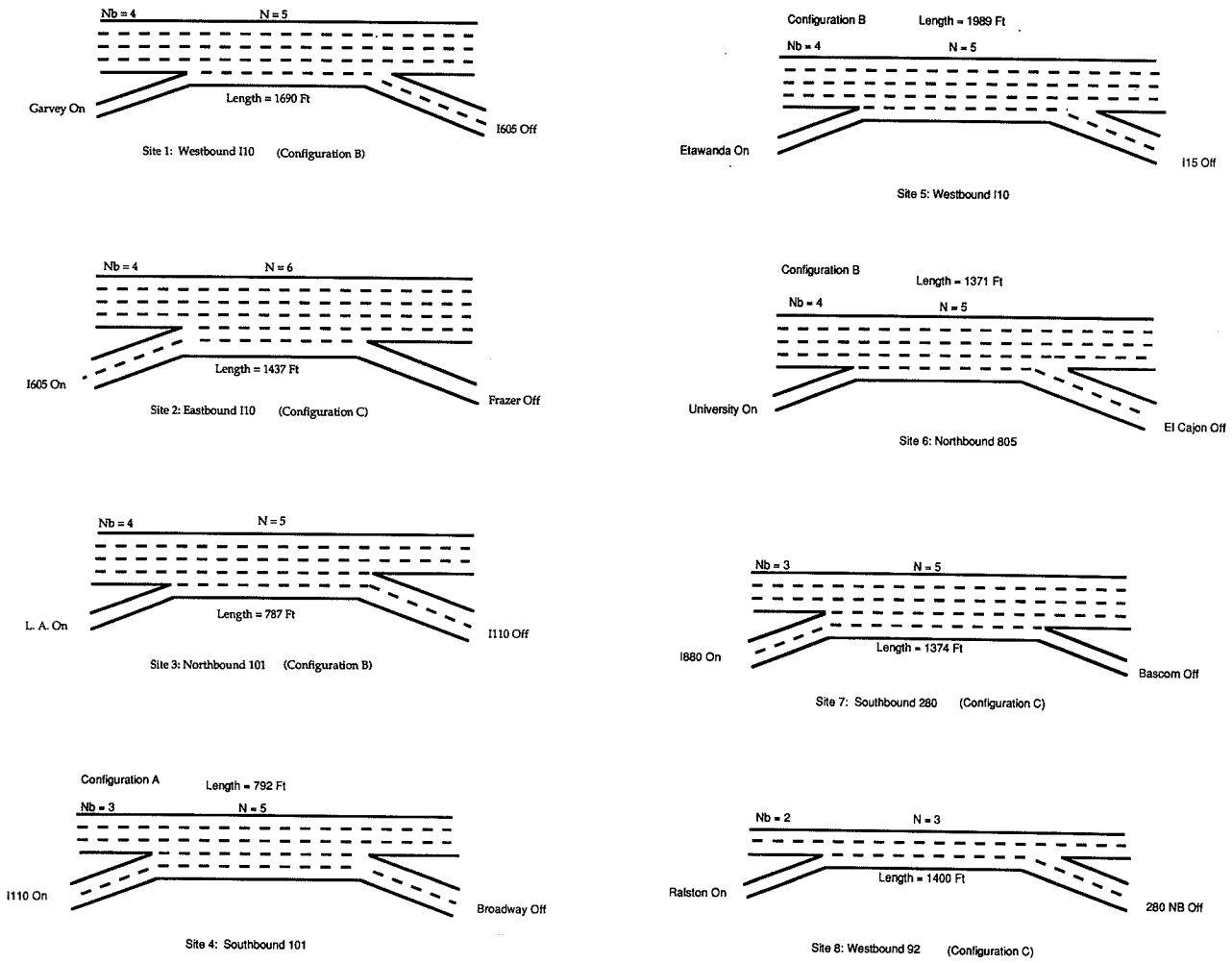


FIGURE 2 Selected test sites.

final data base consisted of 36 data points considering hourly volumes (in pce/hr), and represented a range of operating conditions at each test site.

Application to the Test Sites

The INTRAS model was applied to all the data sets using the field-collected data as input parameters. Regarding the application of INTRAS in this study:

- No adjustment of the internal model parameters was performed to get the best possible match of field measurements with the model predictions on each individual site. Therefore, the results represent the straightforward application of INTRAS and not the findings from calibrating the model on particular sites.
- INTRAS is a stochastic model, i.e., through random numbers, driver/vehicle characteristics are assigned. Therefore, the results from a simulation run may vary with the input random number seed for otherwise identical input data. Ideally, repeated simulations with different input random seeds should be made to gauge the effect of different random numbers. The results from tests performed on a number of data sets indicated that this variation in the predicted speeds of vehicles was between 1 and 2 percent.
- The results from the INTRAS model may also vary with the length of simulation time, especially for congested conditions. In this study, 30-min simulation runs were performed on all data sets. Intermediate cumulative outputs (every 10 min) were also printed, and the results were examined to test the stability of the simulation results. On most of the data sets, this variation was minimal (less than 1 percent), indicating that the model results are stable.

Analysis of the Results

Comparison with Field Measurements

Figure 3 shows a comparison of the measured and INTRAS-predicted speeds of all vehicles in the weaving area under different flow levels in different weaving sections. Each data point represents the average speeds of all vehicles for a 1-hr time period. The differences between the measured and INTRAS-predicted speeds were within 10 percent (approximately 5 to 6 mph) for the entire range of simulated operating conditions. Also, the mean percent difference was only about

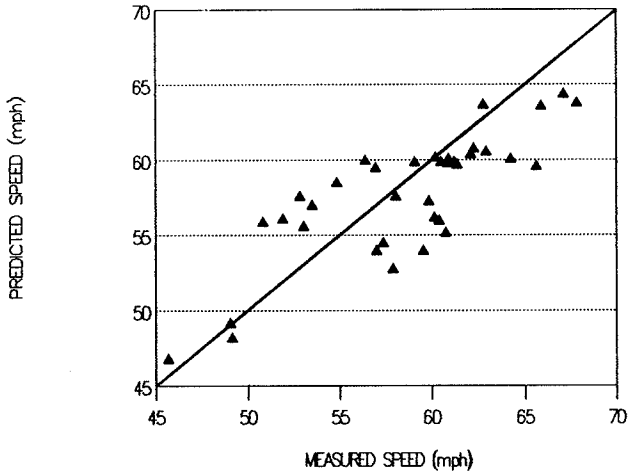


FIGURE 3 Average speeds, all vehicles—measured versus predicted.

1 percent, considering the absolute percent differences on each data set.

The largest differences between measured and INTRAS-predicted speeds occurred on test sites that had considerable variation in the measured average speeds for similar traffic volumes and patterns (see Figure 4). The field data exhibits a much larger scatter than the INTRAS-predicted values. This scatter would not likely be replicated in a simulation model, which uses car-following algorithms equivalent to fundamental speed-flow relationships and apparently does not include sufficient random variations.

The measured and INTRAS-predicted average speeds were compared separately for the weaving and nonweaving vehicles (see Figure 5). The estimated mean percent differences between measured and predicted speeds were very small (1 percent for the nonweaving vehicles and about 3 percent for the weaving vehicles). Again, the mean differences were calculated based on the absolute percent differences on each data set. These differences are insignificant, considering the stochastic nature of the INTRAS model.

The differences between measured and predicted values were within 10 percent in most of the data sets. Larger discrepancies were observed for a few data points, especially for the weaving vehicles. This occurred again at test sites for which the measured speeds were higher for higher volumes, contrary to what is normally expected. Such variabilities in the field data, also noted in the speed/volume scatter plots shown in Figure 4, could be due to the differences in driver behavior for different times of day (peak versus off-peak conditions). Comparisons were made between the predicted and measured speeds, considering only the peak period data (24 data points) (see Figure 6). Closer agreement was obtained between measured and predicted values (most of the differences were between 5 and 7 percent) for most of the data points, and the average differences were insignificant.

Comparison with Existing Analytical Procedures

The existing procedures for the design and analysis of freeway weaving sections generally fall into two major categories: those

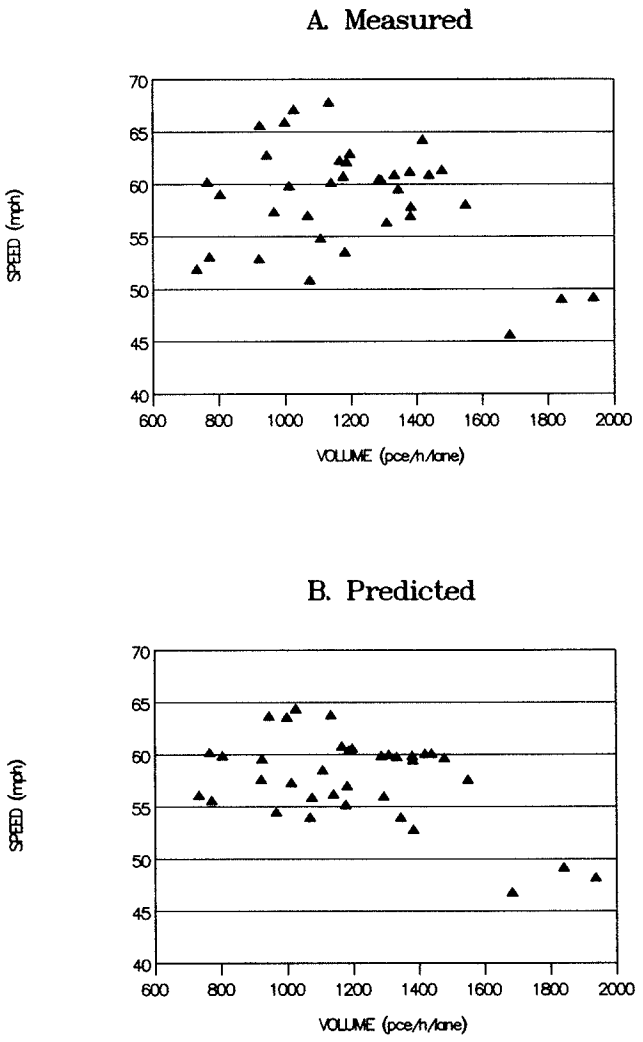


FIGURE 4 Speed/volume relationships.

based on traffic volumes (*5,8*), and those based on the speeds of vehicles in the weaving section (*1,12–14*). The volume-based approaches are generally design oriented, whereas the speed-based procedures are more suitable for operational analysis. All methods, except the HCM-65 technique, estimate the average speeds of weaving and nonweaving vehicles and determine the LOS of the weaving section based on those predicted speeds. All methods use the same basic geometric and traffic data (number of lanes and length of the weaving section, and volume for each movement) to estimate the average speed of vehicles. The basic difference in the existing methods is the way the weaving section configuration is considered to account for the number of lane changes (*2*).

All the existing analytical methods were applied to the same eight test sites used in the simulation experiments. The predicted and measured performance measures were then compared with field measurements and the INTRAS predictions.

Figure 7 shows the mean and standard deviation of the differences between the measured and predicted speeds for all the data sets. The results from the HCM-65 method are not shown since this method does not provide the speeds of weaving and nonweaving vehicles. The results shown in Figure 7 indicate that fairly large discrepancies exist between the

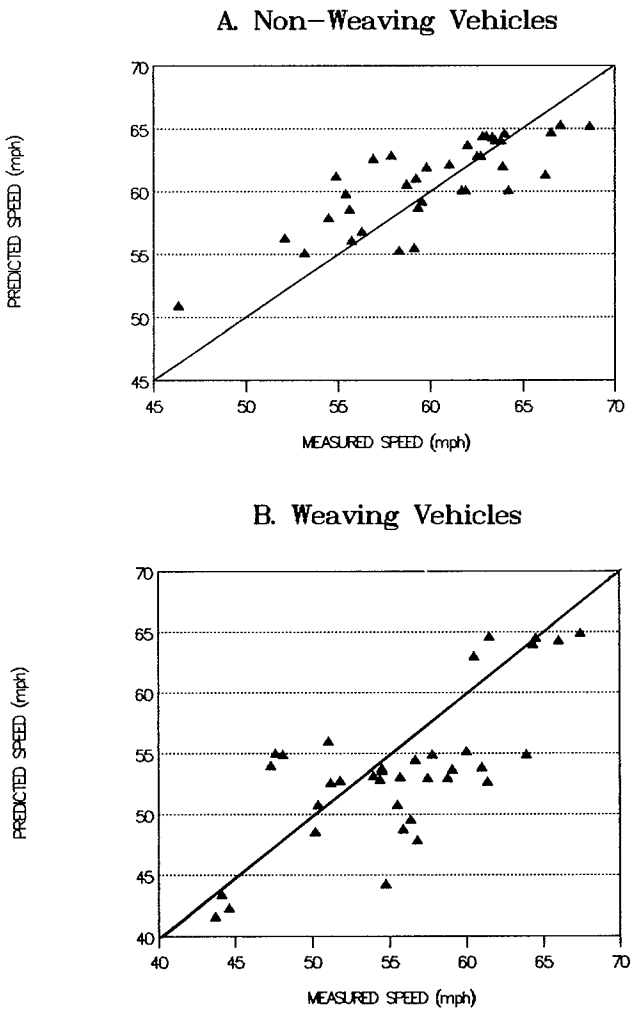


FIGURE 5 Measured versus predicted speeds.

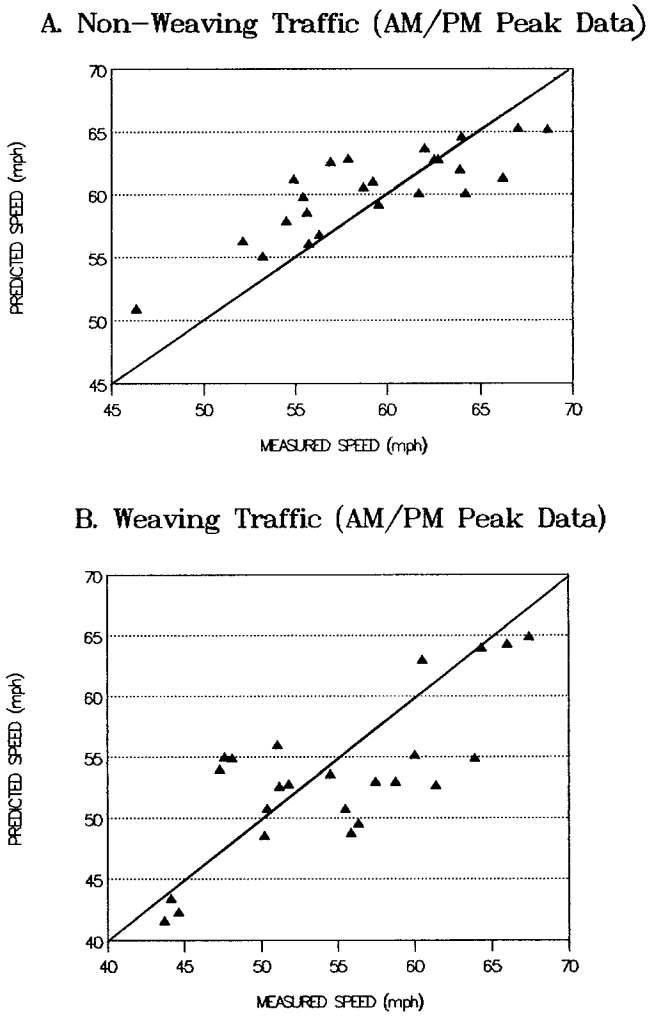


FIGURE 6 Measured versus predicted speeds (a.m./p.m. peak data).

measured and predicted values from the existing analytical procedures. INTRAS estimates, in contrast, are fairly close to the field data. The following comments could also be made from the application of existing methods:

- All methods underestimate the speeds in the weaving section. No consistent patterns were found in the differences between predicted and observed speeds. Large differences were noted between sites.
- Several of the existing methods have limits for certain geometric and traffic parameters (including section length, total or proportion of weaving traffic) that preclude their application on a number of sites with commonly occurring conditions.

PREDICTING WEAVING AREA OPERATIONS

The comparison of the INTRAS-predicted and field-measured performance measures indicates that simulation can reasonably replicate field conditions at weaving areas. Therefore, it can potentially be used to assist in the development of design and analysis procedures by predicting traffic performance under different geometric and traffic conditions.

The next step in this study was to investigate the potential of INTRAS to predict the "capacity" of a weaving section

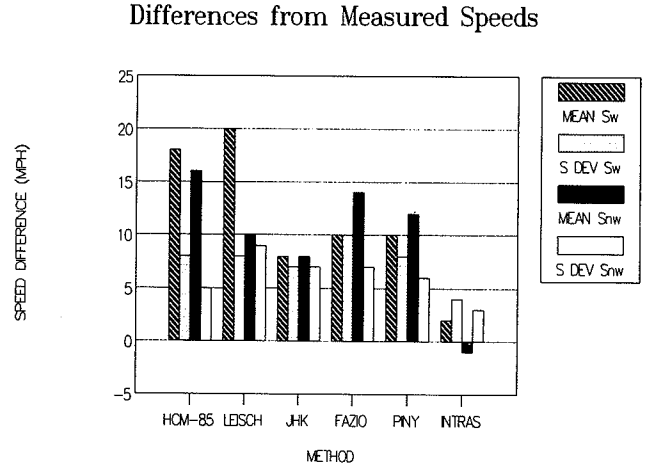


FIGURE 7 Comparison of methods.

(i.e., at which level of traffic volumes the traffic performance becomes unacceptable, considering the speeds of vehicles and other MOEs. It is extremely difficult to estimate the capacity of weaving sections due to the range of the configurations of major weaving sections, the proportion of weaving vehicles in the traffic stream, and several other factors. Therefore, such application of the model should be considered as explor-

ing the model's potential rather than developing capacity estimates.

Two test sites were selected for the simulation experiments: westbound I-10 in Los Angeles (WB10; Site 1) and northbound 805 in San Diego (NB805; Site 6). As shown in Figure 2, both sites have five lanes in the weaving section and have the same configuration (Type B sections according to the definition in HCM-85). The basic difference between those sites is the weaving ratio (*WR*), which is the ratio of the smaller weaving volume to the total weaving volume. On the NB805 test site, the weaving movements are balanced throughout all time periods ($WR = 0.45$), but they are unbalanced on the WB10 weaving section ($WR = 0.07$). In addition, the WB10 site has a longer length (1,690 ft) than the NB805 section (1,371 ft).

The simulation experiments were performed as follows:

- Input data on geometrics, free-flow speeds, the ratio of weaving and nonweaving traffic, the $\bar{W}R$, and the proportion of ramp traffic in the total volume were held constant for all simulation runs. The data on traffic characteristics were taken to equal the average values from the different traffic conditions at each test site.
- The total volume was allowed to vary on each run to yield different volume/capacity (v/c) ratios at the weaving section. A typical capacity of 2,000 pce/hr/lane was assumed to determine the input total volume for each run.

A total of 11 simulation runs were performed on each test site for different volumes corresponding to v/c values between 0.2 and 1.25. The model outputs were then analyzed to obtain the average speeds of vehicles, traffic volumes, density, and the number of lane changes at the weaving section.

The results for each test site are shown in Figure 8. The speeds plotted are the average speeds of all vehicles in the weaving section for a 1-hr time period. The INTRAS-predicted speed against v/c curves are very similar for both sites, despite the large differences in lane-changing activity between the two sites, and show that the capacity is higher than 2,000 pce/hr/lane. The model predictions indicate that a value of 2,200 to 2,300 pce/hr/lane might be used for capacity that corresponds to a v/c ratio between 1.1 and 1.15. Figure 8 also shows that the average speeds remain about the same, approximately 35 mph, for even higher v/c ratios. However, examination of the model outputs indicated the results were unstable, and the volumes predicted to pass through the weav-

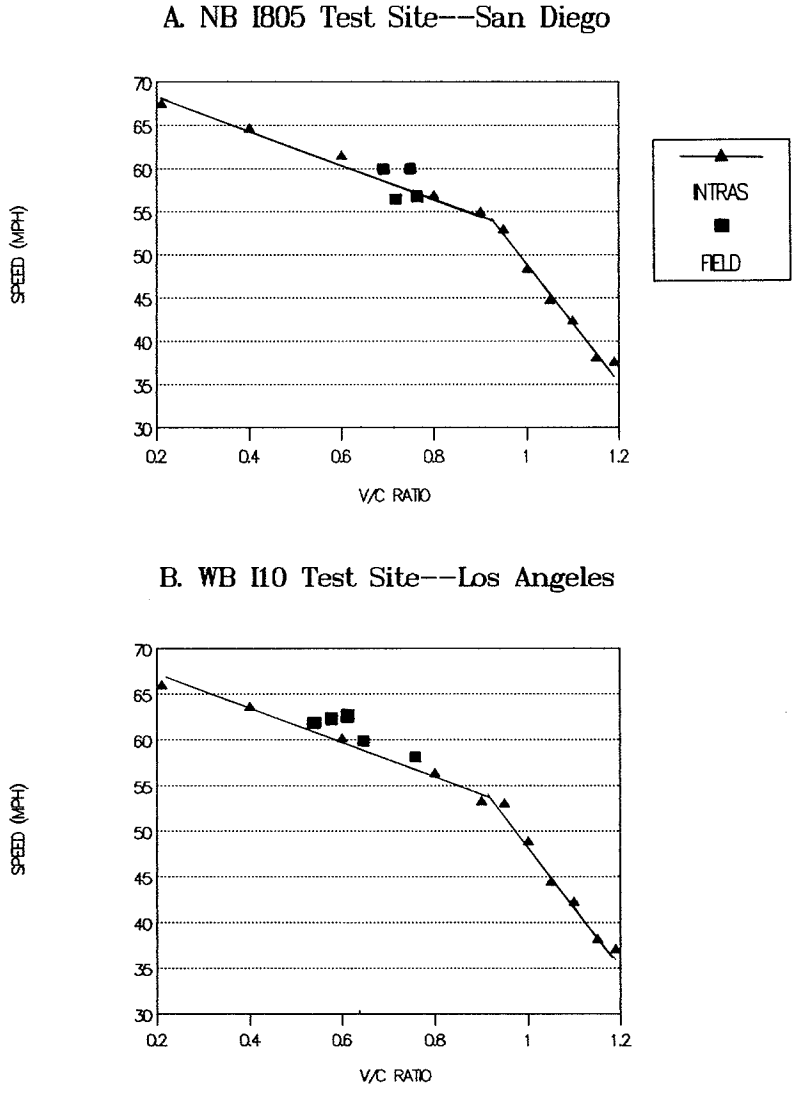


FIGURE 8 Relationship between speed and v/c ratio.

ing section were lower than the input values, indicating upstream congestion.

CONCLUSIONS

Summary of the Study Findings

The following significant findings were obtained from the application of the INTRAS microscopic simulation model, as modified for this study, on eight major freeway weaving sections:

- The INTRAS model reasonably replicated traffic operations on all eight weaving sites. In most of the data sets, the INTRAS-predicted average speeds of both weaving and nonweaving vehicles were within 10 percent of the field-measured values. Larger discrepancies (approximately 15 percent differences between measured and predicted values) were mostly due to the inherent variability in the field-collected data.
- The patterns of the simulation results were consistent for the entire range of traffic conditions on all sites. Good agreement between measured and predicted values was obtained for all combinations of design characteristics and demand patterns.
- The INTRAS-predicted performance measures are considerably closer to the field data than the estimated values from all existing analytical methods for the design/analysis of freeway weaving sections. In addition, the existing analytical procedures produced inconsistent results for several data sets, indicating that they may not be applicable for the entire range of commonly occurring field conditions.
- The results from the model application to predict the operation of the weaving areas for a range of traffic levels were very promising. The results were consistent for both sites, and the predicted capacities reflect recent measured values. Thus, the model can be used to predict performance for weaving sections when data are not available or are difficult to obtain.

Future Research

The findings reported in this paper, as well as other results from the ongoing research on major freeway weaving sections, clearly indicate that a better understanding of the operation of weaving areas and development of improved design and analysis procedures are high priority research needs. The simulation approach holds considerable potential for assisting in the development of such improved procedures.

The following steps are proposed for future research:

1. Operation of weaving sections near or at capacity conditions. Application of the INTRAS model and comparison of the results with field data from sections operating near or at capacity conditions will provide supplementary evidence on the ability of the model to predict performance accurately for a wide range of conditions.
2. Capacity versus weaving section configuration. The relationships between the length and configuration of the weaving section, lane changing, and upstream volume distribution are not well understood, and their implications for design and capacity analysis are not clear. Simulation will be used, in conjunction with field data, to explore these relationships and their effects in the operation of weaving areas.
3. Performance measures. The results from the INTRAS simulation runs, for the existing and additional test sites, will be further analyzed and compared with field data to determine if other performance measures (e.g., number of lane changes, density, and percent time delay) are more appropriate for design and operational analysis of weaving sections.

ACKNOWLEDGMENTS

This research study was sponsored by the California Department of Transportation (Caltrans). The authors would like to thank Fred Rooney and Howard Fong at Caltrans Headquarters for their cooperation and support throughout the study.

REFERENCES

1. *Special Report No. 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. A. Skabardonis, M. Cassidy, and A. D. May. Evaluation of Existing Methods for the Design and Analysis of Freeway Weaving Sections. Presented at 67th Annual Meeting of the Transportation Research Board, Washington, D.C., 1988.
3. *Transportation Research Circular 319: Research Problem Statement: Highway Capacity.* TRB, National Research Council, Washington, D.C., 1987.
4. H. K. Fong and F. D. Rooney. *Weaving Data.* Interim Reports 1-11. State of California, Department of Transportation, 1987/1988.
5. J. E. Leisch et al. Procedure for Analysis and Design of Weaving Sections. In *User Guide, Vol. 2.* Report FHWA/RD-85/083. FHWA, U.S. Department of Transportation, 1985.
6. K. Moskowitz and L. Newman. Notes on Freeway Capacity. In *Highway Research Record 27*, HRB, National Research Council, Washington, D.C., 1963, pp. 44–68.
7. A. D. May. Freeway Simulation Models Revisited. In *Transportation Research Record 1132*, TRB, National Research Council, Washington, D.C., 1987.
8. *Special Report 87: Highway Capacity Manual.* HRB, National Research Council, Washington, D.C., 1965, pp. 160–186.
9. D. A. Wicks and E. B. Liebermann. Development and Testing of INTRAS, A Microscopic Freeway Simulation Model. In *Vol. 1: Program Design, Parameter Calibration and Freeway Dynamics Component Development.* Final Report FHWA/RD-80/106. FHWA, U.S. Department of Transportation, 1980.
10. D. A. Wicks and B. J. Andrew. Development and Testing of INTRAS: A Microscopic Freeway Simulation Model. In *Vol. 2: User's Manual.* Final Report FHWA/RD-80/L07. FHWA, U.S. Department of Transportation, 1980.
11. S. L. Cohen and J. Clark. Analysis of Freeway Reconstruction Alternatives Using Traffic Simulation. In *Transportation Research Record 1132*, TRB, National Research Council, Washington, D.C., 1987.
12. J. Fazio and N. M. Rouphail. Freeway Weaving Section: Comparison and Refinement of Design Analysis Procedures. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986.
13. W. Reilly, J. H. Kell, and P. J. Johnson. *Weaving Analysis Procedures for the New Highway Capacity Manual.* JHK and Associates, 1984.
14. L. J. Pignataro et al. *NCHRP Report 159: Weaving Areas—Design and Analysis.* Polytechnic Institute of New York; TRB, National Research Council, Washington, D.C., 1975.

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

Three-Dimensional Relationships Among Traffic Flow Theory Variables

ROBERT S. GILCHRIST AND FRED L. HALL

This paper is an investigation of the relationships among speed, flow, and occupancy, representing the three variables of traditional theory for uninterrupted traffic flow. The variables were examined in three-dimensional space, rather than two at a time as has previously been the case. Scatter plots showing connected data points were positioned in space using a three-dimensional rectangular coordinate system. Oblique views of the data were projected as two-dimensional plots for presentation purposes. The resulting pictures were evaluated for points of agreement with traditional traffic flow theory and with a possible new approach based on the cusp catastrophe theory. The results suggest that conventional theory is insufficient to explain the data and that the plotted data are visually consistent with the catastrophe theory model of uninterrupted traffic flow.

Traditional traffic flow theory presents speed-flow-concentration relationships as shown in Figure 1. This diagram (*1*) portrays the theoretical relationship between two variables. This representation implies that the underlying data can be adequately represented as three line functions using two-dimensional relationships and that the underlying data also represent a line function in three dimensions. This function has been represented as a horseshoe shape, set at an angle to each of the three projections (*2*,p. 50).

This underlying three-dimensional representation has not recently been tested, although the data collection capabilities of current freeway traffic management systems (FTMSs) make this feasible now. In particular, several systems now collect speed as well as volume and occupancy data, making the use of three independent variables possible. Earlier systems could obtain speed data only through a calculation that relied on the relationship investigated in this research.

The purposes of this investigation were as follows:

- To investigate empirically the three-dimensional relationship among speed, flow, and occupancy;
- To determine whether speed, flow, and occupancy relationships are best described using two-dimensional functions, a surface, or some other approach; and
- To identify the points of agreement between observed data and traditional theory, and between the data and a recently proposed catastrophe theory representation of traffic flow.

The first section of this paper briefly describes the difficulty researchers have encountered in matching traffic flow theory with data. The second provides a description of the data and its location source, collection method, and preparation. Following that, projections of the data are presented and described. Next, there is a discussion of the relationship of the data to both traditional theory and catastrophe theory. Finally, a number of conclusions are presented.

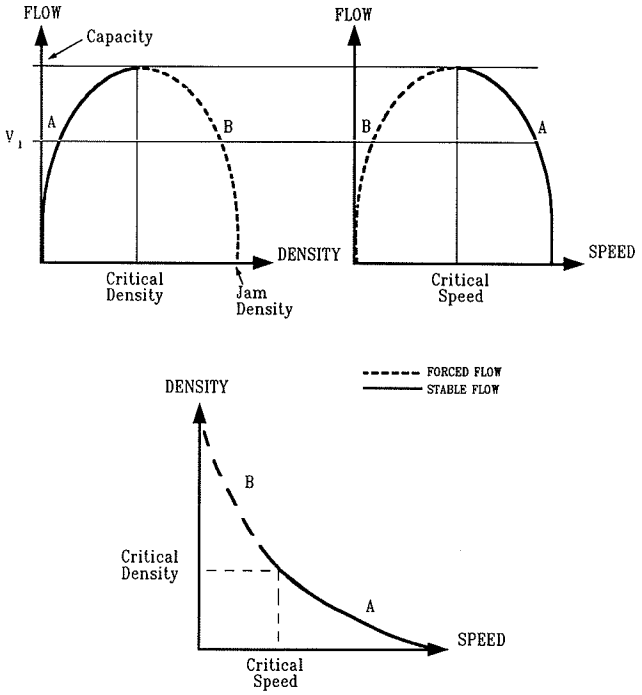


FIGURE 1 Relationships among speed, density, and rate of flow on uninterrupted flow facilities (*1*).

BACKGROUND

Improved data collection methods have often led to serious questioning of traditional theory. This has happened for uninterrupted traffic flow condition theories as well as other fields, such as particle physics. The problem is the inability of traditional theory to explain the data obtained with new equipment or techniques.

Three examples of this in the development of traffic flow theory represent this type of concern. In 1965, use of newly acquired data from Chicago freeways was made in an effort to calibrate the standard two-variable models (*3*). The results showed that the current theory did not match the data well. However, an adaptation of the conventional theory was found that allowed data and theory to coexist peacefully.

R. S. Gilchrist, RGP Transtech, Mississauga, Ontario, Canada L5N 1P7. F. L. Hall, Department of Civil Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4L7.

In another study (*4–6*), the data did not match the conventional theory well, but modifications were found to salvage the conventional theory. One such modification is the concept of two-regime models. Unfortunately, even with this modification, it was found that two sets of data collected at the same location required different sets of parameters. This remains a concern in the modified theory.

The third example comes from the 1985 *Highway Capacity Manual* (HCM) (*1*). Although not explicitly recognized as a problem caused by new data acquisition techniques, the HCM does contain a discussion of the mismatch between data and theory in its coverage of the speed-flow relationship. There is explicit mention of previous attempts to fit a curve to data and to test a variety of theoretical curves (*7*). There is also implicit recognition of this ongoing concern since one section discusses work with discontinuous, or two-regime, models while other sections ignore this and rely on representations (see Figure 1).

These approaches to matching data with theory have all attempted to model the data using some form of a curve in two-dimensional space, so three sets of curves must be developed to explain traffic flow theory fully. It has been suggested that speed-flow-concentration relationships are better explained using a three-dimensional surface defined by the cusp catastrophe theory (*8–10*). This model of traffic flow indicates that data points representing speed, flow, and concentration lie on a surface similar to a sheet of paper with a tear in one section (see Figure 2). The portion of the sheet to one side of the tear is raised relative to the portion on the other side. The raised portion represents higher speeds (uncongested data), while the lower portion depicts lower speeds (occurring during congestion). Transitions between uncongested and congested conditions can occur either by crossing the tear or going around the end of it. The former movement produces a sudden jump in one of the three variables at the same time that the other two undergo continuous change. If the transition occurs by traversing the portion of the surface that has no tear, all three variables undergo continuous change.

However, the work undertaken so far to test this theory has relied on two-dimensional projections of the information, even though the underlying theory is explicitly three-dimensional. Inspection of the three variables concurrently is as necessary to validate this new model as it is to provide a more complete test of conventional theory.

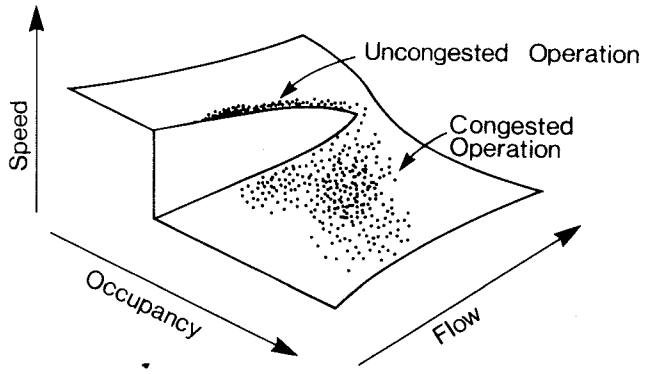


FIGURE 2 Hall's perception of catastrophe theory surface (*8–10*).

DATA ACQUISITION

The data for this analysis were obtained from the Mississauga FTMS on the Queen Elizabeth Way (QEW), just west of Toronto. The area of the QEW studied is a suburban freeway linking a sizable commuting population in Mississauga, Oakville, and points west with major employment centers in metropolitan Toronto. The six-lane freeway is situated in generally level terrain. The data presented in this paper are obtained from Station 16, on the east side of the Credit River, between the Mississauga Road entrance ramp and the exit ramp at Highway 10 (see Figure 3).

Eastbound travel into Toronto generates recurrent congestion between about 7:00 a.m. and 9:00 a.m. Interchange locations at Highway 10 and Cawthra Road contribute to congestion as traffic attempts to enter the main line from the ramps when the freeway is operating at or close to capacity. Ramp metering is in effect, but the ability of the metering process to serve mainline traffic is limited by the availability of queue space for the ramp traffic. Mainline queuing extends well west of Station 16.

Data used in this examination were taken from the FTMS computer tapes for Friday, September 30, 1987. All data are from the median, or left-most, side of the eastbound lanes. A single lane was chosen to reduce the amount of data to be examined, given that a large number of different views might be desired for analysis. The median lane was selected because it has been indicated (*10*) that median-lane traffic operations provide a sensitive and reliable indicator of freeway operations.

This station has paired detectors in each lane, from which speed, flow, and occupancy are transmitted to the central

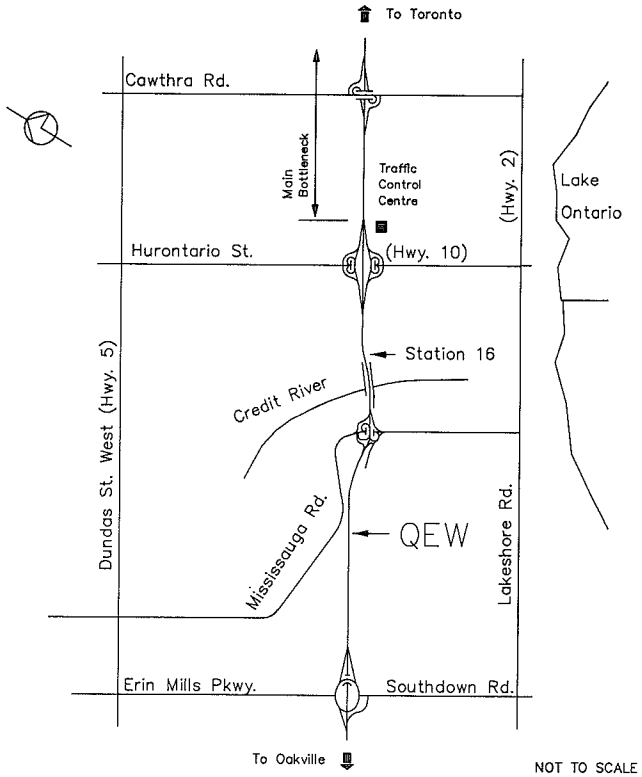


FIGURE 3 QEW Freeway in Mississauga, Ontario.

computer every 30 sec. The data used for plotting purposes included 410 30-sec intervals. The first point plotted was 6:00:30 a.m.; the last was 9:25:30. Between Interval 1 and Interval 75 (6:37:00), speeds were consistently over 80 km/hr. At about Interval 75, there was a sudden drop in speed as traffic operations became congested. At Interval 397 (about 9:16:30), speeds again climbed above 80 km/hr. From about that interval on, the freeway again operated under uncongested conditions.

A review of the data showed that speeds were missing during 15 of the 410 intervals. However, the upstream detector was able to report flows and occupancies in all these cases, and occupancies exceeded 40 percent. These 15 observations were simply deleted from the data file. Some potentially important information may be missing because of this, particularly since the diagrams rely on connections between successive points to help make sense of the pattern. However, in view of the low number of intervals for which speeds were missing, and the fact that consecutive intervals with missing speeds exceeded two only once, it was assumed that missing data would not significantly affect this investigation.

ANALYSIS

In the following subsections, data are first presented as views in the plane of the axis system (Figure 4) to obtain the traditional view, as in Figure 1. Following that are parallel oblique projections of the data where one of the planes in the *xyz*-coordinate system is placed in the plane of the page. (Figure 5 shows this with the flow-occupancy plane in the plane of the page.) The third dimension is viewed obliquely (speed in Figure 5). Projection lines in the third dimension are plotted parallel to each other (i.e., there is no apparent perspective to the view as there would be if the third dimension lines tended toward a "vanishing point").

The three-dimensional plotting is left-handed; hence, with this type of reference system, the initial origin is located at the bottom left corner. The positive *x*-axis extends horizontally to the right. The positive *y*-axis extends vertically "upward" on the page. The positive *z*-axis extends into the plane of the page, "away" from the reader.

All of the plots show data points connected with lines. The original computer work and the photography for the original report (*11*) were in color, with five different colors representing different speed ranges. For this paper, black and white were alternated for the five speed ranges, which allows each of them to stand out clearly in many figures. Figure 5 is a good example; in Figure 5a, only three ranges are clear, but all five stand out in Figure 5b. Area A contains the data with speeds above 80 km/hr. Area B (light lines) covers the range 70 to 80 km/hr; area C (dark lines) the range 60 to 70 km/hr; area D (light lines) the range 50 to 60 km/hr; and area E the range below 50 km/hr. (Interpretation of these figures is discussed later in this paper.)

The data are drawn in a reference cube; horizontal and vertical scales become ineffective when the data are rotated in three-dimensional space and where relative reference points

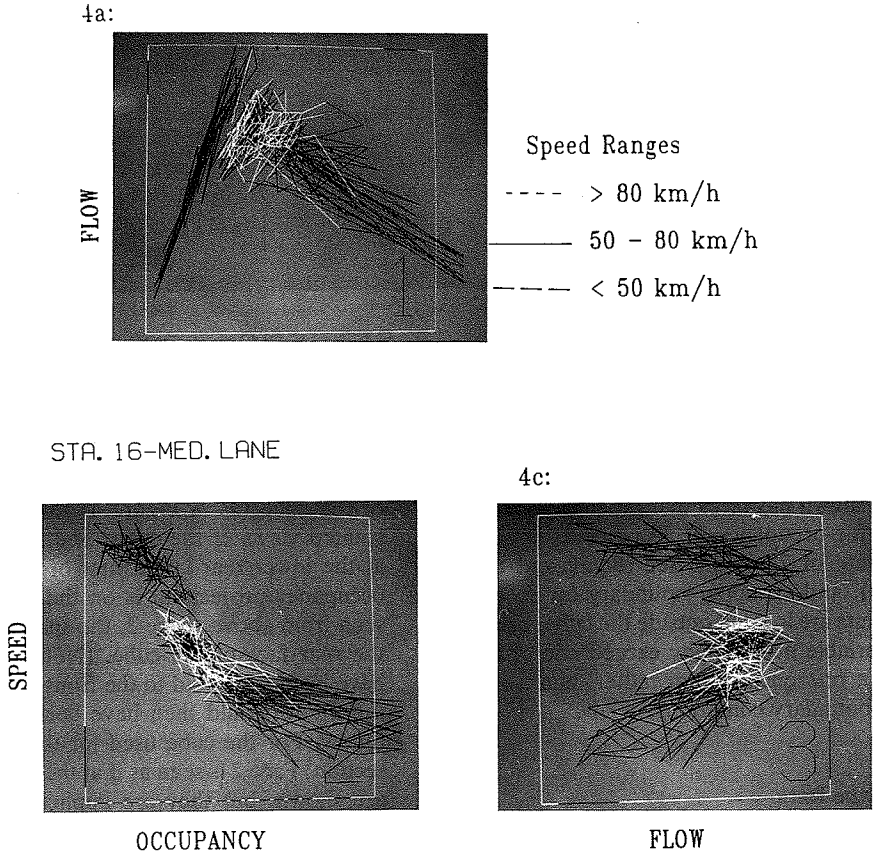


FIGURE 4 Traditional views of the Station 16 data.

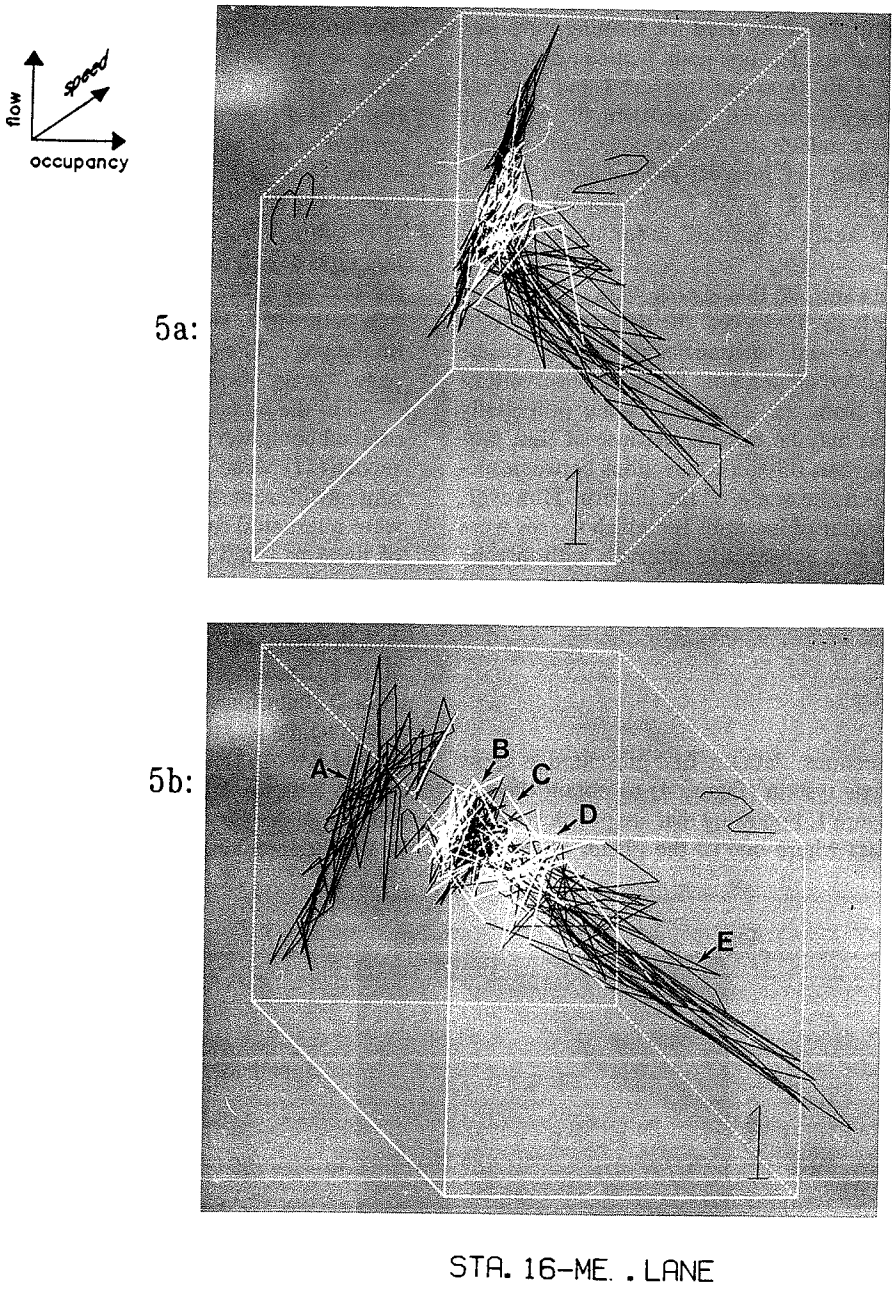


FIGURE 5 Oblique view of Station 16 flow-occupancy data.

change from view to view; hence, scale and labels are not displayed. The cube shown has faces that range from zero to an upper boundary for occupancy of 70 percent, an upper boundary for flow rate of 22 veh/30 sec (equivalent to an hourly rate of 2,640), and an upper boundary for speed of 120 km/hr. A few points are located outside the cube, because actual values of a variable occasionally exceed one of these values. The cube is a reference system rather than a set of maximum values or boundary conditions.

Traditional Views of the Data

Views of the speed-flow-occupancy relationship are obtained by observing the three-dimensional data plots from different positions in space. The traditional views are shown in Figure 6. The three numbers are reference points to label the faces of the cube (as in Figure 4) for assistance in following the rotated views of the cube. The labeled axes in Figure 6 are at the lower rear left corner of the cube. Face 1, the flow-occupancy projection, is on the bottom of the cube, and is seen from above. Face 2, the speed-occupancy projection, is the rear face of the cube. (Note that this curve could also have been projected to the front face of this box, except that the figure would then have been more difficult to draw or follow. For the cube used for the oblique views in the next section, Face 2 is seen as if at the front.) Only the projections are shown, but they are implicitly from a data set located within the cube.

Figure 4 shows these traditional views for the actual data (with the replacement of density by occupancy). It is worth noting several characteristics of the data in these figures, to

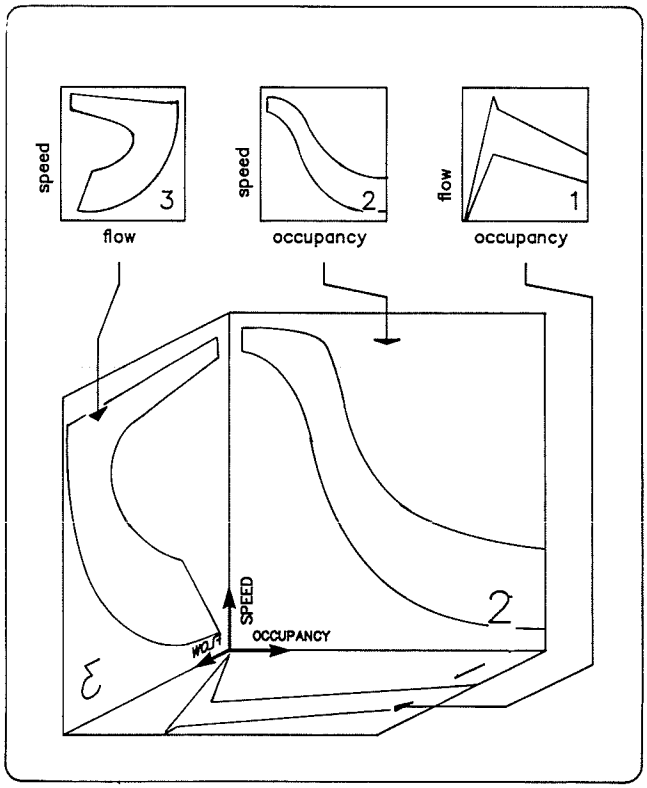


FIGURE 6 Projections of data used to obtain traditional views.

help find the speed ranges in the oblique parallel views of the next subsection. Particularly in the flow-occupancy view (Figure 4a), uncongested data are grouped within a narrow band, extending from a point close to the origin upward at an angle of about 70°. Speeds are consistently in excess of 80 km/hr, and this area of data is characterized by larger longitudinal oscillations over a relatively narrow band width. The congested data are shown by a wide band extending downward and to the right, at about 40° (from the horizontal). Data at the left end of the band are characterized by substantial vertical movement perpendicular to the longitudinal axis of the band. Speeds in this area of the band are greater than 70 km/hr. At speeds less than 50 km/hr, there appears to be a more prominent movement along the longitudinal axis of the band. At speeds between 50 km/hr and 70 km/hr, there seems to be a transition in the data from the vertical oscillations of higher speeds to the longitudinal oscillations of lower speeds. There is a small data gap between the uncongested and congested data. This gap is crossed only twice, and is more obvious in the speed-flow diagram (Figure 4c) than in the flow-occupancy figure.

The other main point to note, in these "conventional" views of the data, is the "error" pattern, i.e., the scatter of the data about what is thought to be the functional relationship (as in Figure 1). In the flow-occupancy plot (Figure 4a), there is little scatter about the left-hand branch of the curve but a considerable amount about the right-hand branch. In the other two figures, there is a very wide scatter of data about the entire curve, perhaps even sufficient to cast doubt on the narrow linear representation. It is worth considering how this error pattern affects the interpretation of statistical inference. Random variation with constant variance is expected but does not seem to be present. Further, the error for each of the three curves is not independent of the other two.

Oblique Views of the Station 16 Data

For the oblique views, the center of the view is the center of the cube, and the cube is rotated about this center point. In all of the views, the rotation of the cube is the same relative to the plane of the page. The horizontal rotation of the cube is about 28° to the left or right of a center line through the cube, and about 28° either upward or downward. View projections are effectively taken from near the corners of the cube looking toward the center of the cube. The three axes passing through the origin ($x = 0$, $y = 0$, $z = 0$) are shown as solid lines, while all other lines are dashed. Labels have been replaced by the three numbers on three faces of the cube, as displayed in Figure 4.

Figure 5 shows the data set as an oblique view of the flow-occupancy projection. (For this view, the revised cube of Figure 6 has been rotated 90° about the occupancy, or horizontal, axis.) The upper picture (Figure 5a) is a view taken from the upper right corner of the cube. The true origin is located at the lower left front corner of the view, and the cube surface with the number 1 is closest to the reader. The origin is located at the front of the cube, so points corresponding to high speeds are positioned at the "rear" of the cube (and are on the left side of the data). Low-speed points are located toward the "front" of the cube (on the right side of the data). The data in the 60 to 70 km/hr range are situated in the center, almost obscured behind other data. (These data are much more visible in Figure 5b).

In fact, all of the data for speeds above 60 km/hr seem to form a narrow band, clustered even more tightly than the uncongested data in the normal flow-occupancy view. Oscillations within the band seem to be longitudinal. All of the data for speeds greater than 50 km/hr seem to lie in a tilted plane. The view looks toward the edge of the plane, which may or may not have a slight twist to it. (Subsequent views seem to confirm the presence of this twist.)

The data for speeds of less than 50 km/hr are not as well defined, at least in this view. These data, again, form a band extending from about the middle of the cube downward and to the right. Movement between data points appears to exhibit a pattern of oscillation along the axis of the band. This suggests that very low speeds (and high occupancies) do not often occur for two consecutive intervals within this data set.

Figure 5b shows the flow-occupancy data set as an oblique view taken from the left upper corner, rather than the right. The surface with the number "1" is still closest to the reader. High-speed uncongested data are again located near the back of the cube (A). In this view, the high-speed uncongested data appear to be well grouped at higher speeds but spread out at the lower speeds and higher occupancies. The band width of these data is much wider than in the upper photograph, and more lateral movement within the band is visible. The gap in the data, which is not evident in Figure 5a, is evident in this picture.

Considering the two parts of Figure 5, there is a strong indication that the uncongested data, and, indeed, the rest of the data above perhaps 60 km/hr, fall on a surface. From the right side of the cube (Figure 5a), the lines are very compact, suggesting the surface is viewed on edge. From the left side

(Figure 5b), the surface is viewed from a much less acute angle and gives the impression that the data are much more dispersed. The lines connecting data points suggest that, while significant variation in uncongested data is present, variation appears to arise from oscillations taking place on a planar surface.

In the lower speed range during congested operation, the pattern is not as evident. It is clear, however, that these data are not on the same plane as the higher speed data. There may also be an indication of a surface for the low-speed data. At speeds close to 70 km/hr (B and C), the data exhibit a pattern similar to higher speeds. At the lowest speeds of congested data (E in Figure 5b), there is a gradual tapering; thus, the band width at lower speeds is quite narrow—not unlike the narrow band width of the uncongested data in Figure 5a. An examination of the very-low-speed data (e.g., much less than 50 km/hr) in Figure 5a indicates a wider band width. In other words, the upper view appears to "disperse" the very-low-speed congested data relative to the lower view. Hence, there is some evidence that these data may also lie on a surface, although the nature of that is less clearly defined than the surface for the uncongested data.

Figure 5 then suggests that the data may lie on a pair of planes or on a surface. More specifically, the data appear to take the form of a U or a V, with fairly broad arms, with the plane of one of the arms having been rotated away from the plane of the other arm. The result is a twisted U.

Figure 7 shows the speed-occupancy plane parallel to the page. Figure 7a is a view taken from the upper right corner of the cube looking down at the data set. High speeds are near the top of the cube, and low flows are located at the rear of the cube. The number 2 is located on the front of the

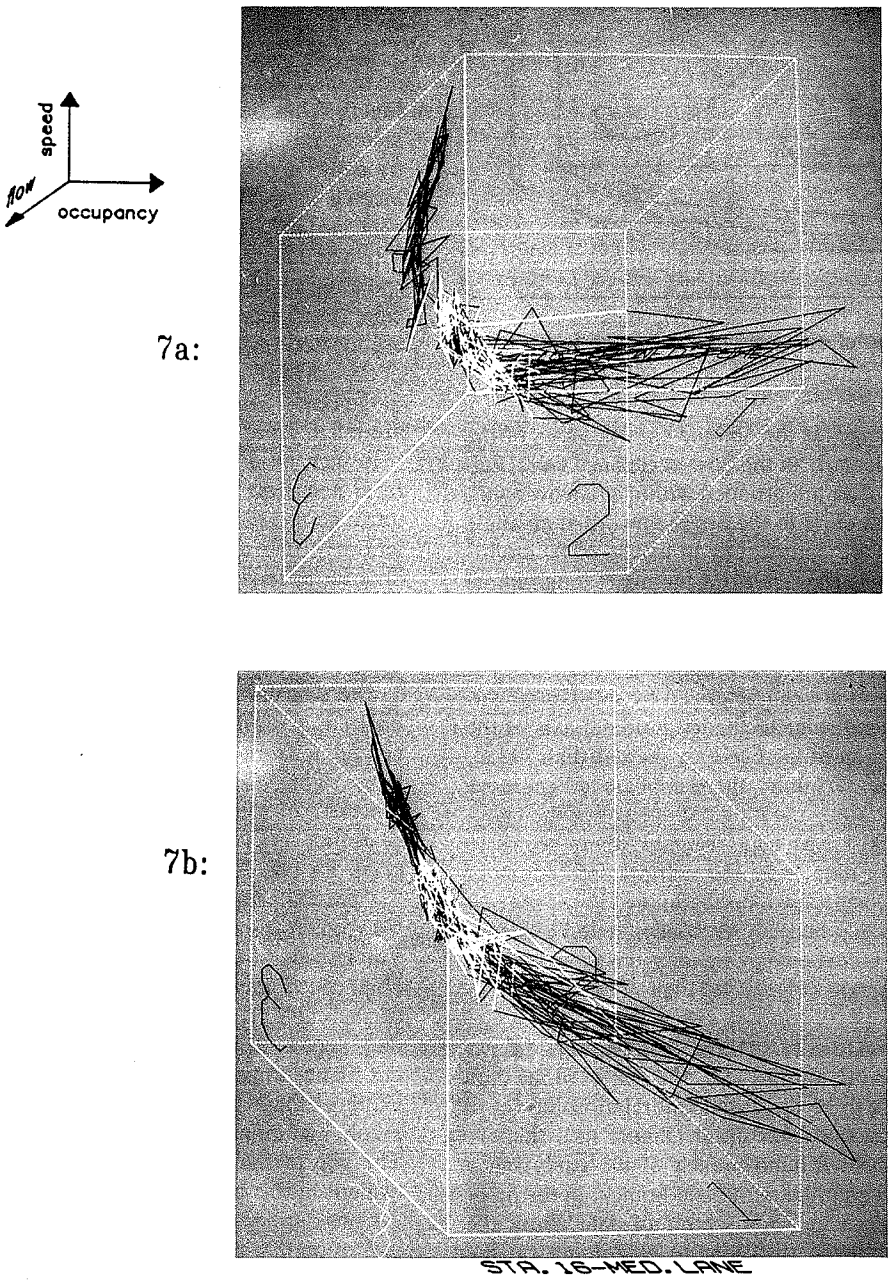


FIGURE 7 Oblique view of Station 16 speed-occupancy data.

cube (i.e., the high-flow side). The cube is positioned so the true origin is located at the back of the cube.

The most striking aspect of this illustration is that it presents a U shape in the data, as just suggested, and there are grounds for finding the twist in the U as well. Figure 5b showed considerable scatter in the high-speed data, but that is greatly reduced in this view. This means the plane is being viewed close to edge-on. This is not true for the low-speed data, however. Hence, the twist may well be visible somewhere in the vicinity of the 60 to 70 km/hr data or just beyond it in the 50 to 60 km/hr range.

Figure 7b shows the speed-occupancy view from the lower corner on the same side of the cube pictured in Figure 7a. This view looks up at the data set, and higher speeds are still at the top. High flows are located at the front of the cube, as is the face with the number 2. The true origin is located at the back of the cube on the left side.

This view of the data set appears to confirm the idea that the surface is rotating, or twisted. Data with speeds in excess of 60 km/hr are compact laterally but spread out linearly along the length of a narrow band. The general form of the data suggests the view is near the edge of a surface for the high-speed data range. The surface rotates in the lower speed range (to the right of the dark cluster) so the view is no longer along the edge of a surface but at an oblique angle instead.

Figure 8 presents the speed-flow plane parallel to the page. Figure 8a is a view taken from the lower left corner of the cube, looking up at the data. The number 3 appears reversed because it is on the back face of the cube. The direction of the U is backward from the traditional view because the view is from the "rear" of the typical projection. High occupancies lie closer to the reader. The U appears tilted into the page, and the view rotates the data so the arms of the "U" seem closer together. The two arms of the U would be expected to overlap if the front of the cube were raised slightly. The uncongested data trend upward to the left, while the low-speed congested data trend upward to the right. This appears to support the earlier suggestion of a twist in the surface.

Figure 8b is also taken from the "rear" of the typical speed-flow projection. The data are viewed from above, and the view is similar to that of the upper photograph, but the data have been rotated about 60° to the left and about 60° vertically. The front of the cube is defined by the large square in the lower left corner of the cube, which is the same face that was to the front in Figure 8a. The number 3 is obscured by the data but is located in the center of the photo, in reverse orientation. The true origin is at the rear of the cube on the right side.

Both of these views present the high-speed uncongested data in a very dispersed form compared with previous views. The uncongested data are at the top of the cube. The original photographs showed the congested data in the mid-speed ranges to be well stratified by color. Even in Figure 8, without color, the data in the mid-speed range (60 to 70 km/hr) are clearly visible. The view of the uncongested data in Figure 8b is such that the surface appears almost perpendicular to the line of vision. The area occupied by these data seems quite large from this angle.

The low-speed congested data have primarily vertical oscillations in Figure 8b, particularly in the lowest speed range (points at the bottom of the figure). The lines formed by these oscillations are parallel in nature, indicating short dwell times for these low speeds. However, in this view, the very-low-speed congested data take place over about the same range as the higher-speed congested data; both band widths have approximately the same lateral dimension. While this presents a further consistency between all speed ranges in the congested data, it does not confirm the presence of a surface for low-speed congested data.

Figure 8c is a view taken from the upper right corner of the cube, looking down at the data. The view is from the "front" of the speed-flow projection so that the U is in the normal position. The front of the cube is defined by the large square in the lower left corner of the photograph and has the number 3 in its lower right corner. The uncongested data, showing a pronounced horizontal oscillation over a wide band width, are located at the top of that front face. This is similar to the positioning of the data for the traditional speed-flow plot. However, the congested data are very compact and complex. In this view, all of the congested data lie within a mass having similar horizontal and vertical dimensions. The view is effectively looking at the congested data along the axis of oscillation—at least for the low-speed congested data. Higher speed congested data appear to be grouped within a very narrow range closest to the reader, while lower speed congested data are the farthest away.

The compression of the very-low-speed data in the projections of Figures 8b and 8c illustrates a strong consistency with the higher speed congested data. Figure 8c is an end view of the vertical oscillations for the very-low-speed congested data indicated in Figure 8b. It is not clear from the projections whether the low-speed congested data lie on a surface.

DISCUSSION

Three items are discussed in this section. The first is one possible physical description of the Station 16 data. The second deals with the match between these results and conventional theory, and the third looks at how these results match the model of traffic flow suggested by the cusp catastrophe theory surface.

Figure 9 shows one possible schematic presentation of the shape of a surface for the Station 16 data set. The surface is projected onto the sides and bottom of a partial cube (see Figure 6) to show how the traditional views of the data are obtained. In Figure 9, the surface for uncongested data is almost vertical and extends from the "rear" of the cube to the front—like a board standing on edge. At the high-flow end, the surface falls downward and to the right. As speeds and flows decline, the edge of the surface nearest the reader rotates downward for higher flows compared with low flows over a narrow range of occupancy. The broken lines of the figure indicate a possible explanation for the speed-flow projection in which occupancies are outside the plotted range.

Comparison with Traditional Traffic Flow Theory

Traditional traffic flow theory has attempted to model these data using line functions and conventional statistical estimation techniques. This analysis implies that the data should not be represented by the standard line functions, with errors

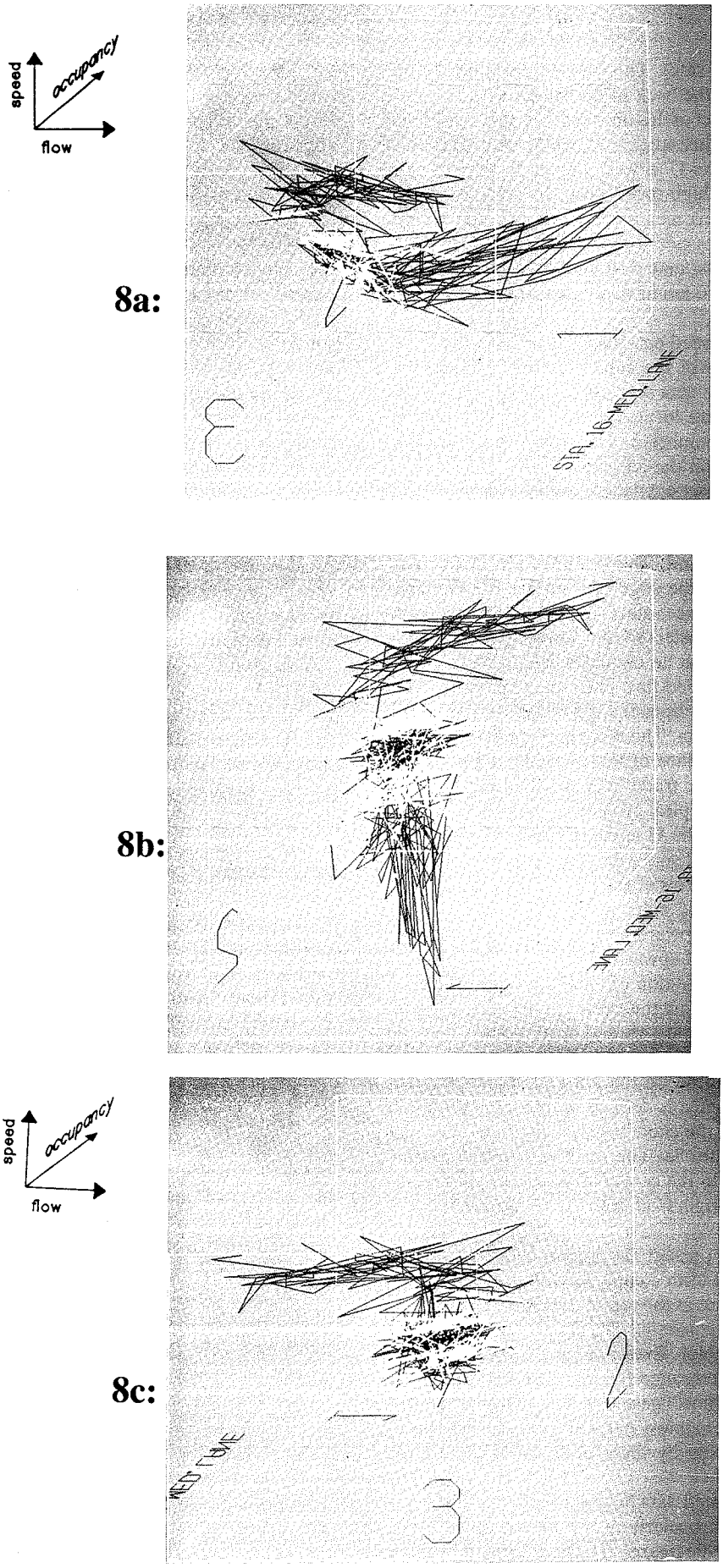


FIGURE 8 Oblique view of Station 16 speed-flow data.

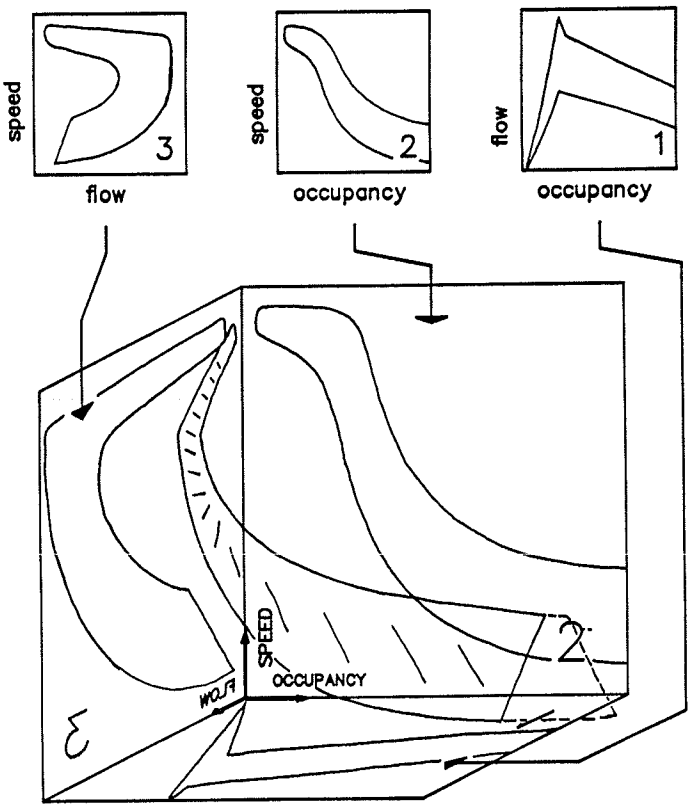


FIGURE 9 Schematic view of surface providing explanation for data scatter.

attributed to random variation in the data or errors in the estimation technique.

Scatter in the data is more correctly related to the three-dimensional model rather than to a single variable. It appears as though the scatter, particularly in the uncongested data, remains on a particular surface and is, therefore, not random with respect to a particular projection made at an angle to that surface. For this reason, statistical estimation using two-dimensional approaches is unlikely to yield consistent results in calibrating the three line functions. More complex three-dimensional estimation techniques may provide better results; however, such an approach would have to distinguish legitimate variation (i.e., on the surface) and error (i.e., variation above or below the surface). Therefore, it would seem nearly impossible to obtain a realistic representation of uninterrupted traffic flow theory using line functions.

Comparison with the Catastrophe Theory Model

This analysis provides more support for the catastrophe theory model of traffic flow than for the conventional model, but there are points of disagreement. There is strong support for three aspects of the model:

1. The clear indication of variation along what looks to be a plane for the uncongested data;
2. The fact that the congested or low-speed operations clearly do not occur on the same plane; and
3. The presence of apparent discontinuities in the form of two jumps from one side of the curve to the other.

However, there are two key points of disagreement as well:

1. The congested data do not clearly lie on a surface; and
2. There is some indication that congested operations, even down to 50 or 60 km/hr may still be occurring on the same plane as the uncongested operations.

These points of disagreement are not enough to reject the catastrophe theory model but do call for careful deliberation. As a result of this rethinking, one peculiarity of the data set used has been recognized. This particular station on the QEW lies between two onramps, both of which create bottlenecks. It is suspected that the congested data arise in two ways and that this affects their appearance. Some of the time, the queue from Highway 10 extends back through Station 16; this would give rise to the low-speed congested data. At other times, the queue from Highway 10 does not reach quite this far back but there is a queue that has formed upstream of Mississauga Road. At those times, Station 16 is recording vehicles that are accelerating from the upstream stop-and-go conditions, which would give rise to a pattern similar to that in the high-speed uncongested data. Persaud and Hurdle (*12*) describe this phenomenon clearly. Hence, the second point of disagreement with the catastrophe theory model has an explanation, and only the first point calls the model into question.

It seems that the best way to picture the three-dimensional representation is as a twisted U or V, in which both arms are more ribbons than lines. The ribbon for uncongested operations is at an oblique angle to all three conventional representations, but that angle is smallest for the flow-occupancy face of the cube. The ribbon for low-speed congested oper-

ations is clearly at a different angle than the first one, but more analysis needs to be done to determine if that arm is planar or a conical section of some kind and to decide exactly where the high-speed congested data fit.

There is an interesting overlap between these results and another current paper. Hall and Persaud (*13*) report that the so-called fundamental equation of identity (flow equals the product of speed and density) seems to hold true only for uncongested data. That equation also defines a surface that may coincide with the uncongested portion of the catastrophe theory surface. The fact that the congested data appear in this research to be situated on quite a different surface (or portion of one) is consistent with Hall and Persaud's results.

CONCLUSIONS

In traffic flow work as well as in other disciplines, new measurement techniques and instrumentation have frequently produced observations that could not easily be explained by conventional theory. The data available from current FTMSs would seem to be such a case. These systems measure all three key traffic variables (speed, flow, and concentration in the form of roadway occupancy), whereas earlier systems could often measure only two. Simple three-dimensional graphical treatment of the measured variables shows major weaknesses with conventional traffic flow theory, as depicted in standard diagrams of two-variable relationships.

Three specific problems have arisen. First, the assumptions about the distribution of error underlying normal statistical estimation techniques for two-variable equations clearly are inappropriate. Three variables are involved, and there is an interrelation of the error among them. The scatter shown in a standard two-variable portrayal of the data is perhaps more a function of the angle between that projection and the actual plane of the data than it is of the scatter in the data themselves.

Second, at least for the uncongested data, there does seem to be a plane along which all of the data fall. There is considerable scatter of data over the plane, but there seems to be little or no deviation from its surface. This observation is simply not addressed by conventional theory but may be of considerable benefit in analyzing what happens on freeways.

Third, it is clear from this analysis that the low-speed congested data do not lie on the same plane as the uncongested data. These data may or may not form a plane themselves, but a shift in the underlying relationship definitely occurs with the move into severe congestion.

As a consequence of these three problems, it would seem that the simple bivariate curves that have conventionally been used to represent traffic flow relationships are inadequate for depicting the underlying three-dimensional relationship. Rather than a thin U set at an angle in 3-space to the axes, a twisted, broad-armed U is needed, whose angle, with respect to the axes, changes with a move into congestion.

Although conventional theory cannot be supported with these results, some aspects of the recently proposed catastrophe theory model of traffic flow are confirmed. There remain some mismatches between these data and this theory, but it provides a closer fit than the conventional model.

These results have been obtained from the analysis of only one lane at one station on one roadway. While other data sets could lead to a different interpretation, that is doubtful. The data produced should not make this system unique. If there is to be a further test of the same kind, however, it is imperative that all three variables be measured. Calculation of the third variable assumes the very thing that is to be tested here, and a paper by Hall and Persaud (*13*) suggests the calculations are suspect. If all three variables must be measured, then there is a strong reason for using occupancy, rather than density, as the third variable for traffic theory.

On the whole, these results suggest that it is time to rethink some of the fundamentals of traffic flow theory and make better use of the data currently available.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada and the assistance of the Freeway Traffic Management Section, Ministry of Transportation of Ontario, in making the data available.

REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
2. *Special Report 165: Traffic Flow Theory*. TRB, National Research Council, Washington, D.C., 1976.
3. J. S. Drake, J. L. Schofer, and A. D. May, Jr. A Statistical Analysis of Speed Density Hypotheses. In *Highway Research Record 154*, HRB, National Research Council, Washington, D.C., 1967, pp. 53–87.
4. A. Ceder. *Investigation of Two-Regime Traffic Flow Models at the Micro- and Macroscopic Levels*. Unpublished Ph.D. dissertation. University of California, Berkeley, 1975.
5. 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.
6. A. Ceder. A Deterministic Traffic Flow Model for the Two-Regime Approach. In *Transportation Research Record 567*, TRB, National Research Council, Washington, D.C., 1976, pp. 16–30.
7. V. F. Hurdle and P. K. Datta. Speeds and Flows on an Urban Freeway: Some Measurements and a Hypotheses. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 127–137.
8. F. L. Hall. An Interpretation of Speed-Flow-Concentration Relationships Using Catastrophe Theory. *Transportation Research*, Vol. 21A, 1987, pp. 191–201.
9. F. L. Hall and D. S. Dillon. Freeway Operations and the Cusp Catastrophe: An Empirical Analysis. In *Transportation Research Record 1132*, TRB, National Research Council, Washington, D.C., 1987.
10. B. N. Persaud and F. L. Hall. Catastrophe Theory and Patterns in 30 Second Freeway Traffic Data—Implications for Incident Detection. *Transportation Research*, Vol. 23A, No. 2, 1989, pp. 103–113.
11. R. S. Gilchrist. Three Dimensional Relationships Among Traffic Flow Variables. M. Eng. project report. Department of Civil Engineering, McMaster University, 1988.
12. 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.
13. 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*, 1989.

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

Traffic Flow Theory and Chaotic Behavior

JOHN E. DISBRO AND MICHAEL FRAME

Many commonly occurring natural systems are modeled with mathematical expressions and exhibit a certain stability. The inherent stability of these equations allows them to serve as the basis for engineering predictions. More complex models, such as those for modeling traffic flow, lack stability and thus require considerable care when used as a basis for predictions. In 1960, Gazis, Herman, and Rothery introduced their generalized car-follow equation for modeling traffic flow. Experience has shown that this equation may not be continuous for the entire range of input parameters. The discontinuous behavior and nonlinearity of the equation suggest chaotic solutions for certain ranges of input parameters. Understanding the chaotic tendencies of this equation allows engineers to improve the reliability of models and predictions based on those models. This paper describes chaotic behavior and briefly discusses the methodology of the algorithm used to detect its presence in the car-follow equation. Also discussed are two systems modeled with the equation and their associated chaotic properties.

Classical mathematical models for natural systems, most often linear, provide well-behaved results for a wide range of input parameters. These models, such as Greenshield's for traffic flow (*1*), are characterized as predictable, deterministic, and exhibiting a kind of stability:

$$u = u_f (1 - k/k_j) \tag{1}$$

where

u = speed,
u_f = free flow speed,
k = density, and
k_j = jam density.

Another physical example of a system with inherent stability is a pendulum displaced 5.001° from vertical and released; its motion will closely follow that of a pendulum displaced 5° from vertical. The point is that small changes in initial conditions should produce small changes in resulting motion. The same assumption is often made concerning models of more complex systems, and that assumption brings certain freedoms: if small changes produce small changes, then there is stability inherent in making predictions from a given mathematical model; if small changes in initial conditions produce large changes, care must be taken when predicting based on that mathematical model.

Other models consisting of mostly nonlinear relationships, often in the form of differential and iterative equations, provide exceptions to behavior patterns typical of the classical models. Two characteristics of mathematical models are their ability to be predictable and deterministic. Any nonlinear equation can possess both, one, or none of these characteristics. An example illustrating the difference between the predictable and deterministic would include the differential equation $dx/dt = Rx(1 - x)$, and the iterative equation $x_{n+1} = Rx_n(1 - x_n)$. The equation $dx/dt = Rx(1 - x)$, is classified as a predictable and deterministic equation—knowing $x(0)$, the value of x at any time t is $[x(0)exp(Rt)]/\{1 + [exp(Rt) - 1]\, x(0)\}$. The iterative equation $x_{n+1} = Rx_n(1 - x_n)$ is deterministic—knowing x_0 precisely gives x_1, but from some values of R it is not predictable, because the only way to find $x_{1,000,000}$ from x_0 is to iterate the equation 1 million times. This example illustrates another feature of some systems, "sensitive dependence on initial conditions." A small uncertainty in $x(0)$ will produce a small change in $x(t)$ for the differential equation, while a small change in x_0 for the iterative equation, for certain values of R, produces complete uncertainty. Specifically, if $R = 3.9$ and x_0 is between 0 and 1, every term x_n in sequence also lies between 0 and 1. Taking $x_0 = 0.4$ yields $x_{28} = 0.259$, while taking $x_0 = 0.4000001$ yields $x_{28} = 0.870$. This clearly demonstrates that this equation is sensitive to initial conditions and that small—0.0000001—changes in the input parameter can produce large changes in the results.

Unpredictability does not imply that any values for the variables can occur, and for some systems a subset of variables called an "attractor" exists to which the system evolves. Although constrained to lie on the attractor, the unpredictability arises from not knowing the long-term position on the attractor. Such behavior is often reflected in the complicated geometry of the attractor. An example of a system with a simple attractor would be (in polar coordinates) $dr/dt = r(1 - r)$ $d\theta/dt = 1$. The attractor is the unit circle $r = 1$; a point $0 < r < 1$ spirals outward toward $r = 1$ and a point $r > 1$ spirals inward toward $r = 1$. Regardless of initial position, except $r = 0$, all paths are eventually arbitrarily close, traveling counterclockwise around $r = 1$ at a constant rate.

A fluid turbulence model, developed by Lorenz, was a system of three differential equations with three parameters:

$$dx/dt = -\sigma x + \sigma y$$

$$dy/dt = rx - y - xz$$

$$dz/dt = xy - bz$$

For parameter values $\sigma = 10$, $r = 28$, and $b = 2.7$, Figure 1 shows the attractor for the system. As intertracings between the two lobes show, the path does not lie on a two-dimensional surface, nor does it fill any three-dimensional region in space. This means that the attractor is a "fractal" and not a standard mathematical object. Another feature of

J. E. Disbro, Design Group, CE 1, Department of Transportation, Dulles State Office Building, 317 Washington Ave., Watertown, N.Y. 13601. M. Framer, Union College, Schenectady, N.Y. 12308.

fractals is that magnification of any portion of the attractor reveals increasingly finer structures, in direct contrast to standard geometrical shapes—for example, the circle, which on magnification becomes even simpler, more like a straight line.

A simple example of a fractal is the Koch curve, constructed by repeated applications of a certain geometrical process. This process involves subdividing a line segment into three equal lengths, erecting an equilateral triangle over the middle third, and removing the base of the triangle.

This process is repeated in the x-axis plane for the four segments of one-third length (Figure 2). The self-similarity of the Koch curve is apparent, and sufficiently magnifying any portion of the curve reproduces the entire shape.

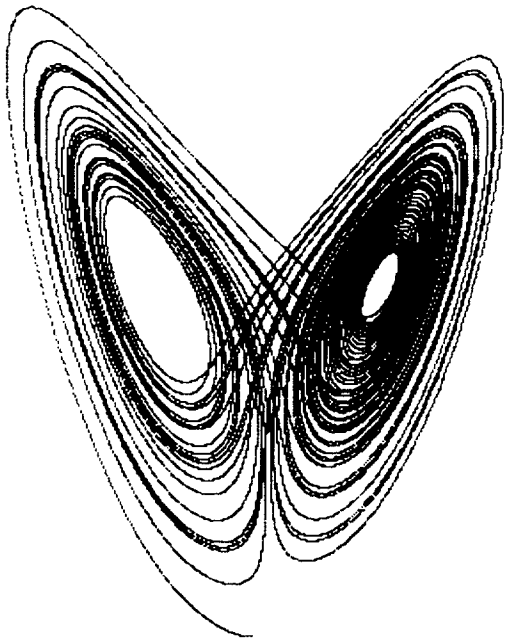


FIGURE 1 The Lorenz Attractor (x-axis is horizontal, y-axis is vertical).

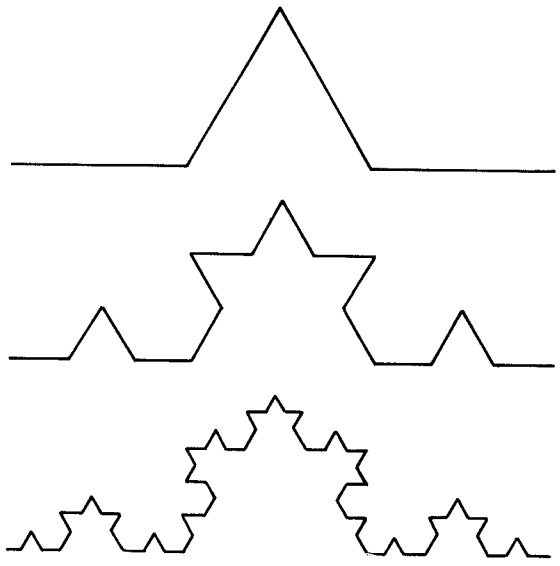


FIGURE 2 The first four stages of the Koch curve.

A "strange attractor" is one that is fractal, and chaotic dynamics are often a manifestation of a strange attractor. To determine whether a system is chaotic, strange attractors must be detected and quantified.

In the realm of dynamics, chaotic systems have three primary characteristics:

1. There is sensitive dependence on initial conditions,
2. The attractor cannot be decomposed into smaller attractors that do not interact, and
3. Any trajectory is arbitrarily close to a periodic trajectory.

Chaos, primarily associated with a state of disorder and generally considered detrimental to systems, has been discovered as a state of high order based on the geometry of the attractors. Unlike stochastic behavior, which arises from the statistical effects of treating large numbers of interacting particles representing a threshold of indeterminism, chaotic behavior is completely deterministic—but unpredictable—and occurs in systems involving as few as one variable. Predictability of chaotic systems is still limited to knowledge of long-term behavior of the associated attractor. New mathematical techniques allow attractors for nonlinear systems to be evaluated, identified, and quantified. Also, by altering input parameters, the shape of the associated attractor can be controlled and, thus, systems can be designed to produce reliable results even in chaotic states. The ability to identify and quantify attractors provides the initial steps in evaluating nonlinear chaotic systems.

Engineers have been applying the chaotic theory to thousands of systems, including thermodynamics, electrical systems, material engineering, and dynamical systems. In the field of civil engineering, chaos theory has been applied to structural vibrations and hydraulic systems. Traffic flow modeling, which contains many highly nonlinear differential equations, also offers applications for chaotic theory.

As early as 1935, engineers were developing models to describe traffic flow principles, consisting of mathematical expressions to describe basic as well as complex physical, human, and vehicular interactions. Early models for uninterrupted macroscopic traffic flow consisted of explaining functional relationships between speed, flow, and density, disregarding precise interactions between individual vehicles. Later models, called microscopic or car-following models, were developed to describe behavior of a traffic stream by the complex interrelationships involved as one vehicle follows another, and by behavior of pairs of vehicles.

In 1960, Gazis, Herman, and Rothery (GHR) developed a generalized car-following model (*2*) in which driver response is inversely proportional to the spacing between vehicles, as follows:

$$\ddot{X}_{n+1}(t + T) = \frac{\alpha\,[\dot{X}_{n+1}(t + T)]^m}{[X_n(t) - X_{n+1}(t)]^l}\,[\dot{X}_n(t) - \dot{X}_{n+1}(t)] \qquad (2)$$

where

= speed,
= accelerations,

X_n = position of the leading vehicle,
X_{n+1} = position of the following vehicle,
T = lag time, and
α, m, and l = constant parameters.

This model was well accepted and has since been reintroduced with various modifications. Experience with the equation has shown that it may not be continuous for the entire range of input parameters. The discontinuous behavior and nonlinearity of the GHR traffic-flow equation suggest chaotic solutions for certain ranges of input parameters. The chaotic realms represent areas where disturbances may not be dampened and predictability is limited. By identifying the range of chaotic solutions and the input parameters yielding such solutions, engineers can make greater use of these models. Also, engineers can control reliability of results in the chaotic realm by altering the shape of the associated attractor through modifications of input parameters—but this avenue must be reserved for future investigation.

This paper discusses application of chaotic theory to the GHR traffic-flow equation. It includes a brief discussion of the methodology used to detect chaos in the GHR equation—a more detailed description of the methodology can be obtained from the authors—and two examples of systems modeled using the GHR equation and their associated chaotic properties. A variety of input parameters are evaluated in a detailed system and the resulting chaotic properties are discussed.

DISCUSSION OF CHAOS

In recent years, "chaos," a new method of evaluating nonlinear dynamics, has arisen and received wide attention in journal articles on mathematics, physics, chemistry, biology, and engineering (*3,4*). Many nonlinear differential and difference equations with an adjustable parameter exhibit chaotic behavior for some ranges of that parameter. This section describes what constitutes chaotic behavior and the methods used to quantify chaos. In many examples, chaotic dynamics can be characterized by presence of a strange attractor in the state space of the system.

To quantify the complexity of strange attractors, an extension of the familiar notation of dimension is used. Consider a smooth curve C in three-dimensional space. An approximation of the length, L, of C can be obtained by finding the smallest number—$N_C(e)$—of cubes of side length, e, needed to cover C, and computing $N_C(e) \times e$. As e is taken smaller, this approximation improves and the limit $L = \lim_{e\to 0} N_C(e) \times e$. Similarly, for a smooth surface, S, in three-dimensional space, the area A is given by $A = \lim_{e\to 0} N_S(e) \times e^2$. The curve is one-dimensional, and the two-dimensional surface is exhibited by the exponent of e in the expression of length (the one-dimensional measure) or area (the two-dimensional measure).

Consider a simple example, where the curve is the line segment $C = [(x,0,0): 0 \le x \le 1]$ and the surface in the square $S = [(x,y,0): 0 \le x,y \le 1]$. Then for small e, $N_C(e) = 1/e$ and $N_S(e) = 1/e^2$, so $L = 1$ and $A = 1$. Notice that trying to measure the area of C yields

$$\lim_{e\to 0} N_C(e) \times e^2 = \lim_{e\to 0} (1/e) \times e^2 = 0$$

and trying to measure the length of S yields

$$\lim_{e\to 0} N_S(e) \times e = \lim_{e\to 0} (1/e^2) \times e = \infty$$

Considering just the curve C, observe that for any number $d < 1$, $\lim_{e\to 0} N_C(e) \times e^d = \infty$, and for any $d > 1$, $\lim_{e\to 0} N_C(e) \times e^d = 0$. Thus, the d-dimensional measure of curve C has the following properties: it is infinite for $d < 1$ but 0 for $d > 1$, and the length for $d = 1$. Similarly, the d-dimensional measure of surface S is infinite for $d < 2$, 0 for $d > 2$, and the area for $d = 2$.

For the Koch curve, the computation is more interesting. Taking $e = (1/3)^n$, it follows that $N(e) = 4^n$ and so the Koch curve has length

$$\lim_{n\to\infty} 4^n(1/3^n) = \infty$$

and has area

$$\lim_{n\to\infty} 4^n(1/3^n)^2 = 0$$

Thus the dimension of the Koch curve lies between 1 and 2. A straightforward calculation shows that the exponent d for which

$$0 < \lim_{e\to\infty} N(e) \times e^d < \infty$$

is given by

$$d = \lim_{e\to 0} \ln(N(e))/\ln(1/e)$$

This is the capacity dimension of the set and is closely related (and often equal) to the Hausdorff dimension. (All possible countable coverings of the set must be considered for the Hausdorff dimension, not simply those by cubes.) Observe that the Koch curve has a dimension of ln 4/ln 3.

If the dimension of a set is not an integer, then the set is a fractal, but some sets have integer dimensions that are fractals. The precise definition of fractal involves defining yet another dimension—the topological dimension—which is beyond the scope of this paper.

METHODOLOGY AND RESULTS

Methodology

This section describes the methodology used in developing a computer algorithm to test for presence of chaos in nonlinear systems. In measuring the capacity dimension of differential equation systems, counting boxes $N(e)$ can cost a lot in computer memory and time. These problems can be avoided by using Liapunov exponents. An infinitesimal sphere, centered about a point on a solution curve of the differential equation, evolves after a short time into an ellipsoid. The Liapunov exponents are natural logarithms of the ratios of the semimajor axes of the ellipsoid to the radius of the sphere, time-averaged over the trajectory.

A relationship between the capacity dimension and the Liapunov exponents is expressed in a conjecture of Kaplan and Yorke (*5*). They arrange the Liapunov exponents in nonincreasing order and allow k to be the largest integer for which

the sum of the exponents is greater than 0. The Kaplan-Yorke conjecture is that

$$\sigma = k + ((\delta_1 + \delta_2 + \cdots + \delta_k)/\delta_{k+1})$$

Although there are counterexamples to this conjecture, it is often true and holds rigorously under very general conditions $\sigma < k + (\delta_1 + \cdots + \delta_k)/\delta_{k+1})$. Determining the Liapunov exponents requires some care. The authors use a method developed by Shimada and Nagashima (*6*), and also independently by Bennetin, Galgani, and Strelcyn (*7*). Together with the Kaplan-Yorke conjecture, this method gives computational access to the dimension of attractors of high-dimensional systems.

Computing the first Liapunov exponent is sufficient to test for the presence of chaos. A positive Liapunov exponent indicates stretching of nearby trajectories, thus guaranteeing the sensitive dependence on initial conditions that characterizes chaos.

As a test, this method (algorithm) was used to compute the dimension of the Lorenz attractor (Figure 1), and the accepted value of 2.06 was obtained. Because of the complexity of the calculations and the agreement to two decimal places, the algorithm used in this report was considered accurate.

Results

The GHR equation was solved by a four-point Runge-Kutta method, modified for a delay differential equation. Tangent vectors also were processed as an array, their evolutions being governed by the Jacobian of the GHR equation. To prevent focusing of the transported tangent vectors to the direction of that with the largest Liapunov exponent, the Gram-Schmidt method was applied to produce a new orthonormal basis (*7,8*). The Liapunov exponents are the natural logarithms of the lengths of the transported tangent vectors, time-averaged along the trajectory. The Kaplan-Yorke conjecture then is applied to determine the Hausdorff dimension. The initial traffic model, consisting of eight vehicles and no disturbances (i.e., intersections, signals, bottlenecks, etc.) was developed with the GHR traffic flow equation (Equation 2) and tested for the presence of chaotic behavior. The following parameter values were selected for the system:

Variable	*Description*	*Value*
n	Number of Vehicles	8
T	Lag Time	1 sec
k_j	Jam Density	260 vehicle/mile
u_f	Free Flow Speed	55 mph
u_o	Steady State Speed	40 mph
l	Constant Parameter	2

The value of l was selected, based on ranges previously used by Ceder and May (*9*). Values for two additional variables m and α were calculated, as a subroutine in the program, using equations derived from the GHR equation:

$$m = 1 - \frac{\ln[1 - k/k_j]l - 1}{\ln[u_o/u_f]}$$

and

$$\alpha = \frac{(l - 1) \times u_f (1 - m)}{(1 - m)\, k_j\, (l - 1)}$$

The step size selected was 0.01 sec, requiring the algorithm to generate matrices of 100 rows and 16 columns to compute and store values. The program was written in Pascal and designed to compute only the first Liapunov exponent, which is sufficient to detect chaotic behavior. The simplicity of this problem, as well as the cost of computer time, did not warrant calculation of the capacity dimension; that will be reserved for the next system to be discussed.

Calculation of the first Liapunov exponents for 5,000 sec required about 8 hr of CPU time on a VAX 11-785 computer. The resulting Liapunov exponents were positive, indicating sensitive dependence on initial conditions, and thus showing the presence of chaotic behavior in the GHR traffic flow equation for these parameters, even for a simple system.

Figure 3 shows change in the first Liapunov exponents for the first 500 sec. It shows oscillations that occur due to transient behavior or system noise, caused by numerical rounding.

Figure 4 illustrates change in Liapunov exponents over time for the first 5,000 sec. No oscillations are apparent because the graph scale does not allow for sufficient detail. The large positive value (about 375) of the first Liapunov exponent, resulting after the transients have died, indicates sensitive dependence on initial conditions. The magnitude of the first Liapunov exponent should not be used as an indicator of quantitative degree of chaos in the GHR equation. No mathematical evidence exists directly relating magnitude of the first Liapunov exponent to the degree of chaotic behavior present.

To further clarify this equation's sensitive dependence on initial conditions, a small sinusoidal perturbation (range between 0 and 0.1) was added to the velocity parameter of the lead vehicle. The graph of the first Liapunov exponent versus time for the sinusoidal perturbation (Figure 4) has a more pronounced peak in the curve and a lower resulting value for the first Liapunov exponent (about 355) after all transients have died out (near 5,000 sec). This indicates that the system with perturbation settles more quickly to an attractor than the undisturbed system. This system's sensitive dependence on initial conditions is clearly illustrated by a comparison of the two graphs, showing how a small change in the adjustable parameter significantly affects the shape of the solution curve for the first Liapunov exponent.

A second system, consisting of a coordinated signal network, was modeled with the GHR traffic-flow equation (Figure 5). The network had five signals spaced at intervals ranging from 500 to 1,500 ft. The network was coordinated with a 60-sec cycle, and offsets between consecutive signals were computed accordingly. It was loaded with eight vehicles at the design speed of 30 mph (44 ft/sec). Initial vehicle positions were selected so that no vehicle was located within an intersection or directly affected by a signal indication for the first second. This was necessary to allow the computer algorithm to initialize the matrices necessary to compute and store position and velocity values. Also, the network was designed so that the entrance and exit rates of vehicles were identical. This was accomplished by including 1,500 ft of additional roadway from Signal 5 to Signal 1 and simplified the modeling.

The program was modified for the network model so that each vehicle constantly looked at the light ahead of it. If the light was green, the acceleration term for that vehicle was not changed. If the light was yellow or red, a negative term was added to the acceleration, if necessary, to stop the vehicle at the light. For example, if, when the light turned yellow, the vehicle was close enough to the light to pass through the

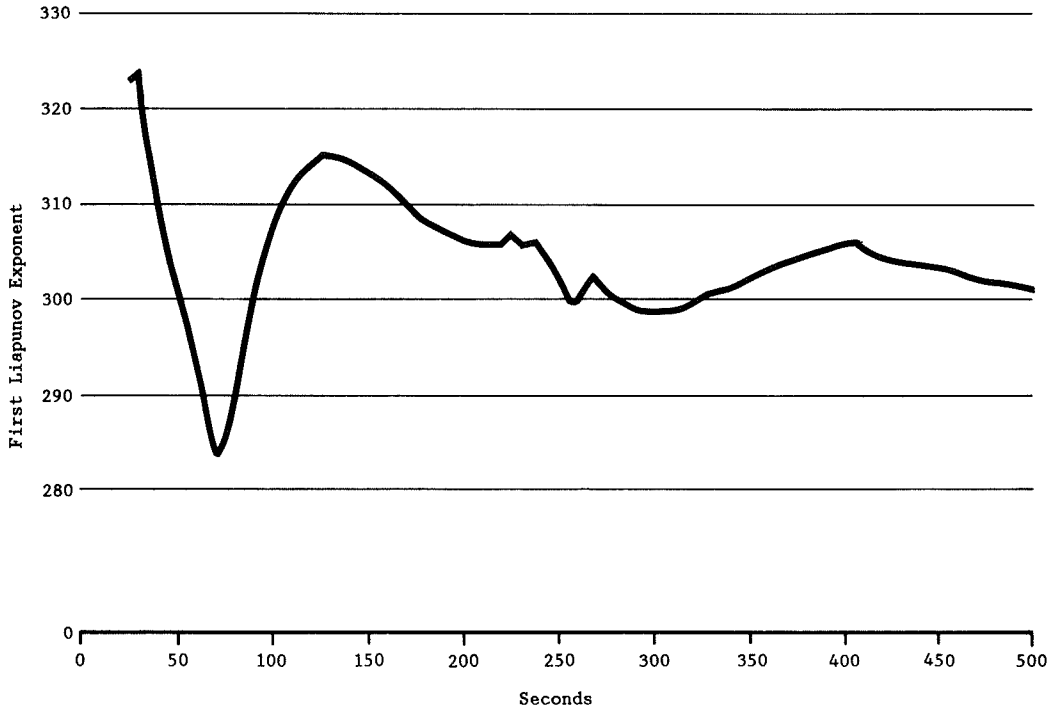


FIGURE 3 First Liapunov exponents versus time (500 seconds).

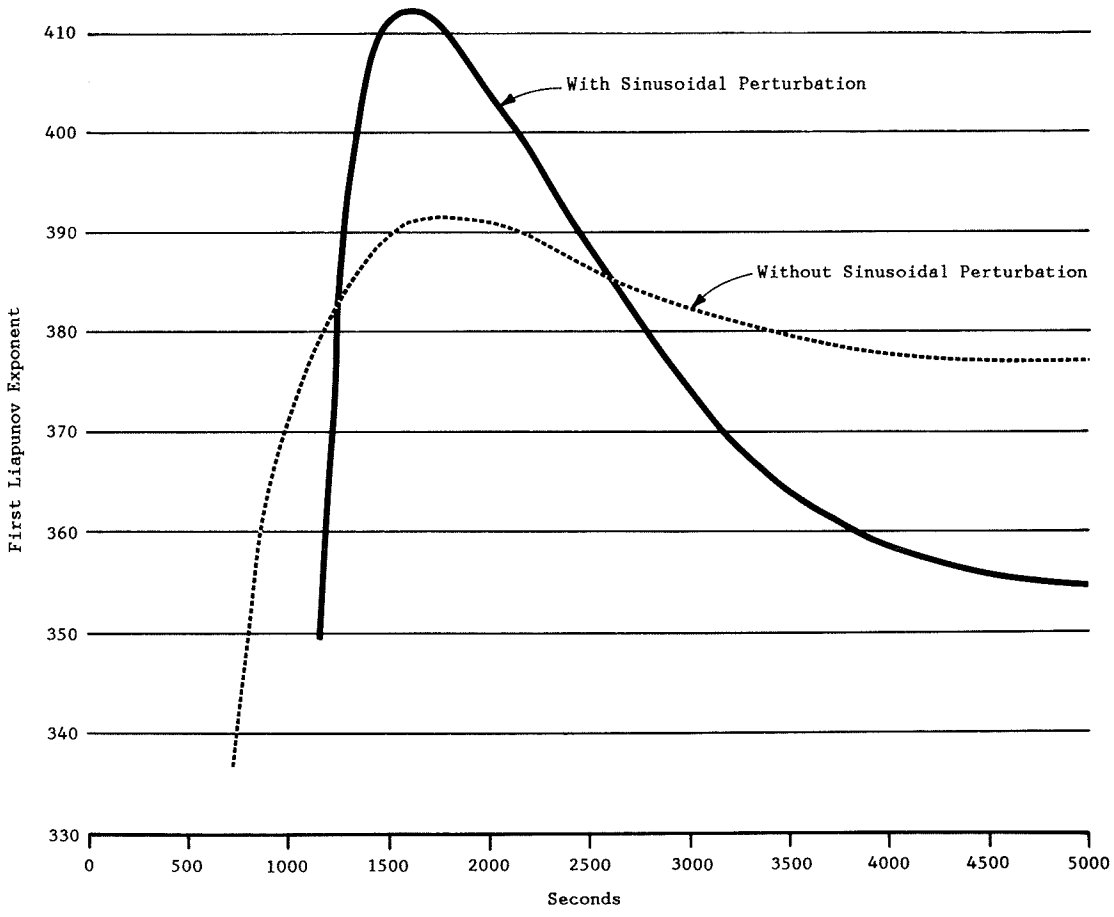


FIGURE 4 First Liapunov exponents versus time (5,000 seconds).

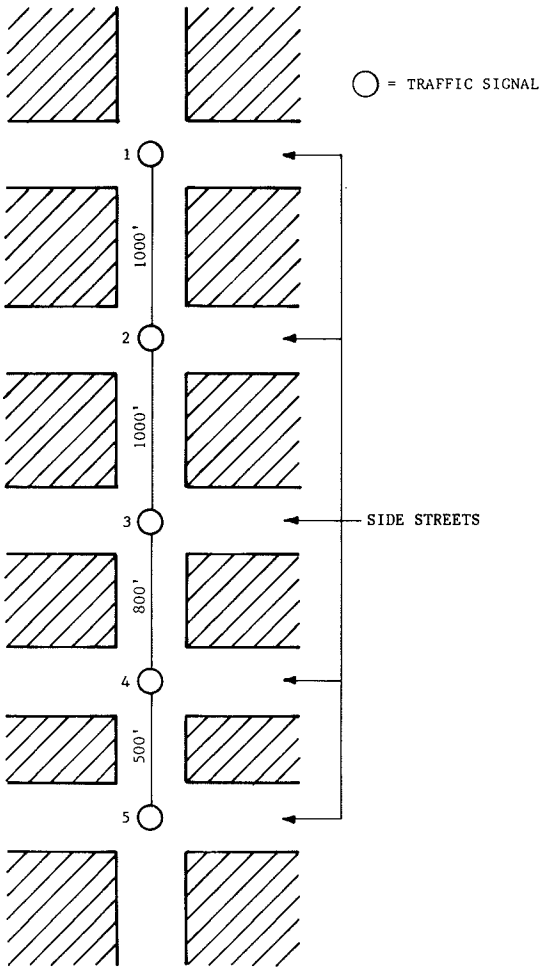


FIGURE 5 Traffic signal network.

intersection before the light turned red, then the acceleration term was not modified. If a vehicle was stopped at a red light, when the light turned green a positive acceleration term was added to bring the vehicle up to the speed limit, provided that this would not result in collision with another vehicle.

Capacity dimensions were calculated for each second for the traffic signal network, using initial speeds of 40, 44 and 50 ft/sec. Figure 6 illustrates the relationship between initial velocities and their resulting capacity dimensions. The capacity dimension for the design speed of 44 ft/sec was 14.25, indicating the presence of a strange attractor—an attractor that is fractal—to which the system can be reduced. This also shows that for an initial velocity of 44 ft/sec, 14 degrees of freedom (14 variables) are necessary to examine the system at any point in time. However, the resulting capacity dimension for initial speeds of 40 and 50 ft/sec is 16.0 (16 degrees of freedom), the maximum for this system. This further demonstrates the system-sensitive dependence on initial conditions and shows that the system modeled is inherently less complex at the design speed.

CONCLUSIONS

Chaotic behavior has been shown to exist in two relatively simple systems modeled with the GHR traffic flow equation (Equation 2). This was done by demonstrating the equation's sensitive dependence on initial conditions (positive first Liapunov exponents) and the presence of a strange attractor (indicated by noninteger capacity dimension). Two different capacity dimensions resulted from simulations using three different initial velocity parameters. The design speed of 44 ft/sec resulted in a capacity dimension of about 14, and speeds slightly higher and lower resulted in a dimension of 16. This

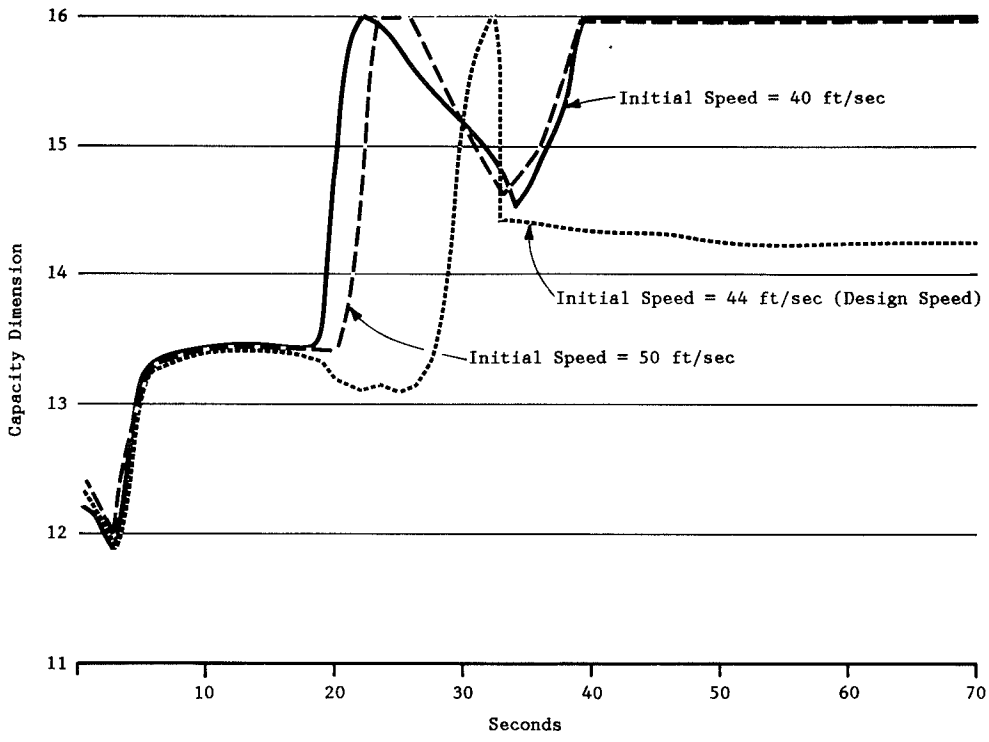


FIGURE 6 Capacity dimension versus time.

finding indicates that the degree of freedom and complexity of the system increase as speeds deviate from the design speed.

As work continues, more details regarding the attractor's geometric properties will be investigated. Knowing the geometric limitations of the attractor will improve predictions. Information on how the attractor changes shape with various input parameters will also be obtained, making more precise predictions possible for greater ranges of input parameters. Finally, attempts will be made to quantify the degree of robustness—effects caused by large changes of input parameters—further improving the reliability of predictions based on the GHR equation.

ACKNOWLEDGMENT

The authors extend special thanks to Adolf D. May, Jr., professor at University of California-Berkeley, for supplying traffic flow data.

REFERENCES

1. B. D. Greenshield. A Study of Traffic Capacity. *HRB Proc.*, Vol. 14, 1935, pp. 448–477.
2. D. C. Gazis, R. Herman, and R. W. Rothery. Non-Linear Following-the-Leader Models of Traffic Flow. *Operations Research*, Vol. 9, No. 4, 1961, pp. 545–567.
3. J. Gleick. *Chaos: Making a New Science*. Viking-Penguin, New York, 1987.
4. J. P. Crutchfield et al. Chaos. *Scientific American*, December 1986, pp. 46–57.
5. J. L. Kaplan and J. A. Yorke. Chaotic Behavior of Multidimensional Difference Equations. In *Lecture Notes in Mathematics 730* (H. O. Peitgen and H. O. Walther, eds.), Springer-Verlag, Berlin, 1979.
6. I. Shimada and T. Nagashima, eds. A Numerical Approach to Ergodic Problem of Dissipative Dynamical Systems. *Progress of Theoretical Physics*, Vol. 61, 1979, pp. 1605–1616.
7. G. Bennetin, L. Galgani, and J. M. Strelcyn. Lyapunov Characteristic Exponents for Smooth Dynamical Systems and Hamiltonian Systems: A Method for Computing All of Them. *Meccanica*, Vol. 15, 1980, pp. 9–20.
8. A. Wolf. Quantifying Chaos with Lyapunov Exponents. In *Chaos* (A. V. Holden, ed.), Princeton University Press, Princeton, N.J., 1986.
9. A. Ceder and A. D. May, Jr. 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.

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

Use of Three-Dimensional Conjugate Directions Search Method To Improve TRANSYT-7F Computational Efficiency

HUEL-SHENG TSAY AND KWO-TSAUR WANG

A modification of the computer program TRANSYT-7F has been developed to reduce computational time and improve the performance index by using the conjugate directions search method in three dimensions. The original TRANSYT-7F uses the hill-climbing method to perform a two-step optimization. This type of optimizing procedure has been used in the TRANSYT program for many years, and even TRANSYT-7F's new version, 6.0, still applies the same algorithm. In this paper, a new search method is developed to obtain simultaneously the final cycle length, split, and offset. It is a one-step optimization algorithm. From tests of 21 cases on a PC/AT, this modified TRANSYT-7F reduces computational time significantly and improves the performance index slightly compared with the new TRANSYT-7F. It also allows the user to consider the spillover effect, perform arterial priority or link maximum-allowed delay, and fix the offsets for designated intersections. Currently, this program can be used not only as a detailed off-line signal-timing analysis tool but also as a part of computing software for four newly developed traffic control systems in Taiwan to generate on-line signal-timing plans.

TRANSYT is a popular computer program used around the world to optimize the signal timing of networks with coordinated intersections. It identifies optimal offsets and phase splits through the minimization of a performance index (PI)—a linear combination of stops and delays. The weight of each stop equivalent to delay is supplied by the user network-wide or on an individual link basis. To find the minimum PI, TRANSYT uses the hill-climbing method to perform a two-step optimization. It usually uses the "quick" step size to obtain the best cycle length and then applies the "normal" step size for the split and offset optimization. The magnitude of the searching step size is given as part of the TRANSYT program but can easily be altered by the user from input Card Type 4. This is an iterative, gradient search technique that requires extensive numerical computation by the computer. It has been used in TRANSYT as the optimization procedure for many years, and the new TRANSYT-7F version, 6.0, still uses the same algorithm. Although the current TRANSYT-7F provides users with a sensitivity parameter to improve its computing efficiency, the hill-climbing method requires considerable computation to obtain the final signal-timing plan.

In 1986, Foulds (*1*) developed another search approach (a modification of the Fibonacci search) in place of the hill-climbing procedure. The Fibonacci search significantly improves TRANSYT, both in terms of the PI and computational time. Later, Chen (*2*) coded a computer program and performed several tests based on this method. He concluded that it is difficult to improve the computational time and PI at the same time through the Fibonacci search. This is probably because the Fibonacci search is a one-dimensional search. The offset and split of coordinated signalized intersections cannot be optimized simultaneously in the second phase of TRANSYT-7F after obtaining the final cycle time through the first step.

Furthermore, the PI function does not always obey the condition of strictly-increase-monotonically or strictly-decrease-monotonically, or the combinations of both. It has some variations in the form of local maxima and minima. This means that the one-dimensional search procedure may not be the best strategy for finding the global optimum of TRANSYT-7F. After numerous tests of six combined strategies proposed by Chen (*2*), it is suggested that the Fibonacci method only be considered in the first step for cycle optimization and that the hill-climbing method still be used to perform the offset and split optimization in the second step. This selected strategy can reduce the computational time by 30 percent but with a worse PI value (3 percent on average) than the new TRANSYT-7F.

In this paper, a three-dimensional, nonlinear search technique—the conjugate directions search method—is developed to replace the traditional hill-climbing method. This new method obtains the final cycle length, offset, and split simultaneously without the two-step process. It is a one-step optimization procedure and can be extended to solve special types of intersections, such as signalized circles, and signalized junctions of expressway offramps and surface arterials. The following sections first focus on the theory of TRANSYT-7F related to signal-timing optimization. Then the framework and theory of the conjugate directions search method in three dimensions is presented and discussed. Finally, this new method is compared with the hill-climbing and modified Fibonacci search methods.

CHARACTERISTICS OF TRANSYT-7F OBJECTIVE FUNCTION

To find a suitable and effective method of optimizing the objective function of TRANSYT-7F, the characteristics of PI must first be discussed. When optimizing, TRANSYT mini-

H-S. Tsay, Department of Transportation, Taipei City Government, Taiwan. K-T. Wang, Institute of Transportation and Traffic, National Chiao Tung University, Taipei, Taiwan.

mizes the PI. The optimization formulation of TRANSYT-7F follows. All variables and symbols are based on the original TRANSYT-7F (*3*), unless otherwise specified.

$$\text{PI} = \sum_{i=1}^{n} (d_i + KS_i) \tag{1}$$

where

d_i = delay on link i (of n links) (veh-hr/hr),
S_i = stops on link i (stops/sec), and
K = a user input coefficient to express the importance of stops relative to delay.

Delay in Equation 1 is composed of uniform delay (d_u), random delay (d_r), and oversaturation delay (d_s). The uniform delay is calculated by averaging the queue length (m_t) over the cycle for any step t times the cycle length, as shown in Equation 2:

$$d_u = \frac{C}{3{,}600N^2} \sum_{t=1}^{n} m_t \tag{2}$$

where

d_u = uniform delay (veh-hr/hr),
c = cycle length (sec),
m_t = queue length during step t, and
N = number of steps in the cycle.

Then random delay accrues due to the random arrivals of vehicles. TRANSYT computes the combined effect of random delay and saturation delay through the following equation:

$$d_s = \left[\left(\frac{B_n}{B_d} \right)^2 + \frac{X^2}{B_d} \right]^{1/2} - \frac{B_n}{B_d} \tag{3}$$

where

d_s = random and saturation delay,
$B_n = 2(1 - X) + XZ$,
$B_d = 4Z - Z^2$,
$Z = (2X/V) * 60/T$,
X = degree of saturation,
V = volume on the link, and
T = period length.

Therefore, the total delay in veh-hr/hr (D) can be computed as

$$D = d_u + d_s \tag{4}$$

The number of stopped vehicles estimated in TRANSYT-7F is equal to the number of vehicles arriving when a queue is present. This is based on empirical studies by the Transport and Road Research Laboratory (TRRL), and even partial stops are counted through a reduction curve if the delay to such vehicles is small (*3*). Based on Equations 1 to 4, the variables considered in the PI include the queue length, degree of saturation, volume, and cycle length. Since volume is an external input provided by the user, the queue length and degree of saturation become two major factors that need to be determined through the cycle length, split, and offset. In other words, the cycle length, split, and offset will be the major elements affecting the value of the objective function PI. These variables, however, have implicit functional forms that should be calculated internally from simulation to obtain the minimum PI value.

Generally speaking, there are four types of methods for solving unconstrained nonlinear optimization problems. These nonlinear problems can be categorized as single variable requiring derivatives (Type A), single variable without derivatives (Type B), several variables requiring derivatives (Type C), and several variables without derivatives (Type D). The solution methods related to each type of nonlinear problem are summarized in Table 1. According to the above discussion, the variables related to the objective function PI of TRANSYT-7F are nonlinear, unconstrained, and have no derivative form with the existence of implicit functions. Therefore, the objective function PI cannot be solved through nonlinear methods that require derivatives.

For nonlinear optimization problems without derivatives, the effective method for dealing with a single variable and several variables may be the Fibonacci search and the conjugate directions search, respectively (*4,5*). Use of the Fibonacci search instead of the hill-climbing procedure used in TRANSYT-7F was proposed by Foulds (*1*) and extensively analyzed by Chen (*2*). Therefore the following section focuses on the theory of the conjugate directions search method in three dimensions and its applications.

CONJUGATE DIRECTIONS SEARCH METHODS IN THREE DIMENSIONS

The conjugate directions search method in three dimensions was first presented by Powell (*4*) and then by Zangwill (*6*) and Brent (*7*). It is a direction set method that has the characteristics of quadratic termination when used to solve quadratic function problems. The procedure takes at most n steps to obtain the final solution when a quadratic function with n variables is considered. This nonlinear search method has an advantage over the other methods with its faster convergence. Since the objective function of TRANSYT-7F considers three implicit variables, the three-dimensional conjugate directions method can further be applied to solve for the cycle length, offset, and split.

Definition of Conjugate Directions

Given an $N \times N$ symmetric matrix C, the directions $S^{(1)}, S^{(2)}, S^{(3)}, \ldots, S^{(r)}$, $r \leq N$, are said to be C conjugate if (a) the directions are linearly independent, and (b) $S^{(i)}CS^{(j)} = 0$ for all $i \neq j$.

Theorem: Parallel Subspace Property (*8*)

Given a quadratic function $q(x)$, two arbitrary but distinct points $x^{(1)}$ and $x^{(2)}$, and a direction d, if $y^{(1)}$ is the solution to min $q(x^{(1)} + d)$ and $y^{(2)}$ is the solution to min $q(x^{(2)} + \lambda d)$, the direction $(y^{(2)} - y^{(1)})$ is conjugate to d.

This property is illustrated in Figure 1 in two dimensions. It can be seen that a single-variable search from $y^{(1)}$ or $y^{(2)}$ along the direction $(y^{(2)} - y^{(1)})$ will produce the minimum (*8*). This method can be further extended to three dimensions to find the optimization value of TRANSYT-7F, which is known as the application of the extended parallel subspace property. Using this extension, the construction shown in Fig-

TABLE 1 SUMMARY OF FOUR TYPES OF NONLINEAR OPTIMIZATION PROBLEMS AND SOLUTION METHODS

Types	Methods
A. Single Variable Requiring Derivatives	1. Newton-Raphson Method 2. Bisection Method 3. Secant Method 4. Cubic Search Method
B. Single Variable Without Derivatives	1. Interval Halving Method 2. Golden Section Search Method 3. Fibonacci Search Method
C. Several Variables Requiring Derivatives	1. Steepest Descent Method 2. Newton's Method 3. Modified Newton's Method 4. Quasi-Newton Method 5. Conjugate Gradient Method
D. Several Variables Without Derivatives	1. Univariate Method 2. Simplex Method 3. Pattern Search Method 4. Conjugate Directions Method

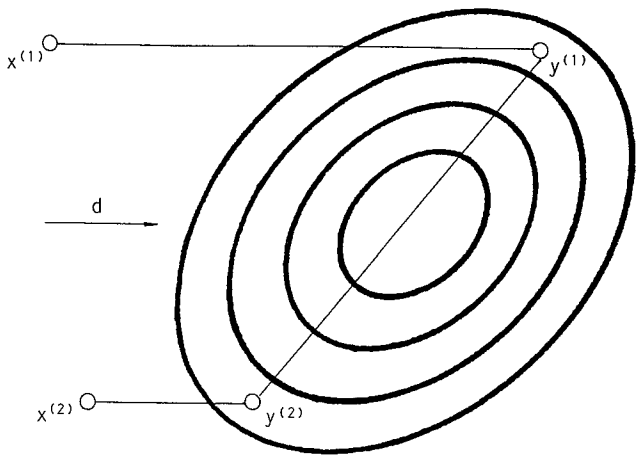


FIGURE 1 Conjugacy in two dimensions.

ure 1 can immediately be generalized to higher dimensions. The search procedure is illustrated through the example given in Figure 2 (*8*).

From the upper part of this figure, the search begins with coordinate directions ①, ②, and ③, which represent the split, offset, and cycle length, respectively. The initial signal timing plan $x^{(0)}$ is obtained from the subroutine STAR1 of TRANSYT-7F by considering the minimization of degree of saturation. From $x^{(0)}$, a series of line searches is made along ③, ①, ②, and again ③. At the conclusion of this cycle, the directions ③ and $x^{(4)}-x^{(1)}$ will be conjugate. The new search direction designated ④ in Figure 2 then replaces ①. A new cycle of line searches is executed using directions ④, ②, ③, and again ④. Through the extended parallel subspace property, the new direction $(x^{(8)}-x^{(5)})$, designated ⑤ in the figure, will be conjugate not only to ④ but also to ③. Hence, the set of directions ③, $(x^{(4)}-x^{(1)})$, and $(x^{(8)}-x^{(5)})$ are mutually conjugate. Therefore, if one additional line search is executed from $x^{(8)}$ along $(x^{(8)}-x^{(5)})$, the point $x^{(9)}$ is found. This point must be the optimum if $f(x)$ is a three-dimensional quadratic, since three mutually conjugate directions have been reached in turn. In other words, nine line searches using only function values are required to determine the exact optimum of a quadratic function in three dimensions. This construction is easily generalized and will require N^2 line searches in N dimensions to optimize a quadratic function.

The proposed method, in fact, comprises three subproblems of a one-dimensional search. It is important to determine the functional characteristics of each variable with the PI value. Several networks of Keelung and Tainan cities were tested comprehensively by the simulation procedure to derive the functional forms of cycle length, offset, and split versus PI. The relationships between the PI and each variable can be observed from Figures 3 through 5.

From the six examples shown in Figure 3, the functional form of PI related to the offsets is a type of discrete function. Most functions obey the characteristics of the sine or cosine pattern, except with some variations. It is not a unimodal function. As far as the computational efficiency is concerned, the Fibonacci search procedure approximately follows the sine or cosine curve. This procedure depends on a numerical

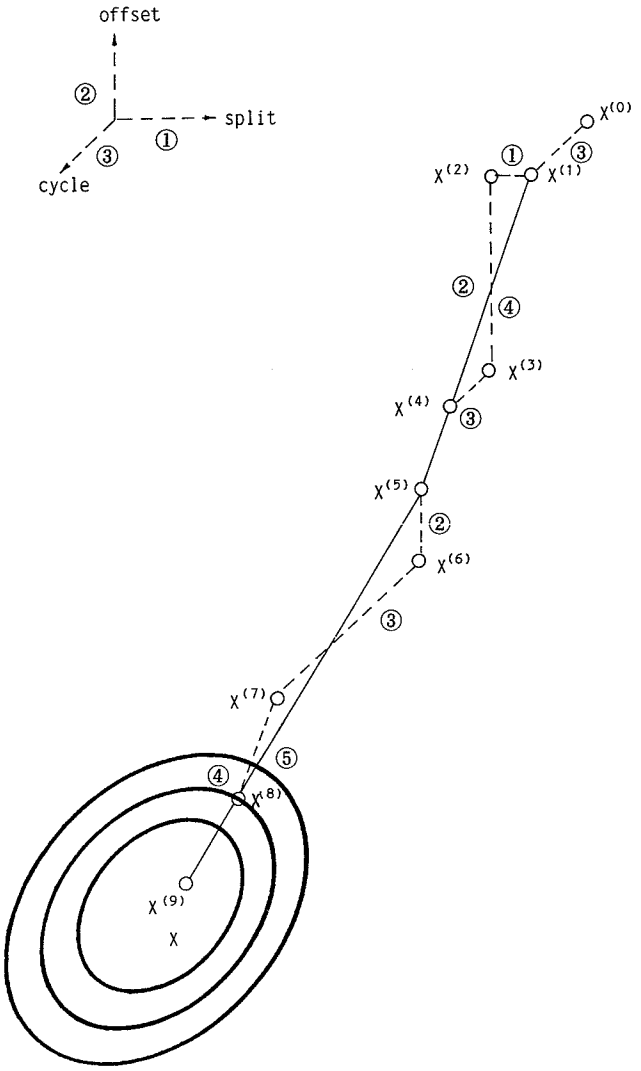


FIGURE 2 Construction of conjugate directions search in three dimensions.

sequence called Fibonacci numbers. The procedure successively reduces the interval in which the minimum of a nonlinear function must lie. The computational procedure requires the function value of only one point instead of two points for each reduced interval after obtaining the first interval from computing the value of two points. Such a searching procedure can reduce the computational time significantly. Since the unimodality of the sine or cosine function does not hold, the Fibonacci search needs to be modified. Full details of the Fibonacci search procedure and its modification were given by Foulds. In this paper, these searching steps are further considered and partially modified to find the final value of offsets in approaching the optimal solution through the conjugate directions search method.

The relationships between PI and different green splits can be seen from Figure 4. In these six cases, the green split seems to have the form of a unimodal function. Since this is a rather important element, more tests are needed before a final conclusion can be drawn. The variables considered in the objective function PI have implicit functions, which are probably the only way to determine the relationships between the green split and PI by using the simulation result. Although a theoretical proof of the unimodal function is difficult, the optimal value of splits can be obtained directly from the Fibonacci search method with the assumption of satisfying unimodality. This nonlinear function may become a constrained condition if the user sets the minimum green split for the designated approach through input Card Type 2X in TRANSYT-7F. Since the Fibonacci search method solves only the unconstrained case, a split-adjusting procedure needs to be considered and has been included in this proposed method to satisfy the constrained situation.

Figure 5 represents the relationships between PI and various cycle lengths. The PI increases gradually (Figures 5A, 5C, and 5D) or decreases first and then increases (Figure 5B) as the cycle length increases. There are some exceptions (Figures 5E and 5F) where PI decreases rapidly to a minimum value when the cycle length increases to a certain extent. To seek an effective search method for obtaining the cycle length, three kinds of searching rules may be considered. First, the Fibonacci search method is used directly to calculate its optimum value. The elimination interval needs to be recomputed if it is greater than 10 sec. Second, the entire cycle length is divided into several 5-sec intervals. In each interval the PI value is calculated; these values are compared to find the smallest PI and its corresponding interval. In the final interval the Fibonacci search procedure is applied to obtain the cycle length. Third, the PI value is evaluated at each cycle length increment, which is part of the input provided by the user.

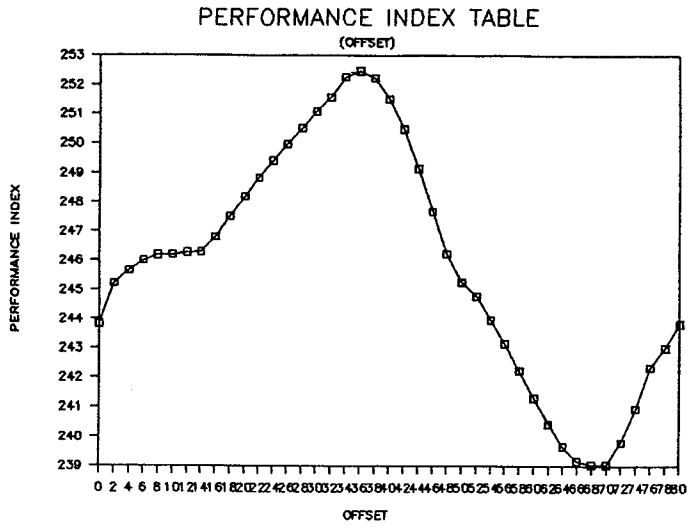
After several tests, it was found that the first and second searching rules need considerable computational time and cannot be guaranteed to give the optimal cycle length. The third rule, however, comes close to the optimum value if the cycle increment is gradually decreased. Hence the proposed method uses the third rule to calculate the final cycle length. More research is needed to obtain the exact functional form of the PI value with various cycle lengths.

The proposed conjugate directions search method is composed of three incremental vectors. To approach the global optimum, each linear search method obviously needs to be modified. Otherwise, the solution may be restricted in the local minimum because it lacks the unimodal property. Using the above searching process, the suitable moving distance must be determined after a good search direction is found. In this paper, the ratios of the offset and split variations versus the cycle length increment are computed first. Then, the offsets and splits equal to each ratio times the actual cycle increment are calculated. The split value is discarded if it violates the preset minimum green split. Finally, the final cycle length, split, and offset can be determined until line searches of all conjugate directions have been completely executed.

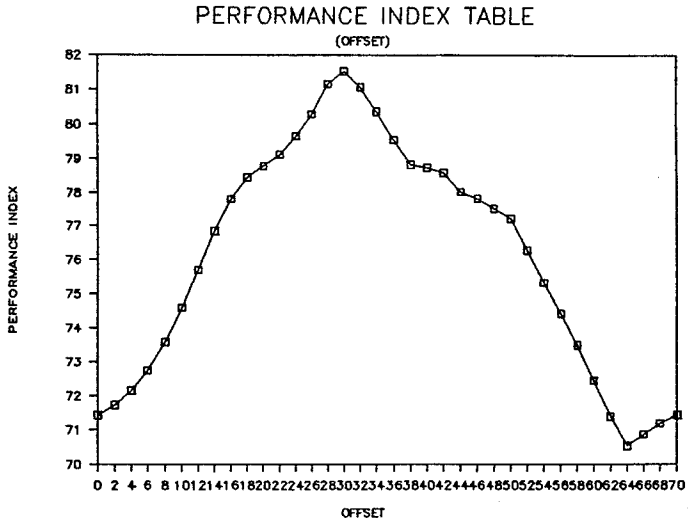
COMPARISONS OF COMPUTATIONAL TIME AND PI VALUE

The hill-climbing search, modification of the Fibonacci search, and three-dimensional conjugate directions search are used to compare 21 cases with different traffic volumes from Keelung, Taichung, and Tainan cities in Taiwan. To make a consistent comparison, the same network information and traffic flows were used to prepare the inputs for the three search methods. TRANSYT-7F has been separately programmed

(A)



(B)



(C)

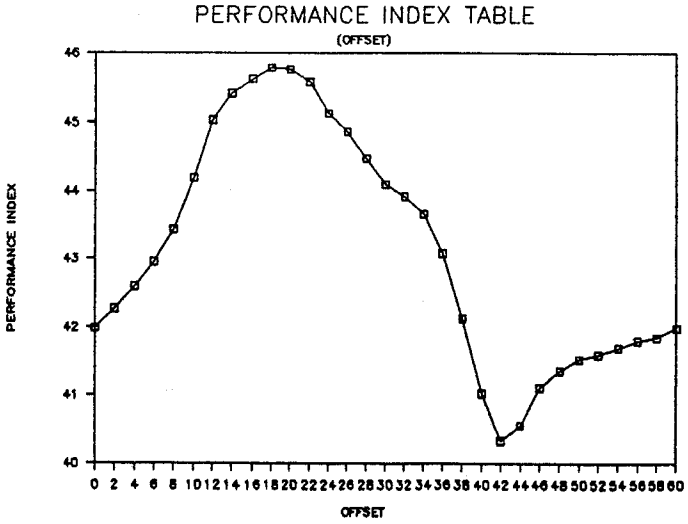
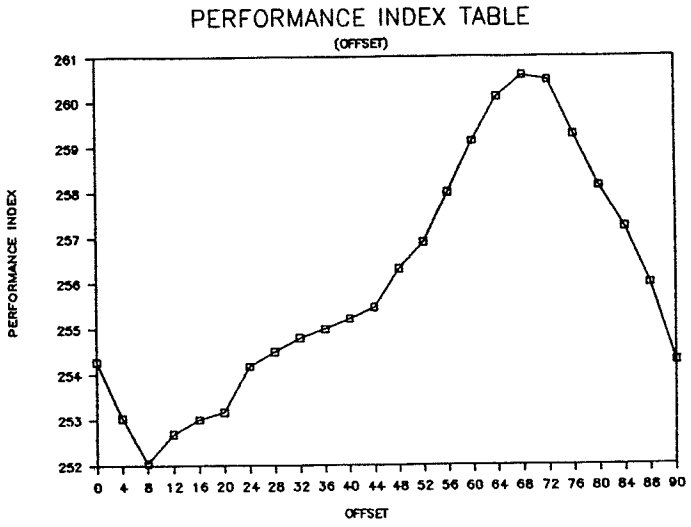
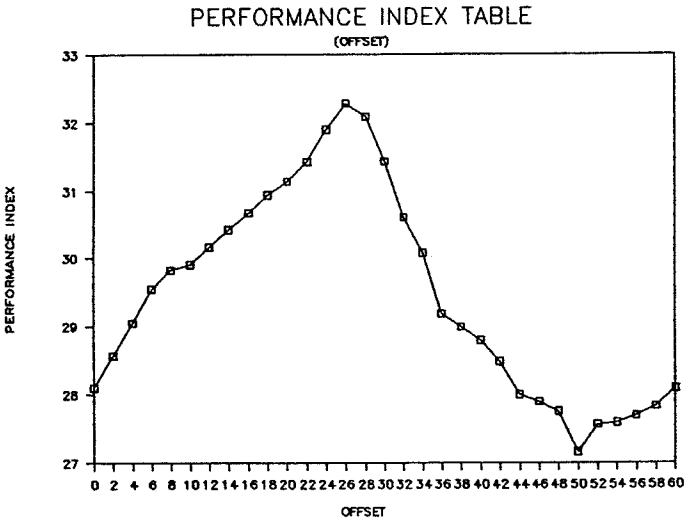


FIGURE 3 Relationships between PI and various offsets.

(D)



(E)



(F)

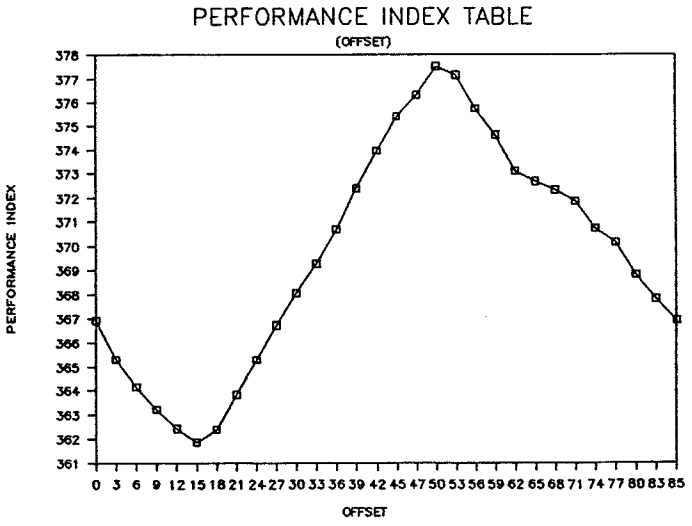
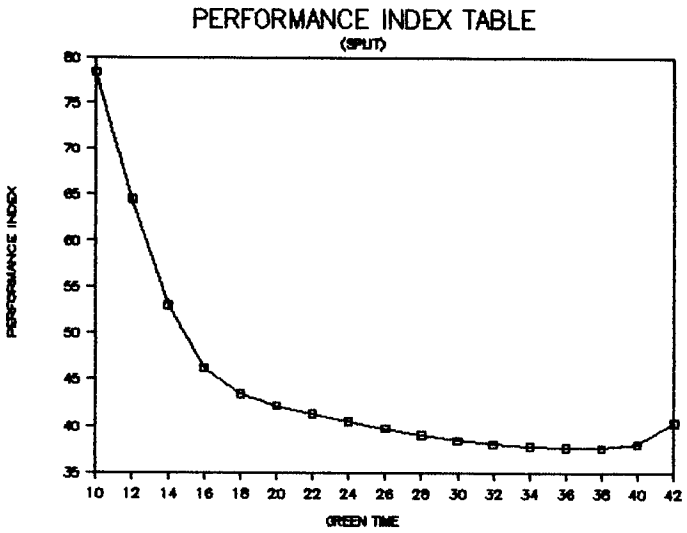
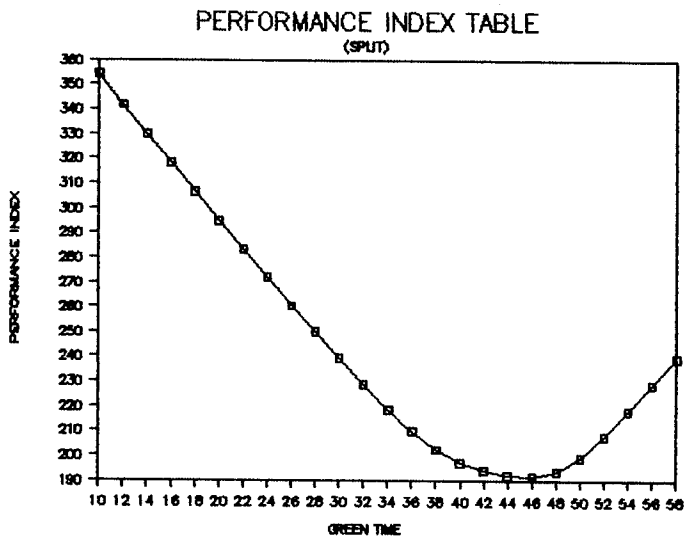


FIGURE 3 (*continued*)

(A)



(B)



(C)

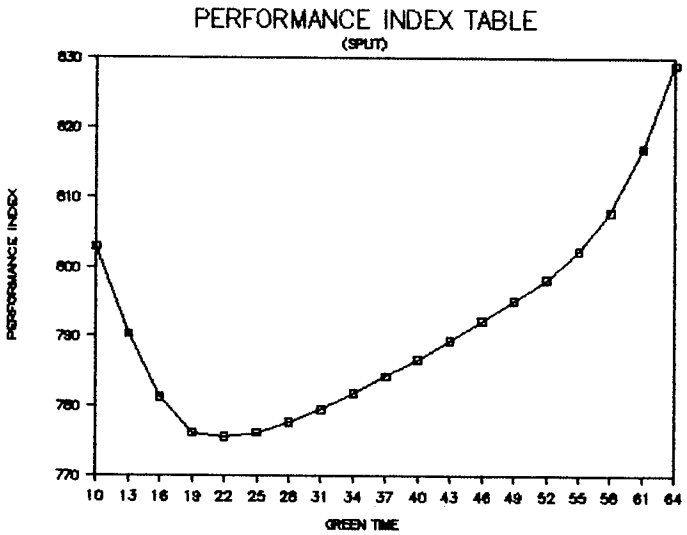
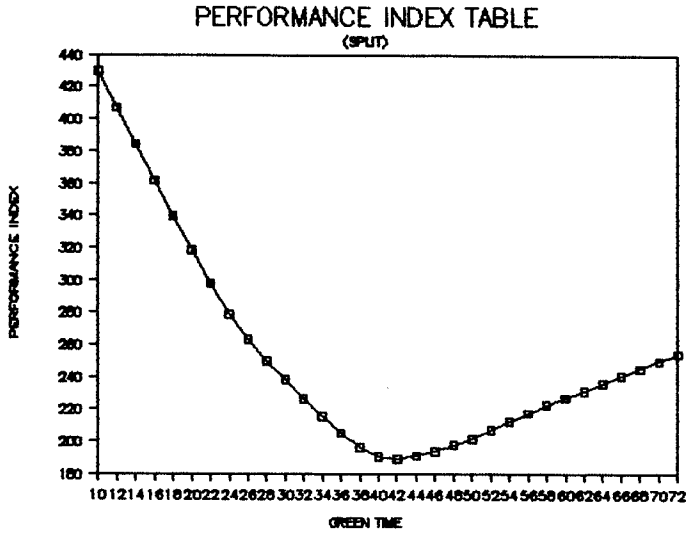
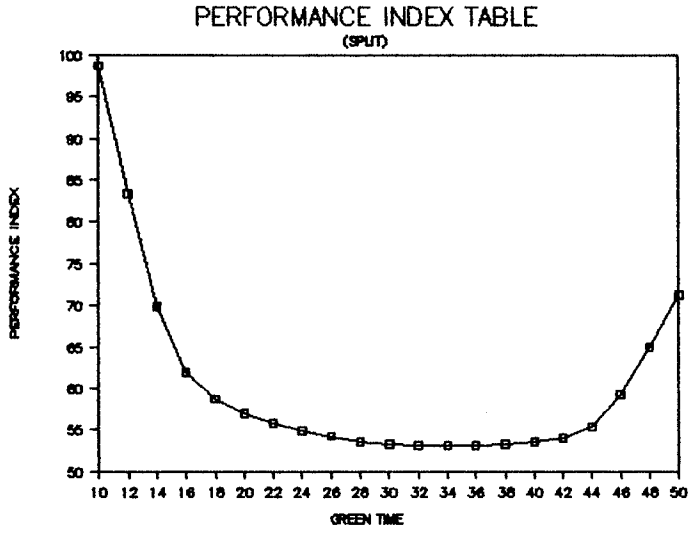


FIGURE 4 Relationships between PI and different green splits.

(D)



(E)



(F)

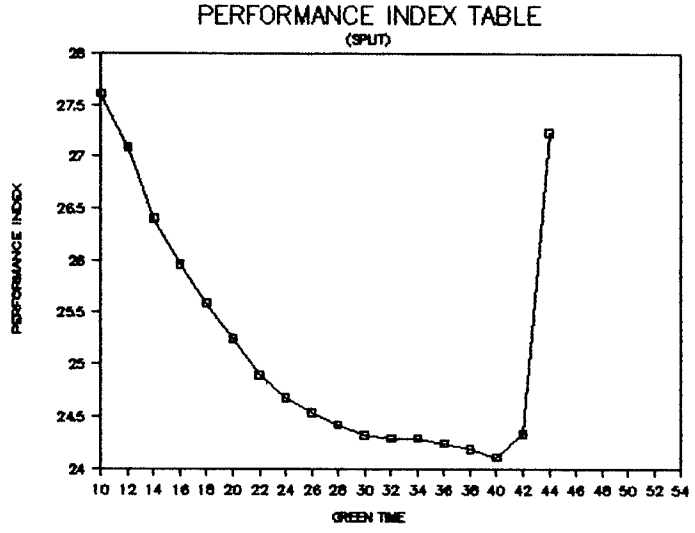
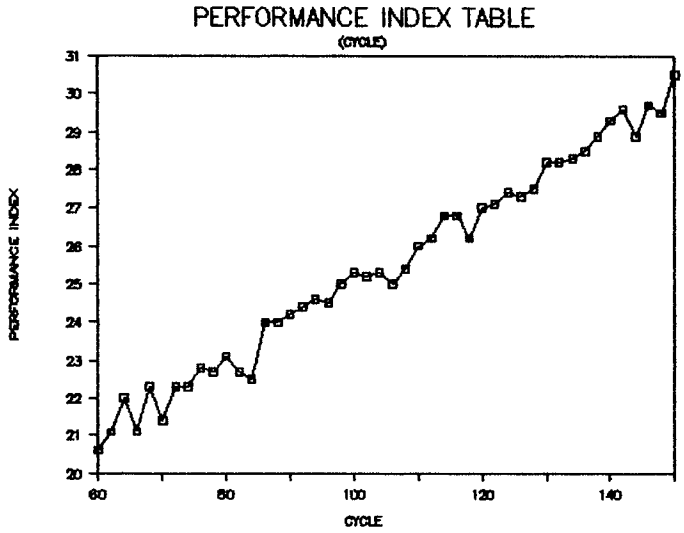
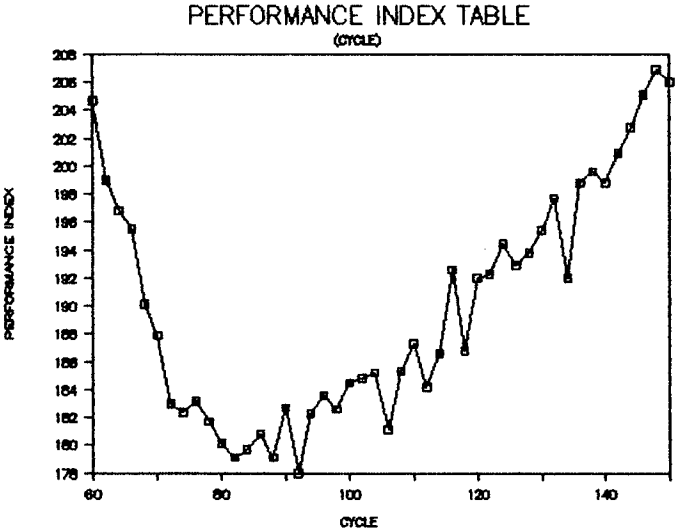


FIGURE 4 (*continued*)

(A)



(B)



(C)

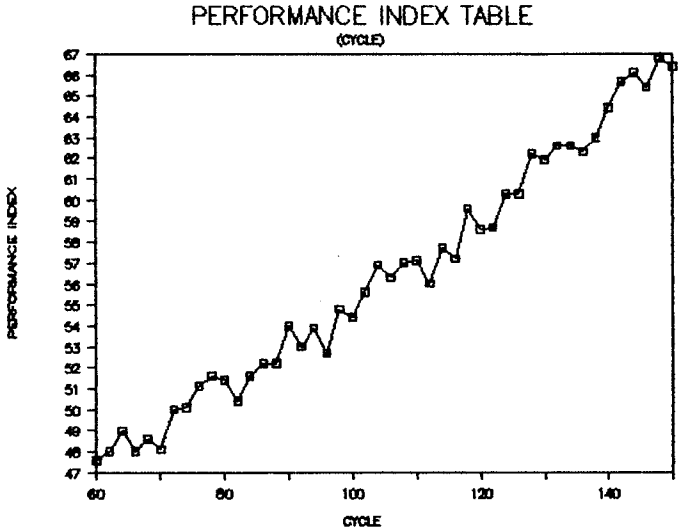
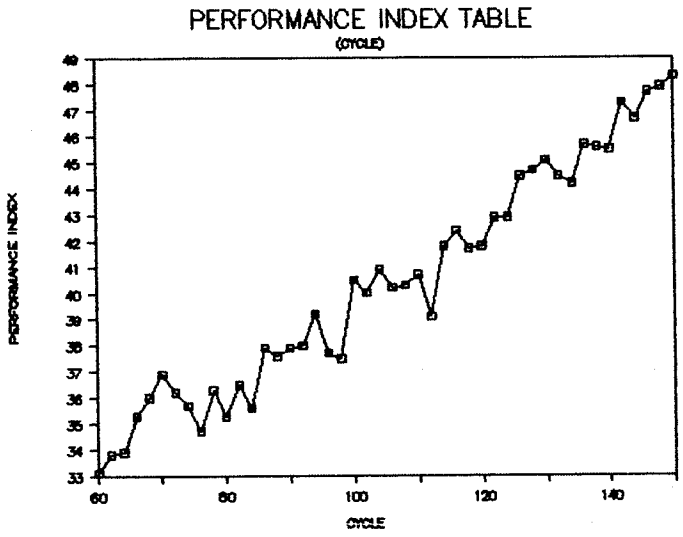
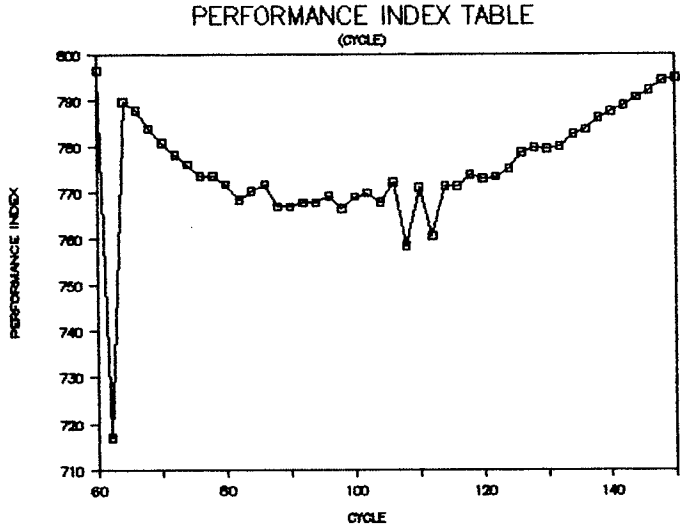


FIGURE 5 Relationships between PI and various cycle lengths.

(D)



(E)



(F)

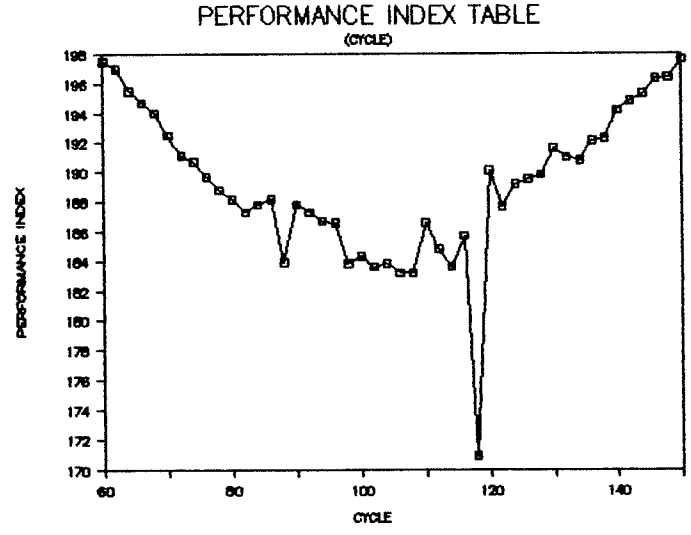


FIGURE 5 (*continued*)

with different search methods and can be run on an IBM PC, CDC, or VAX. In Table 2, the results of the computational time and PI are displayed for 21 coordinated intersections. Times are given for the same PC/AT with math coprocessor 80287-10. In all cases, the conjugate directions method reduces the computational time significantly more than the hill-climbing method but a little less than the Fibonacci method. The conjugate directions search method improves the PI in 20 out of 21 cases when compared with the hill-climbing method. It is also better than the Fibonacci method in terms of the PI in most cases. It should be noted that the modified Fibonacci method and program used here are based on Chen's study (*2*). In other words, the Fibonacci search is considered only at the first step for cycle optimization and the hill-climbing method is still applied to find the final offset and split at the second step.

Since the modified TRANSYT-7F gives a rather fast computing capability and better performance values with the conjugate directions search method, it has been used as a part of computing software to generate on-line signal-timing plans in a newly developed traffic control system known as TRUSTS (Traffic Responsive and Uniform Surveillance Timing System) (*9*). Two TRUSTS systems have been installed in the cities of Keelung and Taichung, and another two are expected to be completed in Taoyuan and Chiayi by the end of June 1989. TRUSTS calculates signal-timing plans through a 32-bit PC with math coprocessor 80387-16, which can handle up to 40 intersections for on-line timing plan generation and table selection. Part of the on-line output for six intersections from this modified TRANSYT-7F program is given in Figure 6.

A NEW FORMULA FOR SPILLOVER PENALTY

The modified TRANSYT-7F also considers the effect of spillover. A new formula for the spillover penalty has been derived

TABLE 2 COMPARISONS OF CONJUGATE DIRECTIONS SEARCH WITH FIBONACCI SEARCH AND HILL-CLIMBING SEARCH FOR 21 CASES

Cases	No. of Intersections	Cycle Ranges (sec)	Cycle Increments (sec)	Time(sec, PC/AT)			PI		
				CDS[a]	FS[b]	HC[c]	CDS	FS	HC
1	4	60-150	2	700	694	992	769.08	770.81	792.75
2	4	60-150	5	332	306	452	767.73	771.98	767.73
3	4	60-150	10	211	187	277	767.73	771.98	767.91
4	4	60-150	2	202	182	305	191.12	199.96	190.83
5	4	60-150	5	135	125	170	189.59	190.19	190.19
6	4	60-150	10	114	88	125	190.19	190.19	190.19
7	4	60-90	2	206	174	291	126.78	144.33	128.57
8	4	60-90	5	139	115	162	126.78	131.60	128.57
9	4	60-90	10	119	90	120	126.78	131.60	128.57
10	7	60-90	2	389	494	681	105.02	105.02	105.02
11	7	60-90	5	238	274	367	106.49	103.73	107.43
12	7	60-90	10	191	200	260	106.49	107.73	107.43
13	7	60-90	2	475	482	684	357.92	359.61	358.20
14	7	60-90	5	314	279	378	357.92	362.32	357.92
15	7	60-90	10	260	215	275	357.92	362.32	357.92
16	8	90-120	2	450	485	550	264.92	277.20	302.50
17	8	90-120	5	301	294	338	264.92	261.16	264.92
18	8	90-120	10	256	227	292	264.92	264.92	264.92
19	15	90-120	2	1457	1957	2274	211.64	222.40	215.85
20	15	90-120	5	875	1052	1226	211.64	211.92	216.25
21	15	90-120	10	699	766	835	211.64	211.92	211.92

[a] CDS (Conjugate Directions Search in Three Dimensions)
[b] FS (Fibonacci Search)
[c] HC (Hill-Climbing Search)

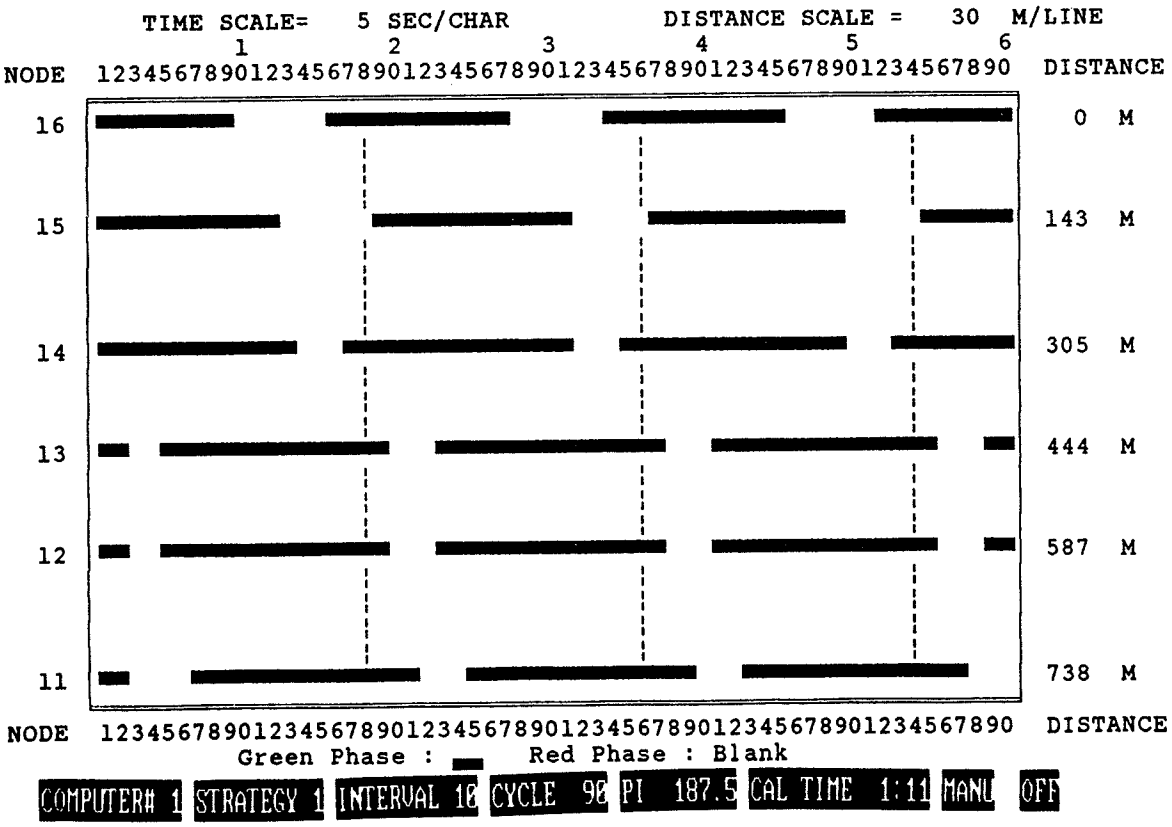
LOAD AUTO MANUL SAVE ERASE QUIT 01:16:59 PM

< SYSTEM PERFORMANCE >

TOTAL DELAY (VEH-H/H)	AVERAGE DELAY (SEC/VEH)	TOTAL UNIFORM STOPS (VEH/H-%)	TOTAL FUEL CONSUM (LI/H)	PERFORMANCE INDEX	SPEED (KM/H)
101.36	14.06	12400.4(48%)	753.13	187.47	16.25

		CYCLE	G1	G2	G3	G4	G5	OFFSET
INTERSECTION	1	90	22	58	0	0	0	50
INTERSECTION	2	90	43	37	0	0	0	0
INTERSECTION	4	90	28	52	0	0	0	2
INTERSECTION	5	90	39	41	0	0	0	2
INTERSECTION	6	90	22	58	0	0	0	88
INTERSECTION	7	90	19	61	0	0	0	6
INTERSECTION	11	90	68	12	0	0	0	30
INTERSECTION	12	90	70	10	0	0	0	21
INTERSECTION	13	90	71	9	0	0	0	21
INTERSECTION	14	90	67	13	0	0	0	82
INTERSECTION	15	90	64	16	0	0	0	88
INTERSECTION	16	90	54	26	0	0	0	77

<< T I M E - S P A C E D I A G R A M >>



COMPUTER# 1 STRATEGY 1 INTERVAL 10 CYCLE 90 PI 187.5 CAL TIME 1:11 MANL OFF

FIGURE 6 On-line output of the modified TRANSYT-7F for six intersections.

and combined with the number of stops and delay to form a new PI:

$$PI = \sum_{i=1}^{n} \left[d_i + kS_i + \delta D_Q \left(\frac{Q_{max} - Q_{cap}}{Q_c} \right)_i \right] \quad (5)$$

where

$(Q_{max})_i$ = maximum queue length within a cycle considering 5 percent variation of flow on link i (veh/lane),
$(Q_{cap})_i$ = maximum allowable queue capacity on link i (veh/lane),
$(Q_c)_i$ = allowable queue vehicles for the width of the upstream intersection on link i (vehicles),
D_Q = penalty value for the spillover, and
δ = 0 if $Q_{max} \leq Q_{cap}$, 1 if $Q_{max} > Q_{cap}$.

The definitions of variables related to spillover can be seen in Figure 7. These values are represented by the equivalent number of queue vehicles instead of the queue length. After comparisons of system performance from several saturated networks, it is concluded that the proposed spillover equation has a better performance value than the one used in SIGOP-III (*10*). Therefore, this new formula is included in the modified TRANSYT-7F to prevent the overflow problem through Field 11 of input Card Type 10.

OTHER SPECIAL DEVELOPMENT

The modified TRANSYT-7F allows the user to determine either arterial priority or maximum allowed link delay. The arterial priority option is similar to recent research (*11*) but uses different concepts of input and performance requirement. The term "arterial priority" is used because the user can specify which arterial links will receive higher priority to proceed under the constraint of setting the maximum allowable delay for the minor street. The unit of the delay value used considers the seconds per vehicle instead of the degree of saturation. To determine the maximum allowed link delay, any link may be assigned with a reasonable and different maximum allowed delay value. The final signal-timing plan satisfies all setting delay restrictions on the designated links. Through this option, some saturated links on arterial and minor streets can be avoided. This is a good way to design arterial signal-timing plans for urban streets.

The modified TRANSYT-7F can fix the offsets for certain intersections through the three-dimensional conjugate directions technique during the optimization phase. It can perform the optimizing search of the cycle length and splits simultaneously under the preset offsets, given network geometry, and traffic flows. This is a rather useful tool when designing special types of intersections, such as signalized circles, signalized junctions of expressway offramps and surface arterials, and progression of coordinated signals. A summary of these special functions related to the newly added input Card Type is shown in Table 3. In addition to the on-line capability, the modified TRANSYT-7F can perform a detailed off-line signal-timing analysis like that of the original TRANSYT-7F but with more functions and less computational time.

CONCLUSIONS

This paper has discussed how a new conjugate directions search method in three dimensions can be used instead of the traditional hill-climbing search method to optimize the signal-timing plan in TRANSYT-7F. After comparing three different search methods, it is concluded that, in most cases, the proposed method significantly reduces computational time and slightly improves PI when compared with the TRANSYT-7F. The modified TRANSYT-7F based on the proposed search

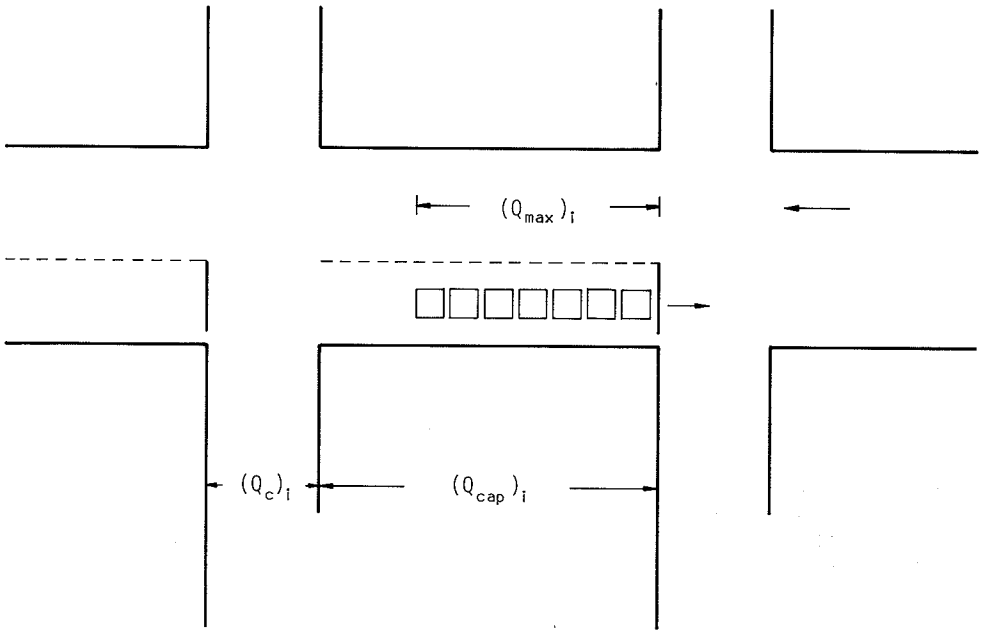


FIGURE 7 Graphic representation of variables in the spillover formula.

TABLE 3 SPECIAL FUNCTIONS OF MODIFIED TRANSYT-7F RELATED TO THE NEW ADDED INPUT CARD TYPE

Card Type	Field	Column	Description
10	11	46-50	0-10000 (Spillover penalty range)
31	1	1-5	31
	2	6-10	0 or 1 ("0" for arterial priority control, "1" for link maximum-allowed delay control)
	3	11-15	Link number
	4	16-20	Link maximum-allowed delay in sec/veh
	5-16	21-80	Same as Fields 3 and 4, Repeat
32	1	1-5	32 (When this Card Type 32 is used, the modified TRANSYT-7F will perform the optimization of the cycle length and split)
	2	6-10	99

technique has been developed and programmed in four different types of computer systems. It has a new PI formula that considers the spillover penalty and allows the user to perform arterial priority or link maximum-allowed delay and to fix the offsets for certain intersections. This program can be used as a detailed off-line analysis like the new TRANSYT-7F. Furthermore, it has been applied on a newly developed traffic control system—TRUSTS—as a part of computing software for generating on-line signal-timing plans. Thus, it is recommended that this modified TRANSYT-7F be used to obtain on-line or off-line signal-timing plans of a given network if the TRANSYT-7F type of PI function is desired. Finally, the search method for each conjugate direction remains a fertile area for future research to further reduce computational time and improve the PI.

ACKNOWLEDGMENT

The authors would like to acknowledge the assistance of Ru-Miaw Hwang and Jen-Yao Wu in preparing this manuscript and in performing several tests of different computer programs.

REFERENCES

1. L. R. Foulds. TRANSYT Traffic Engineering Program Efficiency Improvement Via FIBONACCI Search. *Transportation Research*, Vol. 20A, No. 4, 1986, pp. 331–335.
2. C. H. Chen. *Development of a Modified TRANSYT-7F Program Through the Modification of Fibonacci Search*. Master's thesis. Graduate School of Transportation and Communication Management Science, National Cheng Kung University, Taiwan, 1987.
3. *TRANSYT-7F User's Manual, Release 5.0*. FHWA, U.S. Department of Transportation, 1987.
4. M. J. D. Powell. An Efficiency Method for Finding the Minimum of a Function of Several Variables Without Calculating Derivatives. *Computer Journal*, 1964, pp. 155–162.
5. M. Avriel. *Nonlinear Programming Analysis and Methods*. Prentice-Hall, Inc., 1976.
6. W. I. Zangwill. Minimization a Function Without Calculating Derivatives. *Computer Journal*, 10, 1967, pp. 293–296.
7. R. P. Brent. *Algorithms for Minimization Without Derivatives*. Prentice-Hall, Inc., 1973.
8. G. V. Reklaitis, A. Ravindran, and K. M. Ragsdell. *Engineering Optimization Methods and Applications*. John Wiley & Sons, Inc., 1983.
9. H. S. Tsay. A New Type of Urban On-line Computerized Traffic Control System. *Compendium of Technical Papers*. Institute of Transportation Engineers, 1988, pp. 101–106.
10. *SIGOP-III User's Manual*. FHWA, U.S. Department of Transportation, 1983.
11. M. J. Moskaluk and P. S. Parsonson. Arterial Priority Option for the TRANSYT-7F Traffic-Signal-Timing Program. In *Transportation Research Record 1181*, TRB, National Research Council, Washington, D.C., 1988.

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

Headway and Speed Data Acquisition Using Video

M. A. P. TAYLOR, W. YOUNG, AND R. G. THOMPSON

Accurate knowledge of vehicle speeds and headways on traffic networks is a fundamental part of transport systems modelling. Video and recently developed automatic data-extraction techniques have the potential to provide a cheap, quick, easy, and accurate method of investigating traffic systems. This paper presents two studies that use video-based equipment to investigate the character of vehicle speeds and headways. Investigation of headways on freeway traffic allows the potential of this technology in a high-speed environment to be determined. Its application to the study of speeds in parking lots enabled its usefulness in low-speed environments to be studied. The data obtained from the video was compared to traditional methods of collecting headway and speed data.

Digital image-processing applications offer the potential to automate a large number of traffic surveys. It is, therefore, not surprising that considerable interest has been directed at developing procedures for vehicle classification and number plate recognition (*1*, *2*). However, there appear to be several obstacles preventing these systems from being used regularly in traffic surveys. Some of these difficulties are the relatively high cost of the systems, problems of accuracy, quantity of data collected, and the time required to process the raw data. Fortunately, systems are being developed that can be used for specialized applications. This paper presents one such system. It is a video-based system that allows vehicle speed and headway data to be extracted.

Video-based techniques offer the means to overcome many of the difficulties of collecting information on speeds and headways. The basic idea is to create a video recording of traffic, then extract data from it. The video approach has a number of advantages including being unobtrusive and requiring a small labor component. Its major advantage is the production of a permanent, complete record of the traffic scene. This may be reanalyzed at any stage, and provides an account of each traffic event observed.

The oft-quoted disadvantage of the video technique is the large amount of time and effort needed for data extraction. The use of an automatic data-extraction (image processing) system can overcome this problem. The video vehicle detection system (VADAS) (*3*) developed by the Australian Road Research Board (ARRB) offers a sophisticated, high-technology system useful in data extraction. VADAS was developed for use in studies of the performance of unsignalized intersections, particularly roundabouts where complex traffic maneuvers take place. It had not been applied to a high-speed environment (such as a freeway) before this study, so there was an excellent opportunity to evaluate the system and suggest modifications to it. This equipment also made it feasible to investigate the relationship between vehicle speeds and location in car parks.

Department of Civil Engineering, Monash University, Clayton, Victoria, Australia 3168.

THE VIDEO SYSTEM

Using film equipment to obtain a permanent record of vehicle movements is not a new concept. However, considerable recent developments have occurred in collecting data using video. In particular, ARRB has developed a trailer-mounted video recording system (*3*). This relatively new equipment has until recently experienced only a limited range of applications. It consists of a gas-operated extendible mast (with video camera attached), a power supply, and other video accessories concealed in an inconspicuous, covered trailer. A range of lenses is available that enable it to view various aspects and areas.

The extraction of the data from the video could be carried out by a human observer, but this is time consuming. A preferable method is to extract the data automatically. A procedure for carrying out this task has been developed by Troutbeck and Dods (*3*). The position of the vehicle is recorded as it passes through a particular point on the screen. These points, termed detection points, are identified in the data collection stage. The movement of a vehicle through these points is determined by a change in the intensity of the image and is recorded on computer tape for further analysis. The basics of VADAS are as follows. The image-processing system relies on the observation that most vehicles have a video illuminance level much greater than that of the road pavement on which they are traveling. The general luminance of the scene is monitored by locating a reference point just off the carriageway. Thus, changes in ambient light levels due to cloud cover and the angle of inclination of the sun can be accommodated.

Besides the reference point, VADAS allows the analyst to define up to 16 detection points within the video frame. The illuminance level at each point is then monitored by VADAS, and a change in luminance above a set threshold is taken to mean the passage of a vehicle over that point. The basic steps involved in data extraction using VADAS are the positioning of detection points on the screen and the adjustment of the sensitivity threshold. Once the data have been logged, they can be edited. The output from this process is a computer file of times and detector numbers that can be used in subsequent data analysis. The speeds and headways can be calculated from data on the times when vehicles were detected.

HEADWAY STUDY

A convenient way to describe the inherent variability within a traffic stream is to consider it as a stochastic process and examine the headways between vehicles in a traffic stream. Headways are defined as the time separation between the same point on two successive vehicles in a traffic stream. This definition holds for both single-lane and multilane traffic streams. Statistical distributions can be fitted to represent headways. The reciprocal of mean headway is equal to the mean flow rate.

Headways are perhaps the most important of all traffic variables since a large amount of traffic research and simulation exercises make use of headway models in the analysis of traffic problems. Buckley (*4*) stated that "headways are the building blocks with which the concepts of volume and capacity, and indeed the entire traffic stream are constructed." The ability of traffic models to replicate traffic systems realistically will be enhanced through more efficient and accurate techniques of collecting headway data.

A number of theoretical models of headways exist. Schuhl (*5*) and Buckley (*4*) provided some of the earlier research in this area. The most comprehensive theoretical study of headway distributions was provided by Cowan (*6*), who derived a family of four headway models. The first three members of this family are the negative exponential distribution, the shifted exponential distribution, and the mixed exponential distribution. These three models show increasing levels of sophistication from the simple random traffic model (the negative exponential distribution) through the shifted exponential distribution (which allows for the finite length of vehicles traveling in a single lane by imposing a minimum headway) to M3, which represents traffic as a mixture of free and following vehicles. Cowan fitted the models to data collected in a single lane on an arterial road. He concluded that the M3 model (and the fourth member of the headway model family, M4) could provide a reasonable description of traffic flow.

A number of problems emerge in the application of statistical distributions to traffic headways. There are two theoretical issues of importance. First, the models assume that mean flow rate (q), and hence mean headway, are constant over time (i.e., the traffic process is assumed to be stationary). Although this may not be unreasonable for short time periods, it can cause problems when collecting the large data sets needed to estimate the parameters for the theoretical models and to fit the distributions, particularly when using conventional data-collection methods. Second, the headway distribution models also assume that the headways are independent of each other.

Another difficulty is a practical one, of particular relevance to studies of multilane traffic flows. This is the problem of observing and assessing the full set of vehicle headways between lanes and within lanes. It is this difficulty that a video-based data recording method can help to overcome. In the past, most observational studies of headways [e.g., Cowan (*6*)] have concentrated on a single lane of traffic. The usual method of treating multilane traffic is to assume that vehicle arrivals are independently distributed and that a superposition of single-lane headways can be made. The validity of this approach could be questioned if vehicle arrivals are not independent. Miller (*7*) and Mahalel and Hakkert (*8*) found some evidence of correlations between vehicle arrivals in multilane traffic. Such correlations should be expected on arterial roads where there are signalized intersections. There is no clear evidence one way or the other for freeway traffic flows.

Another phenomenon of interest is the distribution of traffic between lanes on a multilane road. The conceptual model is that traffic uses (indeed, under some highway codes, should use) the outside lanes of a carriageway unless overtaking, when vehicles will use the inside (median) lane(s). Under light traffic volumes, this model works satisfactorily. But, as traffic volume increases, this flow regime breaks down. Vehicles in the outside lane that overhaul a slower vehicle in that lane and wish to overtake find increasing difficulties in changing lanes. For instance, Mahalel and Hakkert (*8*) examined traffic flows on a two-lane carriageway in Israel and found that the median lane carried more than 50 percent of the traffic flow for flow rates above 1,150 vph. Chen (*9*) found that the median lane attracted the largest volume of traffic on a three-lane carriageway in Ohio, for total carriageway flows exceeding 4,500 vph. This tendency for increased proportions of total flow to use the faster lanes as flow builds up leads to a greater tendency for bunching in that lane, and perhaps more instability in the traffic flow.

Although the negative exponential distribution has not proven to be a good descriptor of single-lane headways, there is a strong possibility that it could represent multilane traffic flows since headways close to zero are feasible. Buckley (*4*) and Gerlough and Barnes (*10*) both suggested that the negative exponential model could be used for multilane unidirectional traffic flow.

Given these observations, the data collection scheme was designed to include the defined flow regime; traffic volumes in the identified flow regime were spread out over different times of day and different days to account for the known variations in traffic flow levels over hours of the day and days of the week. Spreading the data-collection periods would provide the means to avoid misinterpretations based on fluctuations or deviations from the normal traffic pattern.

Theories of traffic flow and models of traffic system performance require comprehensive, accurate, and unbiased data for their appraisal. Much theoretical work has been directed at the case of uninterrupted flow, in which the performance and behavior of a traffic stream is dictated by the nature of the stream and the interactions between vehicles in the stream, free of outside influences. The traffic situation most closely approximating these conditions is probably freeway traffic. Data recording by conventional means, involving the placement of vehicle detectors on the carriageway, is difficult and impractical for freeway operations. Further, their presence on the roadway could modify driver behavior, producing biased results.

DATA COLLECTION

The primary objectives of the headway study were to examine the use of VADAS for automatic image processing in a high-speed traffic environment, as well as to fit theoretical headway models to freeway traffic data. In a series of experiments performed in 1986, data were collected using video cameras located at particular vantage points, e.g., freeway overpasses and pedestrian bridges on the Mulgrave Freeway in the east-

ern suburbs of Melbourne. The freeway section was that between the Springvale Road and Wellington Road interchanges (Figure 1). This section has two lanes for each direction of traffic flow. It was selected because of the considerable variations in the level and quality of traffic flow and the presence of the footbridge, which provided a convenient and inconspicuous high point to record the flow. This freeway section represented uninterrupted traffic, with no external influences on the flow.

This survey involved the collection of nine sets of data spanning three separate time periods on each of three days. The first six sets were about 15 min long, while the remaining three sets were about 30 min long. In all, a total of about 3 h of traffic flow was recorded, covering traffic demand from 8 a.m. to 5 p.m. Table 1 shows reference information for the data sets. Only the first six data sets were used in the analysis described in this paper.

The data covered a wide range of medium to high flows. Young et al. (*11*) proposed that different models should be applied to traffic under different flow regimes and that these regimes could be defined in terms of the flow-density diagram for a given road. They proposed three separate regimes: low, medium, and high flows.

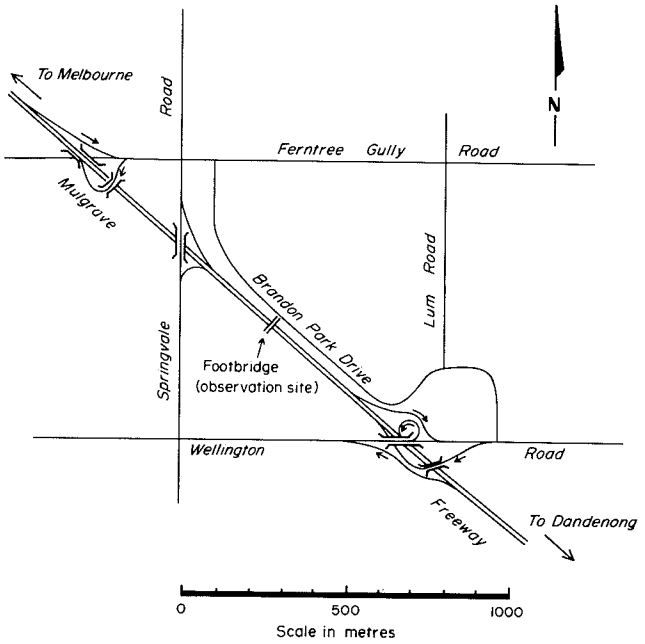


FIGURE 1 Freeway study area.

TABLE 1 DESCRIPTION OF MULGRAVE FREEWAY DATA

Data Set	Date	Time	Direction
1	7/8/86	9:14 a.m.	to city
2	7/8/86	12:00 p.m.	from city
3	7/8/86	3:36 p.m.	from city
4	20/8/86	10:24 a.m.	from city
5	20/8/86	11:46 a.m.	to city
6	20/8/86	2:21 p.m.	to city

DATA EXTRACTION

Two methods were used to extract numerical data from the video recordings. The first was manual extraction, and the second was automatic data extraction using VADAS. The manual method was included to provide a basis for evaluating the automatic procedure.

Manual Analysis

A stopwatch timer was included on the video record to assist in manual data extraction. The recorded tape was played at a quarter speed. An arbitrary mark was identified on the screen (e.g., a corner cube reflector or ceramic lane delineator). Each time a vehicle passed the mark, the time was noted. Different observers watched different lanes. The whole process took up to 90 min for 15 min of video recording. This process required extensive manipulation and was time consuming. The entire processing time (not including the time subsequently spent in data analysis) exceeded 3 hr (6 person-hr) per data set.

VADAS Analysis

Figure 2 shows the configuration of 11 detection points, in two bands across the carriageway, used in the data extraction. The detection points were a fixed distance apart, and the distance corresponded to that between consecutive corner cube reflectors in adjacent groups of the raised pavement markers that form the lane delineators for the freeway. Knowledge of this distance permits speed information to be computed, once the validity of the techniques has been properly established. The present discussion deals solely with headways.

Care was taken when positioning the detection points to ensure that any change in ambient light conditions, e.g., cloud cover or shadows from bridges or foliage, would affect all points equally.

Adjusting the sensitivity threshold for each detection point was vital. The sensitivity was set sufficiently high to detect as many vehicles as possible. This was found to be especially important for those high-speed vehicles with similar luminance to that of the road surface. However, when the sensitivity was too high, some vehicles were detected more than once. This phenomenon was especially pronounced for light-

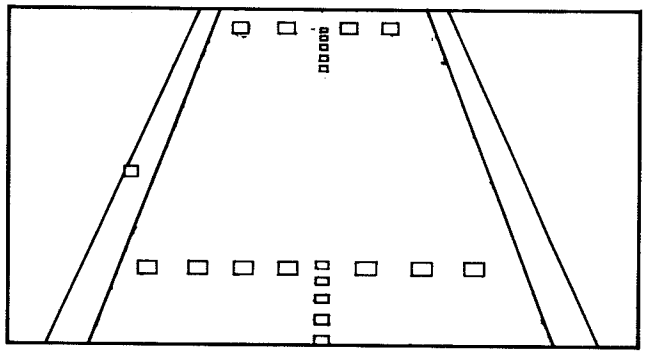


FIGURE 2 Location of VADAS detector points.

colored vehicles, which could trigger the detection point perhaps three times (hood-windscreen-roof). An optimum sensitivity was thus chosen, on the basis that all vehicles would be detected. False points triggered were then detected in the data-editing phase, using a series of logical tests of the time differences between successive events at the one detection point. The method is not perfect. Some false detections may still have been accepted while some genuine detections may have been missed. However, the procedure should be at least as good as manual observations based on viewing the video record at normal speed.

When the video signal at each detection point exceeds the upper and lower limits, an output is produced and read by the VADAS microprocessor (*3*). A preset time delay is used to prevent further detections at the detector. This was set at 70 ms for the freeway study to ensure that all vehicles could be detected. This, however, led to some multiple detections, as indicated above. Some problems were noted in recording motorcycles when these vehicles traveled between the detection points.

The time taken to extract the data from all the video tapes was approximately 5 hr. This involved an initial period of around 4 hr to set up and become proficient with the equipment. The processing of the tapes was completed within 1 hr. Obviously, if data was regularly processed using this technology, the time required for familiarization could be substantially reduced and the data extraction would become quicker. Further time was required for editing the data. The editing was carried out using a series of programs developed by Troutbeck and Dods (*3*). The editing for all the headway data took approximately 1 hr. It was therefore apparent that VADAS offers considerable time savings in headway data collection once users become familiar with the operation of this equipment.

The VADAS headway surveys described here were compared with manual methods. The comparison indicated a VADAS accuracy level of approximately 90 percent, which was generally attributable to the poor quality of the video camera used for the data collection. The accuracy of the VADAS system has also been tested in intersection surveys with results of approximately 95 percent being obtained (*3*).

RESULTS

The data extracted in this study were subjected to three separate analyses: a general view of the traffic flow for each data set and each lane and an examination of lane usage characteristics.

Traffic Flow Characteristics

Comparisons were made of east- and westbound flows for all nine data sets and the range of traffic flow conditions. The results of this comparison are shown in Table 2. This table shows a range of lane flows from 632 vph/lane to 1,476 vph/lane and a range of total flows from 1,420 vph to 2,676 vph. An important result is the consistent observation of small minimum headways (approximately 0.3 to 0.5 sec) for single-lane flows. No distinguishable differences were found between the characteristics of traffic in each flow direction. Figure 3 shows the coefficient of variation of headways plotted against the traffic flow rate, with the observations scattered about the horizontal line representing a unit coefficient of variation. There is a possible suggestion of differences between lane flows, with the outer (slower) lane showing coefficients typically less than unity, while the inner (faster) lane shows coefficients generally exceeding unity. More data are needed to pursue this observation, but it is interesting to reflect that a unit coefficient of variation is one analytical property of the negative exponential distribution. A further comment on Table 2 is the strong indication that trucks were largely confined to the outer lane.

TABLE 2 SUMMARY DATA FOR THE HEADWAY DATA SETS FROM MULGRAVE FREEWAY

	Data Set Number					
Description	1	2	3	4	5	6
Full Carriageway						
Sample size	550	356	668	364	381	375
Flow rate (vph)	2,200	1,424	2,672	1,456	1,524	1,500
Truck percentage	7.8	12.1	5.2	11.2	9.4	8.5
Minimum headway (sec)	0.0	0.0	0.0	0.0	0.0	0.0
Maximum headway (sec)	11.5	17.4	10.1	17.9	22.0	17.8
Outer Lane						
Sample size	245	198	300	204	206	209
Flow rate (vph)	984	796	1,200	820	828	840
Truck percentage	15.9	19.1	9.7	18.5	13.0	13.3
Minimum headway (sec)	0.3	0.2	0.2	0.5	0.3	0.6
Maximum headway (sec)	14.5	20.8	12.0	18.5	22.0	17.8
Median Lane						
Sample size	305	158	368	160	175	166
Flow rate (vph)	1,220	632	1,472	640	700	664
Truck percentage	1.3	3.2	1.6	1.9	5.1	2.4
Minimum headway (sec)	0.3	0.2	0.2	0.3	0.3	0.4
Maximum headway (sec)	18.2	39.5	19.9	41.5	40.0	45.9

Distribution of Lane Flows

Mahalel and Hakkert (*8*) reported Israeli data that showed that a majority of vehicles used the outer lane at low volumes. The reverse occurred at high volumes. This tendency has been observed in other studies from various countries, e.g., Chen (*9*), Lenz and Hotop (*12*), and Rorbech (*13*). A similar tendency was noted in this study. One possible explanation is given below, in terms of bunching tendency. Figure 4 shows the split of traffic between the two lanes of the carriageway. It suggests that equal lane flows occurred at about 1,550 vph. This is consistent with the results of the other studies cited above. Further data are needed to fully test this finding.

Bunching Tendency

An examination of the coefficients of variation in headways (Figure 3) indicates some apparent differences between the patterns of traffic flow in adjacent lanes. In the outer lane, the coefficient was typically less than unity, while it exceeded unity for the median lane. This difference occurred at all levels of traffic flow. Figure 3 shows these results. There was a greater tendency for bunching in the median lane, with a bunching tendency exceeding that of Poisson flow. On the other hand, the vehicles in the outer lane were underdispersed in relation to Poisson arrival patterns. Similar results were reported by Mahalel and Hakkert (*8*). This difference is probably due to the presence of slower-moving vehicles (those with speeds well below the mean speed) in the outer lane. Under medium-to-heavy flows, there are limited opportunities for lane changing and overtaking. To maintain a reasonably constant speed and not be caught behind slow vehicles, drivers prefer to change lanes at the earliest opportunity while still some distance behind the slow vehicle. The long sight distances available on a freeway facilitate this maneuver.

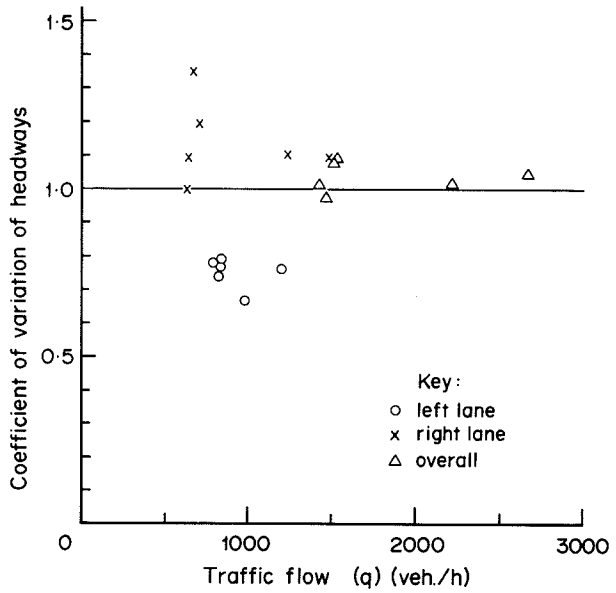


FIGURE 3 Relationship between coefficient of variation and traffic flow.

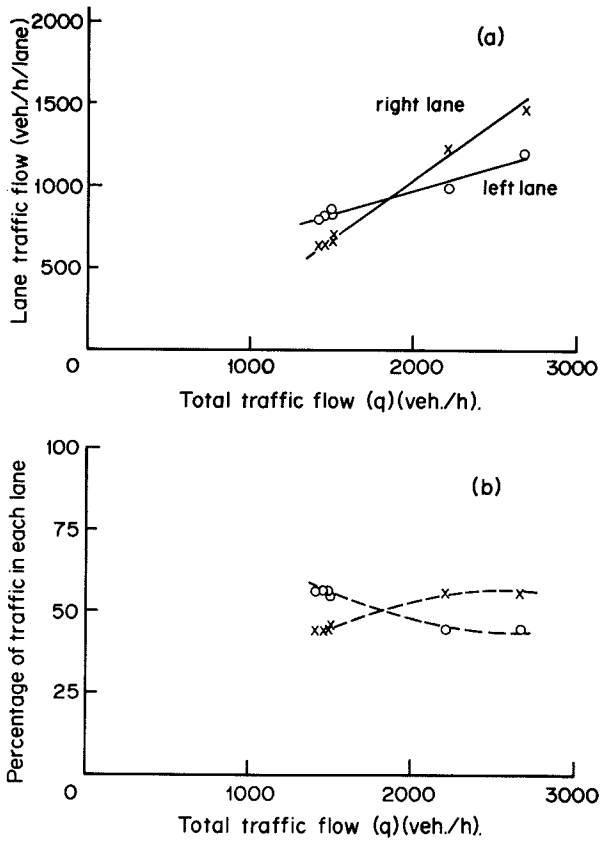


FIGURE 4 Division of traffic lanes.

This hypothesis is supported by the shift in the distribution of vehicles towards the median lane with increasing volume (see Figure 4). At higher volumes, the level of service is lowered to a point where the initiation of overtaking maneuvers becomes difficult. A large proportion of drivers then elect to stay in the median lane to achieve a smoother flow.

Distributions of Headways

Researchers have given considerable attention to headways and suggested many theoretical distributions. The problem with many of these models is that they often require complex iterative procedures to fit them to observed data. Rational subdivision and nonlinear least squares techniques have to be used, which are lengthy and at best approximate. As a result, only the common univariate models offered in the TRANSTAT distribution fitting package (*14*) were fitted to the headway data. The effort to estimate the additional parameters of mixed models is justified only if they substantially improve the fit to the data. Distribution fitting using TRANSTAT suggested that the exponential model best represented headways for the combined lanes, while the log-normal model best represented the individual lanes. Since a large number of the headway models fitted the data to a satisfactory degree, the mixed models, other than M3, were not investigated.

These general conclusions were not valid for all data sets, indicating that further work is required to investigate the conditions under which the different models could apply.

SPEED STUDY

The primary objectives of the speed study were to determine mean desired circulation speeds on parking circulators and aisles, as well as comparing two methods of speed data collection: radar and video.

Initial understanding of the specific nature of the data to be collected is an important component of any traffic survey (*15*). This is crucial in speed studies since the term "vehicle speed" is by no means a simple concept. It can refer to space mean speeds, design speeds, safe driving speeds, desired (free speeds), or spot speeds. In this study, several speed-related concepts are discussed. Space mean speed is the average speed of vehicles taken over a length of road at a particular point in time, whereas time mean speed is the average speed of vehicles taken at a particular point in space over a period of time.

The desired speed is the speed a driver will adopt if not hindered by other vehicles. It is often referred to as the free speed and is measured by observing the speeds of isolated vehicles and vehicles at the head of a platoon of vehicles, since vehicles in a platoon cannot adopt their desired speed. Methods to overcome bias in measuring free speeds are described by Taylor and Young (*15*). The definition of desired speed provides the impression that drivers are not hindered in their desire to travel at a particular speed. Drivers in parking lots are constrained by the geometric characteristics of the roads and their major purpose (looking for a parking space, leaving the parking lot, etc.). Therefore, they may never reach their desired speed. Possibly a more appropriate term would be "desired circulation speed." The desired circulation speed is the speed a driver will adopt when moving through the parking system. Vehicles that are not hindered can adopt this speed.

The spot speed is the speed of a vehicle at a particular point along a road. This speed can be measured using a radar gun. The interval speed is the speed over a small length (or interval) of road and can be determined using video-based equipment. It is calculated by dividing the time taken to traverse a small length of road by the length of the road. The average spot (interval) speed is the arithmetic mean of the individual vehicle speed measurements. Since the speed measurements in this study are taken at a spot (or interval), the average speed is a measure of the space mean speed. However, since the characteristics of the roads considered limit overtaking, the average spot speed is also a reliable estimate of the time mean speed.

A REVIEW OF GENERAL SPEED STUDIES

Knowledge of the speed of vehicles in parking lots is important in ascertaining the safety of the system and the effects of speed-control devices, and as an input into parking lot design models (*11*). Little attention has been focused on this parameter. This is difficult to understand given the considerable conflict between pedestrians and vehicles in parking lots, which is exacerbated by the relative speeds of vehicles and pedestrians. No reported studies of vehicle speeds in parking lots could be found. It is, however, informative to review studies of speed on other parts of the transport network (*16*).

In summary, considerable effort has been directed at the determination of vehicle speeds on urban arterial roads, rural roads (*17*), and residential streets (*18*). The impact of road design standards has also been investigated (*19*). All these studies used radar to collect the data.

SPEEDS IN PARKING NETWORKS

As discussed, no speed studies have been directed at vehicle speeds in parking lots. Therefore, before considering the studies presented in this paper, it is necessary to introduce the components of a parking lot and the concept of desired speed as it relates to vehicles in parking systems. Parking road networks can be likened to the urban road network where a hierarchy of roads are defined, each with different functions (*20*). The lowest level of the hierarchy in a car park is the aisle (*A*). It is the area of pavement used by cars to gain access to parking spaces. The upper end of the hierarchy is defined as the circulator (*C*). It is a roadway that provides access to parking aisles from ingress and egress points of the facility. Parking should not be provided on circulator roads.

As mentioned previously, drivers in parking lots are often constrained by other activities (e.g., looking for a parking space or exit). These constraints may not allow drivers to reach a desired speed. Therefore, this paper introduces the concept of circulation and refers to the desired circulation speed in parking lots. In parking lots, many of the roads are short, and a driver may spend all the time accelerating along the road rather than adopting a desired constant speed. The determination of vehicle speeds in parking facilities is complicated by the need to decelerate to park and carry out other maneuvers. The concept of a desired circulation speed may therefore not be appropriate since the vehicles may never reach a steady speed. Thus, the first step in this study was to study the character of speed profiles of vehicles traveling along roads in parking facilities. The determination of the steadiness of the vehicle speeds was studied in both circulators and aisles.

DATA COLLECTION AND EXTRACTION

The site chosen to conduct the data collection was Vermont South Community Shopping Centre, located on Burwood Highway opposite the Australian Road Research Centre. This site was chosen because it provided separate circulators and parking aisles with unimpeded views of at least two aisles from the video camera. The study was carried out on a Friday morning, since this provided high volumes of traffic. In fact, the car park was around 70 to 90 percent full when the readings were taken. The survey was carried out on a day with perfect weather conditions: sunny with light winds. This was advantageous since poor driving conditions may have affected the results.

Data was collected in two parts of the car park (Figure 5): a major circulator (Site 1) and a parking aisle (Site 2). The video recording system enabled changes in speeds along routes to be measured. By contrast, it is difficult to measure vehicle speeds at particular points along routes using a radar gun. Although radar can be used in conjunction with treadle systems to obtain speed profiles (*21*), the video allowed the

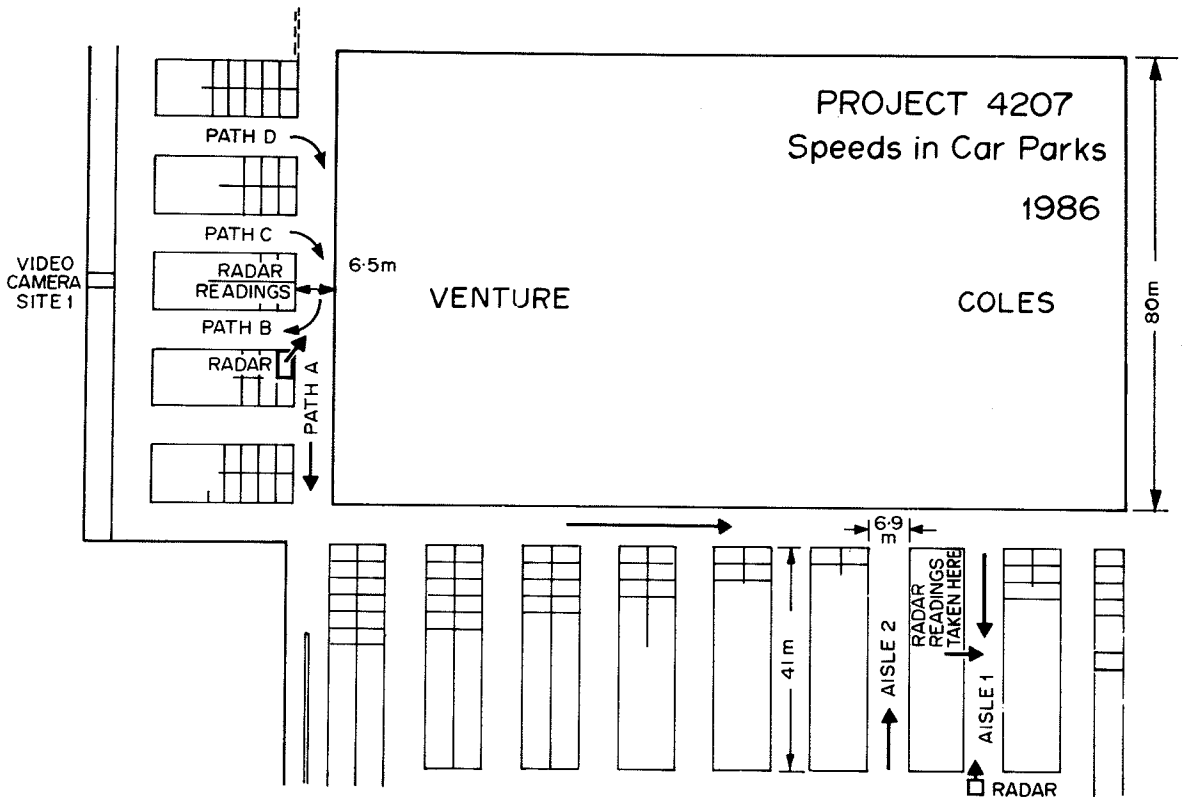


FIGURE 5 Vermont South parking lot layout.

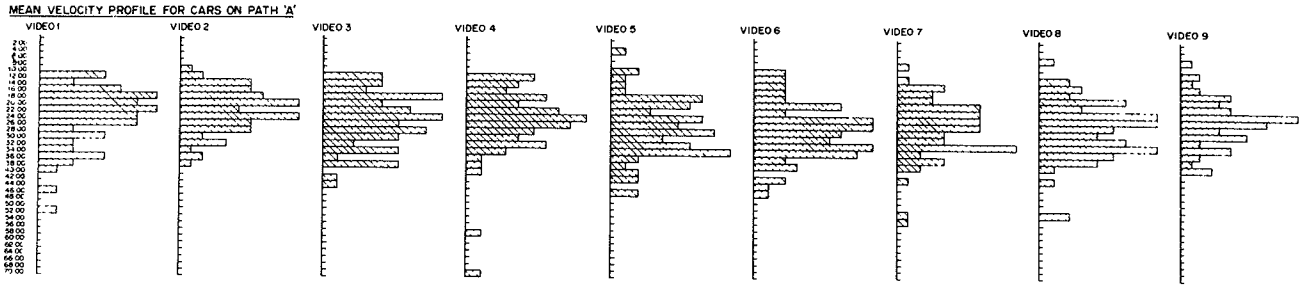


FIGURE 6 Speed profiles for vehicles traveling along circulator (Path A).

flexibility to adjust the location of detectors and to relate vehicle speeds to trip characteristics (e.g., whether or not the vehicle was parking). The trailer housing the video camera was positioned and stabilized perpendicular to the vehicles' movement path. The camera was then focused and raised to the top of the mast (10 m). After the camera had begun recording, the aisles and circulatory roads were marked by witches' hats at 5-m intervals so they could be located when the video tape was processed.

The extraction of the data from the video could be carried out by a human observer, but this is time consuming. A preferable method is to extract the data automatically. VADAS was used to extract the vehicle headways. In this system, the speeds are calculated by measuring the times the recording points are activated and dividing the difference into the distance between detection points. To determine the desired speed, it was assumed that any headway between two cars of less than 5 sec represented car following. In such cases, the second car was rejected.

RESULTS

The vehicles moving along the circulators were divided into four groups (Figure 5): those vehicles that did not enter an aisle (Path A), those that entered Aisle 3 (Path B), and those that exited Aisles 1 (Path C) and 2 (Path D).

The speed profiles of the vehicles that did not enter an aisle (Path A) are shown in Figure 6. The average (space) speeds of these vehicles fluctuated between 21.2 km/hr and 25.2 km/hr. The general evenness of the speed profiles indicates that the drivers do in fact have a desired circulation or cruising

speed, which fluctuates at about 23.2 km/hr. The vehicles in this lane are moving toward another part of the parking lot or are using the parking lot as an alternate route to bypass the adjacent arterial road system. The speed profiles for those vehicles exiting an aisle (Paths C and D) showed that there is a pattern of steady acceleration toward the desired speed similar to those of vehicles moving through the system (*16*).

Vehicle speeds in aisles varied considerably. There are two types of vehicles that use aisles: those that find a parking space and those that are looking for a space. Vehicles of the first type slow down as they approach the parking space (Figure 7). Vehicles of the second type travel along the aisle looking for a parking space or leave the parking lot (Figure 8). A comparison of the speed profiles for both parking aisles showed that there is a slight difference in both the shape of the speed distribution and the mean velocity for each aisle. This could be due to a number of aisle characteristics, including differences in the grades between these aisles and differences in the proportions of the stalls that were used in each aisle.

An important point to establish in the interpretation of speed profiles is whether or not the driver reaches the desired circulation speed. By looking at the evenness of the profiles, it is reasonable to conclude that desired speeds are reached.

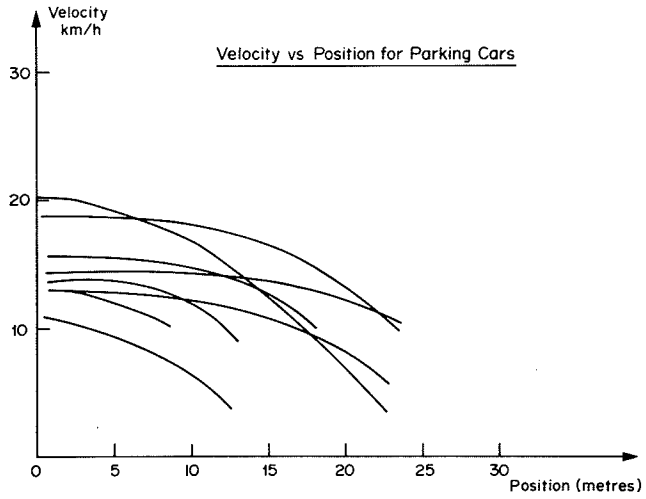


FIGURE 7 Speed profiles for vehicles parking in aisles.

For the aisles, these speeds were around 14 to 15 km/hr; for the circulator, they were around 21 to 25 km/hr.

A difference means test was used to test the difference in speed of vehicles turning from a circulator into an aisle. These two means showed no significant difference, even though they were not on the same road. The mean desired circulation speed of a car turning the corner from a circulator into an aisle was 13 km/hr.

COMPARISON OF RADAR AND VIDEO DATA

Video has a considerable advantage over radar in that it is located perpendicular to the vehicle's motion. Thus, it can record speed readings along a length of road easily, and variations in the speed of the vehicle along the road can be measured. The radar unit is located ahead of the car, which creates problems in locating the position of the vehicle. Further, radar and video measure different speeds. Video measures a speed over a given length of road (space), whereas radar measures it at a particular point (spot). Comparison of the two methods therefore involves the assumption that the speed measured by video is constant over the length chosen. Comparison between the two approaches was required, therefore, to determine if they produced similar results.

Data Recording Technique

Vehicle speeds were also measured with radar using a Kurston H.R. 4 handgun, which had been adjusted to measure low speeds accurately. The radar unit was set up in direct line with the vehicle's path. Two observers were required to collect the speeds of the cars using the radar. One paid particular attention to when the car passed the data collection point and noted the speed at that instance; the other observer recorded the speed of the cars on paper as well as noting the time of occurrence and occasionally noting an identifying characteristic of the car. This enabled the radar measurements to be matched to the same vehicle's speed recorded using the camera.

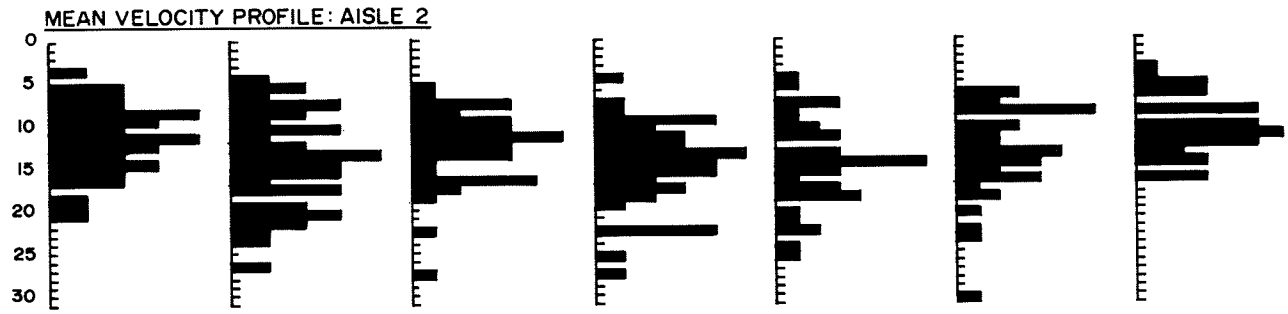


FIGURE 8 Speed profiles for vehicles in aisles.

Data Analysis

Data analysis was carried out by comparing the speeds of the vehicles provided by the radar and the video at the same point. These methods ideally should produce similar results, since they are both measuring the same vehicle speeds. Two tests were conducted to compare the two data sets. One test compared the overall means of each sample, while the other compared the individual data between the two samples.

The comparison of the means is shown in Table 3. The significance of the difference of the means is shown using the probability value. It is the significance level associated with a significant difference in the means. The first probability value indicates the difference in the location of both speed distributions by comparing the sample means. It can be seen that there is little difference at the 1 percent level. The second probability value refers to the paired means test. It indicates whether the difference in the mean differences of individual speeds is not equal to zero. At the 1 percent level this is in fact the case for all but Path B. There is therefore little evidence to support the hypothesis that the individual speeds measured by the two methods are the same. However, the overall mean from the two data sets shows no difference. This is consistent with arterial road speed studies (*3*).

Another comparison of individual speeds was made using regression. If, as assumed, the video and the radar give similar results, the linear regression should give a 45° line with an intercept at the origin. The results of four comparisons are given in Table 4.

The correlation between the two data sets is high, indicating a strong relationship between the radar and the video readings. However, as with the pairs *t*-test, the individual readings are not directly comparable. The video tended to underestimate the low radar speeds and overestimate the high radar speeds. The lack of agreement between the individual readings discussed earlier was therefore confirmed.

Comparison of the radar and the video speed readings indicated that the overall mean speed did not vary but the individual speed readings were not consistent. However, a significant difference in statistical testing may have little practical impact. The magnitude of the differences in the mean speed using the two measuring techniques is 1.17 km/hr with a standard deviation of 3.30. The accuracy of both the radar and the video is approximately 1 km/hr. In the light of the measurement error in the two techniques, a 1.17 km/hr difference is, for practical purposes, small. Due to the small sample sizes of these surveys, further studies would be needed to confirm these initial conclusions.

A comparison of the time taken to collect and extract speeds using video and radar was also undertaken. The radar survey required two people for the duration of the survey (4 hr) to collect the data. The video, once set up, can be left unmanned, and therefore requires only one person for an initial period (approximately 1 hr) to ensure that the trailer and unit are positioned and recording correctly. The data-extraction times for VADAS were similar to that of the headway data (around 6 hr) but would be considerably lower if the equipment was used regularly. The radar data was quickly transcribed into the computer (around 1 hr) due to the small amount of data. It therefore appears that both methods require about the same time to collect and code speed data. The VADAS system would be considerably faster once the users were familiar with the equipment.

AN APPRAISAL OF THE DATA COLLECTION AND ANALYSIS TECHNIQUE

A number of points should be considered when collecting data using the video:

- The position of the sun and its intensity play a significant role in the quality of the recorded image. Glare tends to confuse distant images, and the contrast between the image being observed and the background must be great enough to register a difference when the data is analyzed.
- It may be difficult to determine movements associated with vehicles obscured behind larger vehicles. The positioning of the camera should consider this possibility.
- The area the video can cover is restricted by the lens used in the camera; however, distortion error may result from the curvature of the camera lenses. Consideration should be given to the need to have a site that will allow the video to cover the full study area.
- Parallax errors will also result from the use of the camera since the camera image is two-dimensional while the actual situation is three-dimensional. This two-dimensionality could cause errors in the distance traveled between the two markers on the screen. Corrections must be made.
- Although the extendible mast allows the camera to look down on the traffic flow for near sites, vehicles at distant sites must be viewed somewhere between a top view and a profile. This may cause a problem since the profile of a vehicle is not square and, since vehicles do not always follow the same path, the detector points may be activated by the hood or bonnet, or missed altogether. In such a case, some of the detector points may have to be adjusted to detect the same point on the vehicles consistently.
- Dark cars and motorbikes were very difficult for the data logger to detect because a change in brightness is required to activate the detector. Another effect of dark vehicles is that the blimp may not be activated when the first part of the car passes the blimp.
- Quite often, a car passed through a detector and caused the detector points to flash until the car was totally past. This caused a succession of times to be recorded. In this case, the first time would be considered the correct time.
- The sensitivity of the data logger was continually adjusted by the brightness of the sun. This was a problem in cloudy conditions, for the brightness changed every time the clouds cleared. For the purposes of data logging, it would thus be preferable to collect data on a bright, sunny day. This would also cause more of the darker cars to be detectable due to the sun's reflection off their body panels.

CONCLUSIONS

Video-based data collection systems allow data to be collected that have previously been unavailable using traditional tech-

TABLE 3 COMPARISON OF VIDEO AND RADAR DATA

	Video		Radar		Prob. Value	
Path Type	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Paired Means
A	25.9	7.5	24.7	6.7	0.33	0.0032
B	14.6	4.4	14.5	4.6	0.49	0.4700
C	13.4	3.1	—	—	—	—
D	22.9	5.5	20.3	5.3	0.18	0.0001
Aisles	14.9	4.3	14.4	4.3	0.27	0.0056

TABLE 4 REGRESSION RELATIONSHIPS BETWEEN RADAR AND VIDEO DATA

	Intercept				Slope			
Path Type	Sample Size	Estimate	Standard Error	Sign	Estimate	Standard Error	Sign	Correlation
A	62	3.92	1.34	0.005	0.80	0.05	0.00	0.90
B	14	0.01	1.30	0.99	1.00	0.09	0.99	0.96
D	13	−0.76	1.55	0.63	0.95	0.07	0.25	0.97
Aisles	10	0.01	1.20	0.98	0.95	0.06	0.25	0.95
Total	99	1.58	0.56	0.006	0.88	0.03	0.00	0.95

niques (e.g., headways on multilane freeways and desired circulating speeds in car parks). This technology provides the traffic researcher with an unobtrusive method of observing traffic movements. Video tapes of traffic events contain much more information than manual recorders can collect. The additional information becomes invaluable when modeling complex traffic movements. The flexibility of being able to position the on-screen detector points allows a wide range of traffic parameters to be collected.

The VADAS system offers a reasonably fast method of collecting and extracting traffic survey data. Editing programs are necessary to ensure that the data extracted are accurate. Considerable time savings occur when operators become experienced with this technology. The technology also offers reasonable accuracy compared with traditional techniques, provided weather is appropriate and an adequate vantage point can be found.

REFERENCES

1. M. R. Wigan. Some Investigations of Image Processing Techniques Applied to Road Problems. *Proc., 1st International Conference on the Applications of Advanced Technology in Transportation Engineering*, ASCE, San Diego, Calif., 1989, pp. 424–431.
2. R. Troutbeck and M. R. Wigan. *Workshop on Visual Data Acquisition and Application*. Report ARR 151. Australian Road Research Board, Vermont South, Victoria, 1988.
3. R. Troutbeck and J. Dods. Video-Based Traffic Data Collection. In *Traffic Analysis: New Technology and New Solution* (M. A. P. Taylor and W. Young, eds.), Hargreen Publishing Co., Melbourne, 1988.
4. D. J. Buckley. Road Traffic Headway Distributions. *Proc., 1st ARRB Conf.*, Vol. 1, 1962, pp. 153–183.
5. A. Schuhl. *Poisson and Traffic*. Eno Foundation for Transportation, Saugatuck, Conn., 1955, pp. 59–74.
6. R. J. Cowan. Useful Headway Models. *Transportation Research*, Vol. 9, No. 6, 1975, pp. 371–375.
7. A. J. Miller. An Empirical Model for Multilane Road Traffic Flow. *Transportation Science*, Vol. 4, 1970, pp. 164–186.
8. D. Mahalel and A. S. Hakkert. Traffic Arrival Patterns on a Cross Section of a Multilane Highway. *Transportation Research*, Vol. A 17A, No. 4, 1983, pp.263–270.
9. C-C. Chen. Lane by Lane Speed-Volume Studies on a Six-Lane Freeway. *Proc., 13th ARRB Conf.*, Vol. 13, No. 7, 1986, pp. 1–13.
10. D. L. Gerlough and F. C. Barnes. *Poisson and Other Distributions in Traffic*. Eno Foundation for Transportation, Saugatuck, Conn., 1971.
11. W. Young, M. A. P. Taylor, and P. Gipps. *Microcomputers in Traffic Engineering*. Research Studies Press, Letchworth, 1988.
12. K. H. Lenz and R. Hotop. Some Results Derived From Automatic Traffic Monitoring. *Strassenverkehrstechnik 1*, 1970.
13. J. Rorbech. *The Multilane Traffic Flow Process*. Ph.D. dissertation. Ministry of Public Works, Copenhagen, 1974.
14. R. G. Thompson, M. A. P. Taylor, and W. Young. TRANSTAT: A Microcomputer Program for Fitting Probability Distributions to Traffic Data. *Proc., 14th ARRB Conf.*, 1988 (in press).
15. M. A. P. Taylor and W. Young. *Traffic Analysis: New Technology and New Solutions*. Hargreen Publishing Co., Melbourne, 1988.
16. W. Young. *An Exploratory Study of Speeds in Parking Lots*. Report AIR 466-1. Australian Road Research Board, 1987.
17. M. A. P. Taylor and W. Young. The Distribution of Speeds and Gaps in Traffic Studies. In *Traffic Flow Theory* (P. G. Gipps, ed.), Department of Civil Engineering, Monash University, 1984.
18. M. Armour. Vehicle Speeds in Residential Streets. *Proc., 11th ARRB Conf.*, Vol. 11, No. 5, 1982, pp. 90–205.
19. J. R. McLean. *Two-Lane Road Traffic Flow and Capacity*. Report AIR 359-6. Australian Road Research Board, 1982, chap. 10.
20. *Car Parking Facility: Off-Street Parking*. Standards Association of Australia, AS 2890.1, 1986.
21. M. A. P. Taylor and L. M. Rutherford. Speed Profiles at Slow Points on Residential Streets. *Proc., 13th ARRB Conf.*, Vol. 13, No. 9, 1986, pp. 65–77.

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

Development and Evaluation of a Breadboard Video Imaging System for Wide Area Vehicle Detection

PANOS G. MICHALOPOULOS, ROBERT FITCH, AND BLAKE WOLF

Traffic engineers are constantly seeking new technology and equipment to deal with the problem of urban congestion. Among the most promising concepts available today is the use of video imaging for vehicle detection, automatic surveillance, and extraction of data needed for developing advanced control concepts. A recently developed video detection system is presented in this paper. This system operates on real time, can simultaneously detect traffic at multiple points within the camera's view, and emulates loop detectors. The system was installed and tested both off-line and in real time through taped data and field installations, respectively, and was directly compared to loops. The results suggest similar accuracy levels. In speed measurements, higher accuracies are expected for video systems than for loops. Finally, software is being developed for real-time extraction of traffic parameters, state variables (i.e. queue lengths and size), and measures of effectiveness (delays, stops, energy consumption, etc.) by the same device.

Vehicle detection appears to be the weakest link in traffic surveillance and control. Although accurate equipment is available for detecting vehicle presence on the roadway, it essentially employs technology of the late 1950s, has limited capabilities, presents reliability problems, and often requires massive and expensive installation for true traffic-responsive control. The latter is particularly true in state-of-the-art surveillance and control systems, which often involve large-scale street or freeway corridor networks. Regarding reliability, most cities with mature systems in the United States report that, at any time, 25 to 30 percent of their detectors are not functional or operating properly. Furthermore, discussions with suppliers and manufacturers suggest that often loop detectors, the most widely used detection device, seem to be active but actually produce false or inaccurate actuations. Finally, adverse weather conditions or pavement reconstruction present additional challenges for maintaining these detectors.

Perhaps the most important drawback of existing detectors is their limitation in measuring important traffic parameters and accurately assessing traffic conditions. This is because the technology employed represents a "blind" type of detection; only the presence or absence of vehicles over the detectors can be assessed with high accuracy. Traffic parameters, such as speed and traffic composition queue length, must be derived from presence or passage and require multiple detection, which increases cost and exacerbates the reliability problems mentioned earlier. Furthermore, common detectors (such as loops) do not have surveillance or sufficient vehicle recognition capabilities; most importantly, they are not flexible—they detect traffic only at fixed points. This is an important drawback for traffic control since the detection points should vary with speed, volume, and control objective.

Despite the aforementioned problems, existing detectors cannot be casually dismissed as they represent proven technology that will continue to serve its purpose in the foreseeable future. However, recent advances in image processing, electronic cameras, special-purpose computer architectures, and microprocessor technology have made the machine vision alternative for vehicle detection attractive, economical, and promising. A machine vision system for vehicle detection consists of an electronic camera overlooking a long section of the roadway. A microprocessor or a larger computer determines vehicle presence or passage from the images received by the camera, and derives other traffic parameters, preferably in real time. Vehicle detection can be obtained at specific points of the roadway while other traffic parameters can be derived by analyzing the images of the entire roadway scene. In the system described here, the microprocessor alternative was selected.

The advantages of vehicle detection through image processing are many, as a video detection system (VIDS) has multitasking capabilities. While performing its basic detection functions, it could simultaneously derive traffic measurements locally (using a microprocessor) or at a central location, perform surveillance functions, act as a vehicle counting and classification station, detect incidents and alert a human operator, and recognize special vehicles (ambulances, fire trucks, buses, etc.). There are, of course, other secondary tasks that a VIDS system can perform, such as (a) collecting and preprocessing data to be used in conjunction with existing traffic software packages, (b) revealing the nature of an incident by transmitting images of the scene, and (c) recording data for accident analysis, reconstruction, etc. Finally, it can be used as an evaluation device for measuring and assessing the flow quality or deriving measures of effectiveness for traffic studies.

An imaging detection system does not disturb the pavement and should, therefore, improve reliability, especially during reconstruction operations. Additionally, it can detect traffic at multiple spots of the roadway, within the camera's view, thereby becoming cost effective. For instance, in a previous

P. G. Michalopoulos and B. Wolf, Department of Civil Engineering, University of Minnesota, Minneapolis, Minn. 55455. R. Fitch, Triple Vision, Inc., 777 Harding St., Suite 4, Minneapolis, Minn. 55413.

feasibility study performed in Minnesota (*1*), it was estimated that fully instrumenting an intersection with a VIDS system (4 cameras, 1 microprocessor) would cost less than loop detectors, assuming that at least 3 loops/approach are required. Furthermore, simultaneous detection at 30–40 points using one or more cameras is possible. It was also estimated that the VIDS system design presented here would save 35 percent in maintenance costs and reduce the man-hours required by about 70 percent. Further savings could be realized if the same microprocessor also performed control functions (a viable alternative), thereby eliminating the need for a separate controller.

The flexible detection configuration of VIDS, combined with its ability to extract traffic variables difficult to obtain by conventional detection devices, suggests that the system should be particularly effective for automatic surveillance and control of saturated networks.

Because of these advantages, research on a cost-effective image-processing system for vehicle detection began to evolve during the mid-1970s in the United States, Europe, and Japan. In 1984 research at the University of Minnesota started through projects funded by the Minnesota Department of Transportation and later by the Federal Highway Administration (FHWA). As a result, a real-time multispot breadboard system was just completed, installed, and tested in several actual situations. Placement of the detectors with a mouse device can easily be accomplished by the user in minutes, by placing detection lines on a television monitor in any desirable configuration. Once these "pseudo-detectors" are placed, the system generates presence and passage signals compatible to loops, measures speeds, and generates essential traffic parameters such as volumes, headways, and occupancy. Furthermore, the system allows visual inspection of detection results along with actual traffic conditions for validation purposes and optimization of detector placement. The latter can easily be changed as often as desired either manually or automatically. Special algorithms for treating artifacts such as rain, snow, shadows, pavement reflections, etc., were developed. Also, the system can operate under both day and night conditions. Finally, any ordinary video camera used for surveillance purposes can be hooked to the breadboard system, i.e., no special-purpose cameras are required, although it should be evident that better-quality cameras, without blooming or streaking characteristics, improve the system's accuracy and effectiveness. Unlike earlier experimental units, the breadboard system not only operates in real time and deals effectively with all the aforementioned artifacts (rain, shadows, snow, etc.), but also operates under all traffic conditions. This is an important attribute since background compensation, when congestion sets in, is very difficult and has not been previously researched.

Following the initial algorithm laboratory development, testing was performed for algorithm optimization using videotaped data. Subsequently, the system was installed at the freeway surveillance and control center of the Minnesota Department of Transportation in Minneapolis and tested against live data from several cameras. The results are very encouraging and they suggest performance comparable to loops. For this reason, following additional testing, prototype fabrication is planned in early 1989.

In this paper, the breadboard system is described along with the facility that was also developed for quick algorithm testing and optimization. A brief description of the detection methodology is also presented followed by test results.

BACKGROUND

Research on image processing for vehicle detection began to evolve during the 1970s in the United States, Europe, Japan, and Australia (*2*). In the United States, research on this topic was initiated by the FHWA and conducted by the Jet Propulsion Laboratory (JPL) (*3–5*). Although the major objective of this project was individual vehicle tracking, algorithms for vehicle detection and speed measurement were also developed. The imaging system developed by the JPL, called Wide Area Detection System (WADS), was recently evaluated by Sperry Corporation (*6*). In this study, recommendations were made for improving the hardware and software design of the WADS system (*7*). Briefly, although the work performed by the JPL was pioneering, the WADS system was too primitive for practical applications; however, this should be expected at the initial stages of new technological developments.

Several countries are currently funding research and development on this subject including (a) work in England on image processing applied to traffic at the University of Manchester Institute of Science and Technology (UMIST) (*8*), the University of Sheffield (*9*), and the University College, University of London (*10*), (b) a vehicle tracking system being developed in France by the National Research Institute for Transportation and Security (INRETS) (*11*), and (c) a real-time multispot detection system being developed in Belgium by Devlonics, Ltd.

The UMIST project utilized a solid-state camera generating a 100 × 100 pixel/frame image at 8 frames/sec. The camera was mounted at a height of 22.5 m above a two-way highway and data was collected during a period in which illumination varied by a factor of 4. The output was digitized and averaged. An image corresponding to road background in the absence of any vehicle was stored in the memory of the digital processor. During operation, the digitized image was subtracted from the reference image to generate the road background. In the absence of vehicles, the two images should be similar and therefore their difference was due to noise and changes in illumination. A threshold was then used to compare the differences of the two images. The resulting binary image was compressed and stored on video cassette and processed in the laboratory. This system was not implemented in real time and would only work in ideal conditions where the background did not change significantly and where there were no common artifacts, such as shadows and reflections, to cause false detections.

The system currently being developed at the University of Sheffield operates under the assumption that the roadway background does not change significantly over a period of 1 min, which is considered to represent ideal conditions. This approach is highly prone to errors due to illumination changes, shadows, and reflections. At the University College, University of London, the focus is on implementing vehicle tracking on real-time, parallel image processing computer architectures. This vehicle detection approach requires a background to be manually sampled, which is impractical in field situations, so work is under way to automate this estimation. Once objects are separated from the background, features needed

for vehicle tracking are extracted. These features would result in tracking not only vehicles but also common artifacts which would generate a substantial number of false detections.

In France, INRETS is also developing a real-time vehicle tracking system. The system automatically determines the roadway lane positions and then tracks vehicles down each lane through the entire camera field of view. The major problems with this system are that it can lock onto common artifacts, as well as have problems tracking vehicles through various background changes (e.g., asphalt/cement boundaries, building shadows) and in congested situations where the tracking mechanism breaks down. Finally, some observed test sequences indicated that the primary objects being tracked were dark shadow areas under vehicles and not the vehicle itself.

Recently, Devlonics, Ltd. of Belgium has advertised a real-time system that can accomodate up to 4 detection spots, each covering a 10-m lane area. The approach taken, which originated in cooperation with the Catholic University of Louvain (*12*), was to detect vehicles relative to an automatically determined reference background and track their movement through the 10-m area so as to also determine vehicle speed. Little detailed information about the approach taken is available; however, it was learned that vehicles must move through the 10-m area in less than 2.5 sec or they become part of the background signal. Furthermore, a microcomputer is needed to implement the detection for each spot, so the full 4-detection-spot system requires four microcomputers. Preliminary testing of the system in the Netherlands revealed detection problems in the presence of rain, shadows, congestion, and other artifacts. Additionally, the system does not seem to operate in real time but with a 5-sec constant decision delay, which is too long for critical intersection control applications.

The Japanese government sponsored the Institute of Industrial Science, University of Tokyo, research on measuring traffic flow using real-time video processing (*13–16*). The non-imaging sensor designed by Shigeta and Ooyama is of interest (*17*, *18*). The sensor is an array of photoelectric elements with geometry designed to match the perspective distortion produced by the camera installed at a specific height and angle of view. The photoelectric elements have a spectral response with a maximum of 930 nm that is thought to be optimum during the complete 24 hr day/night cycle. Detection is produced by illumination differences which are discovered by pairs of sensors. The distance between these sensors is known, and by measuring the time difference between detection by the first and the second element in a pair, the speed of the vehicle can be estimated. This system was tested in Tokyo for two years. The Shigeta-Ooyama system, which is the most cost effective, is not truly an imaging system and cannot be extended beyond simple detection as it requires fixed roadway placement geometries and has only fixed and discrete detection points in the field of view.

The Australian Research Board has developed a real-time vehicle presence system (*19*) that allows placement of up to 16 detection spots at any position in the camera field of view via front panel thumbwheel switches. To determine the background level, an additional reference detector is required, which must be placed in an area free of vehicles. This reference is compared with the detection spot outputs and, when fixed thresholds are exceeded, a vehicle is detected. Each detection spot has a manual offset adjustment to compensate for the difference in road surfaces between the reference and detector areas. The approach works adequately for ideal situations, but the system cannot distinguish the difference between vehicles and major artifacts such as vehicle shadows, reflections, and building shadows. Also, since the detection algorithms are hard-wired, there is no flexibility to reprogram and improve the system.

Experience with machine vision over the past 5 years suggests that despite the impressions generated throughout the literature, a reliable, fieldable, real-time multispot vehicle detection system is still lacking. The major problems with existing systems that have been addressed and resolved by the breadboard system are as follows:

1. The inability of existing systems to automatically adapt to a wide variety of backgrounds without reference marks prevents them from running reliably or autonomously. A unique approach to estimating the background at the detection spot was therefore developed; this allows automatic adjustment to any uniform or nonuniform road surface without operator intervention at startup or while running.
2. The operation of prior approaches in the presence of common artifacts such as shadows, illumination changes, and reflections has resulted in these systems having high false alarm rates. In the system presented here, these problems were resolved using a vehicle-signature-based detection approach that can differentiate vehicles from these artifacts.
3. Congested traffic conditions and stopped vehicles have caused the loss of the vehicle and erroneous background estimation in prior approaches. The VIDS system allows vehicles to stop for much longer periods of time without "blending" into the background.
4. Most existing systems only support a small number of fixed position detectors, and not the arbitrary placement of any type of detector in any configuration within the camera's field of view. In contrast, using the VIDS system one is able to place detection spots of any number, size, and shape anywhere in the camera's field of view, and one can reposition these spots dynamically under software control. This is accomplished without requiring the camera to be placed at a fixed geometry (e.g., height or angle).
5. Existing approaches to cost-effective real-time implementations have resulted in oversimplification of the sensor, hard-wiring the detection processing, or using prohibitively costly processors. Cost effectiveness was a major consideration in the development of the VIDS detection system. The system can operate with standard video cameras; no specialized sensors are needed. The approach taken in developing the VIDS breadboard system allows operation in real time while still being fully programmable. By using an IBM AT-compatible personal computer for the breadboard rather than an expensive image processing platform, it is demonstrated that the final system implementation is cost effective.

DEVELOPMENT OF THE BREADBOARD SYSTEM

The objective of this breadboard system was to fully develop and implement, in real time, the functions of vehicle presence, passage detection, and speed estimation with performance

comparable to magnetic loop-based systems. Emulation of loops was considered an essential step for fully demonstrating concept feasibility. Derivation of other traffic parameters and measures of effectiveness can easily follow from these basic detection functions, considering that this is being accomplished at multiple spots within the camera's field of view. This latter limitation, including the range of operation (currently up to 500 ft), is mainly a function of the camera placement which is nominally assumed to be 40 ft; however, these limits will increase during prototype development. In addition to the multispot detection capabilities, the system is applicable to both freeway and intersections. The detection spots can also be dynamically positioned (without having to reposition the camera) by the system's software or by the user. Thus, the system is not locked into a fixed geometry, i.e., it is a true imaging system. To demonstrate cost effectiveness and feasibility, the breadboard system is based on an IBM AT-compatible personal computer which maximizes the use of existing software while minimizing custom hardware. The final prototype will be substantially more compact.

Detection and speed algorithm development were based on videotaped data recorded in real traffic situations, under a variety of environmental conditions, at both intersection and freeway locations. The main purpose in collecting this data was to capture as many different conditions as possible in order to achieve a high confidence in the algorithm performance. The data was recorded over three years at the Minnesota Department of Transportation (MnDOT) Traffic Management Center (TMC) in Minneapolis from video camera sites in the MnDOT freeway surveillance system. Data were also collected at sites in Michigan, Florida, California, and Maryland. The site selection was based on close proximity to both intersections and freeways; in this manner, collection of both types of traffic data was possible.

Over 70 hr of videotaped data sequences were collected covering a number of lighting, vehicle, traffic, weather, and other conditions, as summarized in Table 1. For initial algorithm development, 50 representative sequences were selected from these videotapes (79,000 images) and transferred to optical video disks. Video disks were used because of their rapid, random access, excellent image stability, step framing capabilities, compact archiving, and low cost.

These recorded data were used in the development and evaluation of the presence, passage, and speed algorithms. The development and evaluation of these algorithms were done on an IBM AT-compatible personal computer. The development facility was first used for off-line algorithm development and evaluation (i.e. verifying the presence, passage, and speed algorithm performance on the recorded video data). The same system was then used to evaluate the algorithms in real time while connected on-line to video cameras in the field at the MnDOT TMC. By using this common facility for both off-line and on-line development and evaluation, the time to transition from the lab to the field was greatly reduced.

TABLE 1 SUMMARY OF TYPES OF DATA COLLECTED

Conditions	Description
Lighting	Dawn, day, dusk, night
Traffic composition	Cars, trucks, semi-trucks, buses, ambulances, motorcycles, bicycles
Traffic flow conditions	Normal, congestion, queues, turning, stopped, multi-lane
Environmental factors	Clear, overcast, fog, rain, snow, haze, abrupt lighting change, hot/cold temps, high humidity
Artifacts	Vehicle shadows, building/sign shadows, cloud shadows, sun glare, wind motion, reflection, occlusion, lens spots

DEVELOPMENT FACILITY AND REAL-TIME SYSTEM

Once again, it should be stressed that although the development facility is based on an IBM AT-compatible personal computer, the final design will be a much smaller, less expensive, and self-contained device. The primary reasons for choosing an AT compatible system were the low cost and high availability of software and peripheral hardware (e.g., video digitizers), the high throughput of the processors, the excellent software support, and the dedication to continuing software compatibility; for instance, the system is fully compatible with the newer 80386-based machines such as the Compaq 386, even though initial development began on an IBM PC based on the 8088 microprocessor.

The development facility hardware system is shown in Figure 1. The personal computer was equipped with a real-time video digitizer that accepts video from any standard video source and converts it into a digital format that can be processed by the PC-AT. The video sources include an optical video disk player, a videocassette recorder (VCR), and a video camera or demodulator. Video output from the digitizer was used to display (or record on a VCR) the digitized results overlayed with graphic information. More output to a digital tape recorder was used to record real-time results from the real-time detection algorithms. The PC-AT also contains a 12 Mhz 80286 microprocessor, 3 megabytes of processor memory, two 1.2-megabyte floppy disks, and a 40-megabyte hard disk, along with graphic display terminals for both monochrome and color graphics and text outputs. For real-time

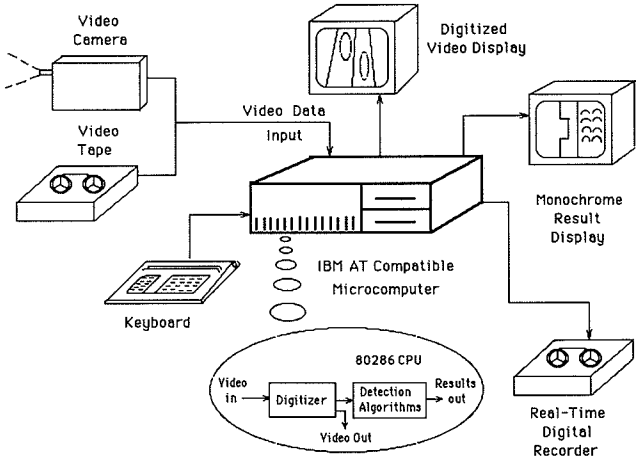


FIGURE 1 Development facility hardware system.

operation, only the real-time video digitizer and one other video image preprocessing circuit card (formatter) are needed in addition to the standard computer.

The development facility software capabilities include

(a) interactively placing multiple detection spots in any number of lanes and at any position along the roadway (in the camera field of view),
(b) ground-truthing sequences of images that are on the video disk (ground truthing is manually determining if a car is present in a detection area for repeated algorithm performance "scoring"),
(c) iterating various processing algorithms on image sequences,
(d) determining algorithm performance automatically,
(e) examining results of processing and experimenting with alternative vehicle detection techniques,
(f) selecting and processing specific video disk image frames or sequences, and
(g) calibrating detector positions to correct for roadway perspective.

ALGORITHM DEVELOPMENT

Presence and passage detection signals were generated at all the detection spots within the field of view. These "pseudo-detectors" were interactively placed by the user at any position in the camera field of view and at any orientation. Examples of possible detector placements are in Figure 2. Pairs of closely spaced detectors were used to estimate vehicle speeds. Detectors across lanes were primarily for vehicle passage, while downlane (or longitudinal) detectors sense vehicle presence. Multiple crosslane detectors were also used for area presence.

Certain spatial and temporal features needed to be extracted for detection and speed estimation for each detector. Spatial features provide information on vehicle signature regardless of the vehicle's speed, while temporal features respond to vehicle motion. Spatial features are the relationships between intensity values across a detector at any instant in time. Temporal features were taken for each detector over a number of time samples (i.e. over a number of image frame times).

The extracted spatial and temporal features were combined using sequential decision processing to generate the background detection and vehicle presence and passage detection signals. Reliable background detection, and its adaption to a wide range of both uniform and nonuniform backgrounds, is a key improvement over earlier approaches; it should be noted that the background is automatically determined by the system—no assumption is made about the road surface signature or its uniformity. Background values were continuously updated and a special logic was developed for updating when a vehicle is present. This logic prevents the background from being "lost" or falsely determined in congested or stopped vehicle traffic situations. Given reliable background estimation, vehicle detection is determined by differences relative to the background level.

Edge-based features are customarily used in research to detect motion. They provide good separation between vehicle signatures and those of common false-alarm-generating artifacts such as shadows (vehicles, clouds, fixed objects), illumination changes (camera AGC, transition periods, lightning), and reflections (headlights, sun glint). As a result, it should be possible to suppress most of the false alarms associated with these common artifacts.

In addition to vehicle detection, individual vehicle speed was also measured. This was accomplished by using pairs of closely spaced detectors and measuring the time it takes the vehicle to move between the detectors. This is shown conceptually in Figure 3. By estimating the time (t) that it takes the vehicle to travel from the first detector ($D1$) to the second ($D2$) and knowing the distance between the detectors (d), the speed (s) was easily estimated. For higher speeds, this time can be reliably measured using the difference in time between the passage signals generated by the vehicle detection algorithms (similar to a speed trap used with loops). But in some situations, such as in congestion, the passage signal generation was not reliable enough to generate an accurate speed measurement. In fact, at lower speeds the accuracy of passage signal generation was increased if speed was used by the vehicle detection algorithms. As a result, a speed estimation technique that works independently of vehicle detection had to be developed.

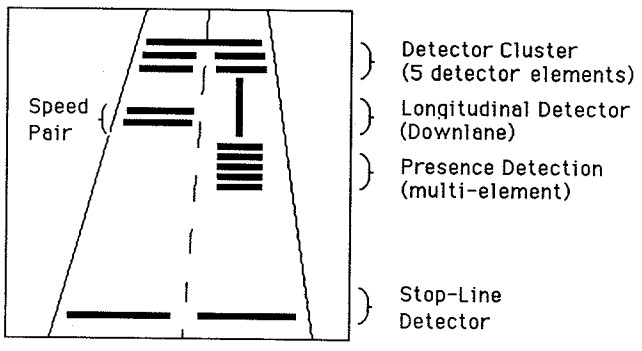


FIGURE 2 Example detector configurations.

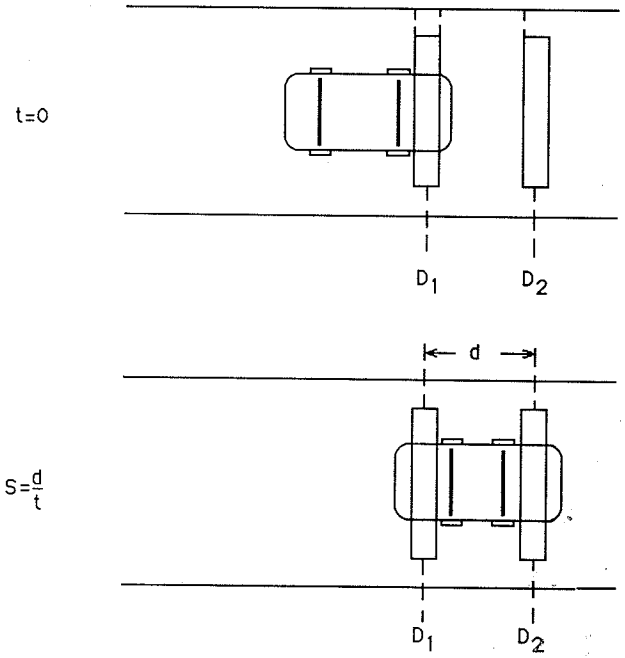


FIGURE 3 Speed estimation using doublets.

Briefly, the technique compares features generated by each detector and registers them in time using a signal correlation technique. This yields estimation of instantaneous speed as opposed to average speed so that the speed estimate can be used directly for improved control purposes. This technique does not have to rely on either the vehicle detection or background detection outputs.

PERFORMANCE TESTING

The performance of the detection and speed estimation algorithms has been evaluated continuously during the development stage as well as after the completion of the real-time breadboard system. The evaluation was accomplished off-line using the algorithm development facility, and was performed using the video disk sequences; this was key to developing vehicle detection and speed estimation algorithm concepts which could be quickly tested on a large data set (over 79,000 image frames on the video disks). Off-line evaluation is currently being performed to optimize the speed-estimation algorithms, but the vehicle-detection algorithms are now past this stage and are being evaluated on-line in real time. On-line evaluation was made possible when real-time implementation of the detection algorithms was completed. This evaluation was performed in real time on both videotapes and used live data from cameras at the MnDOT TMC. The latter required installation of the breadboard system, which monitors freeway traffic in Minneapolis and St. Paul through 36 camera installations. The breadboard system was connected with several of these cameras to allow visual inspection of the detection outputs. The on-line evaluation allowed quick determination of the system's performance over long time periods on many image frames. On-line evaluation is ongoing as the detection algorithms are being improved; evaluation of other traffic parameters and measures of effectiveness, such as occupancy, time headways, queue lengths, stops, and delay, is planned.

VOLUME PERFORMANCE

Three primary evaluation measures were used to determine the performance of the vehicle detection algorithms: detection accuracy, miss error rate, and false alarm error rate. Detection accuracy is the ratio of the number of vehicles correctly detected by the system to the total number of vehicles (determined by visual inspection). This measure indicates how many vehicles actually present were detected by the system. Conversely, the miss error rate is simply 100 percent minus the detection accuracy and is often referred to as just the "error." The false alarm error rate is the ratio of the number of vehicles falsely detected to the total number of vehicles. This measure indicates how many times the system indicated a vehicle was present when there was none. Note that the detection accuracy and the false alarm error rate do not add to 100 percent as do the detection accuracy and miss error rate.

The real-time vehicle detection performance evaluation allowed direct processing of videotape and live camera data that resulted in extremely fast evaluation on a large data set. At the time of this writing, an all-day live evaluation of the system was just completed at two locations monitored by the MnDOT TMC. At the first location, the camera monitored a freeway section, while at the second an intersection was monitored. At the freeway site, passage (counting) detection was evaluated and presence detection was evaluated at the intersection. The system was run from 7:00 a.m. until 10:00 p.m. on July 15, 1988. Every ½ hr, the system was run continuously for 5 min for both cameras. Over 350,000 image frames containing 8,299 vehicles were processed and scored. These images included a number of significant artifacts including shadows, clouds, rain, congestion, and occlusion; over 1,000 shadows were counted. For 987 vehicles at the intersection, the performance over the entire day was a 98.5 percent detection accuracy (972 out of 987 vehicles), a 1.5 percent miss error rate (15 out of 987), and a false alarm error rate of 5.1 percent (detected 50 vehicles when none were present). For the 7,312 vehicles on the freeway section, the performance was a 91.8 percent counting accuracy (6,712 out of 7,312), an 8.2 percent miss error rate (600 out of 7,312), and a false alarm error rate of 2.3 percent (detected 168 vehicles when none were present).

A plot showing the performance for each 5-min time slice is shown in Figure 4. On each plot the vertical axis represents percentage and the horizontal axis the time of day from 07:00 (7:00 a.m.) to 22:00 (10:00 p.m.). Each plot shows the detection accuracy and false alarm error percentages during each 5-min time-slice every ½ hr; to avoid misunderstanding, it should be reiterated that the miss error rate is defined as 100 percent minus the detection accuracy.

As shown in the plot for the intersection, the detection accuracy ranged from 95 percent to 100 percent throughout the entire day. The false alarm error rate remained below 10 percent except between 08:00 and 09:00 when 60-mph winds caused the camera to move the detection area over a lane marker resulting in false vehicle detections. Wind similarly affected performance around both 10:30 and 17:30. This wind problem can be significantly reduced by using smaller and more aerodynamic camera housings (the one used was an older, bulkier design with approximately a 3 ft^2 "sail" area). Recent camera installations are more rigid and compact and do not seem to move significantly. This wind effect can also be compensated for electronically by measuring the scene shift and translating the digitized image; the prototype will have this capability if camera motion is determined to be a problem.

For the freeway case, Figure 4 indicates that the detection accuracy remained above 90 percent for the entire day, except between 15:00 and 17:00 when congestion caused some pairs of vehicles to appear as one vehicle, which resulted in lower vehicle counts. It should be reiterated that scoring on the freeway was done by counting vehicles (passage). If vehicle presence had been scored instead, the performance would have exceeded 90 percent, even during this period, since presence accuracy remains high even in congested situations. The algorithms for counting in congestion are still being optimized. The false alarm errors shown in the plot were less than 10 percent throughout the day. The majority of the errors occurred between 08:00 and 10:00 (peaked at 10 percent) and were due to large numbers of vehicle shadows on the road.

In addition to evaluating the system all day for 5-min periods every hour, the performance of the system operating for an entire hour at two additional freeway locations with near-capacity traffic was determined. In the first location (Case 1 in Table 2) evaluation was performed for four detection spots placed as shown in Figure 5. The camera was viewing I-35W

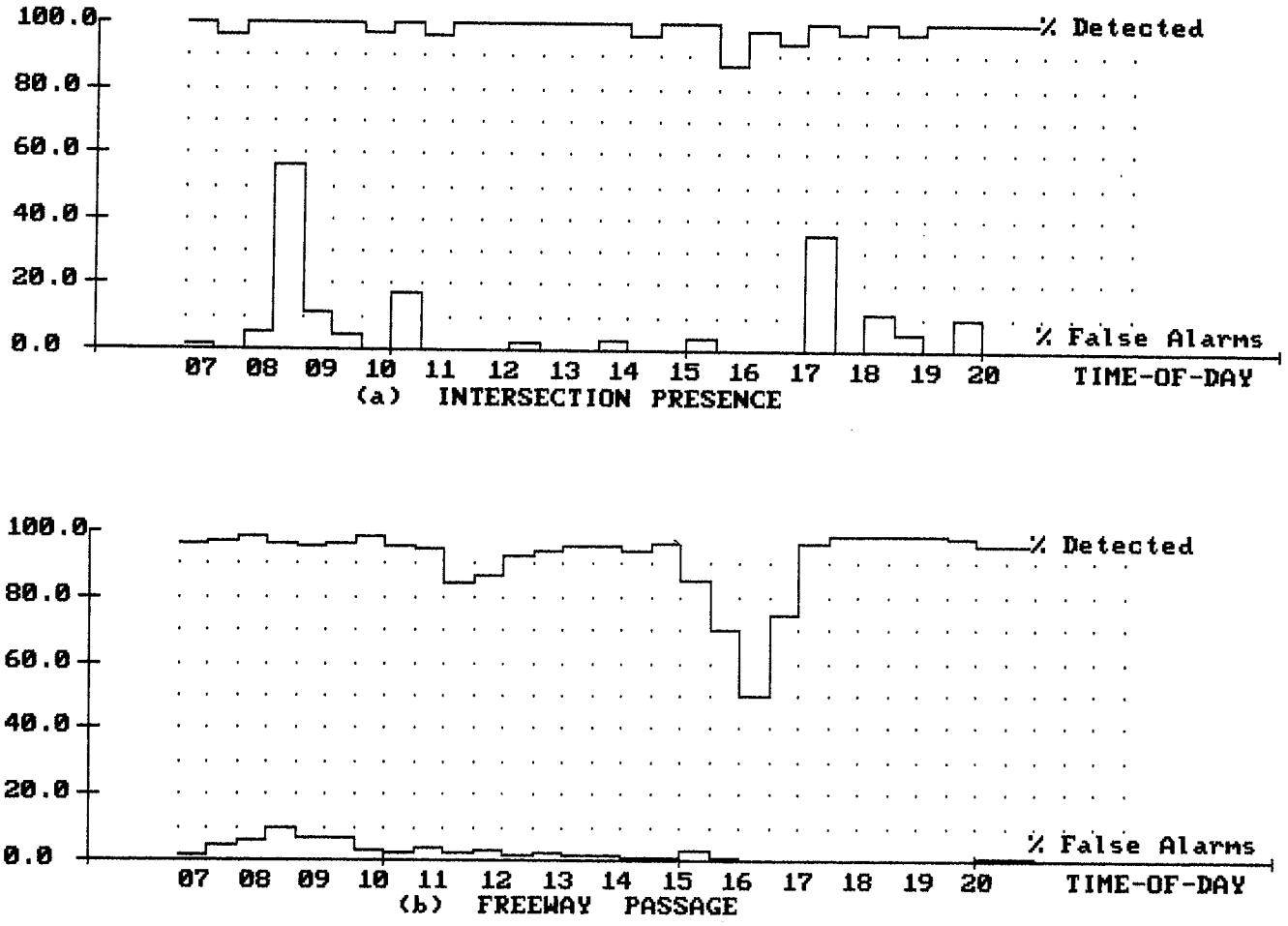


FIGURE 4 All day performance (7 a.m. to 10 p.m.).

in Minneapolis at 60 ft above the road and approximately 150 ground ft from the first detection spot. Vehicle volumes (counts) were determined automatically by the VIDS system for twelve 5-min time slices from 16:15 to 17:15 (rush hour). Actual volumes were determined by manual counting. The average volume to capacity ratio was approximately .86 assuming a capacity of 2,400 veh/hr/lane.

The performance results for each of the four detectors is shown in Table 2 (Case 1), which indicates the actual and measured volumes on each detector, the root mean square error (RMSE), and the mean absolute error. The RMSE for the entire hour was computed from

$$\left[\frac{\Sigma\ (\text{manual count} - \text{VIDS count})^2}{\text{no. of 5-minute time slices}}\right]^{1/2}$$

As Table 2 suggests, the RMSE ranged from 5.68 to 8.89 vehicles per 5-min interval while the mean absolute error ranged from 2.85 percent for Detector 1 to 4.78 percent for Detector 2. These error levels are negligible despite the high percentage of trucks during the period of data collection. Naturally, traffic composition affects performance since trucks and tall vehicles tend to occlude the camera's view, thereby reducing detection accuracy. In the second test site, the error levels were further reduced when truck composition was closer to normal levels. In general, errors increase with distance from the camera and in the farther lanes due to taller vehicles occluding smaller vehicles. Increased range also causes consecutive vehicles to appear contiguous—two vehicles counted as one. However, algorithms dealing with this problem will be developed, and further testing will be performed to determine whether the errors due to this artifact are significant enough to justify further algorithm development.

TABLE 2 PERFORMANCE IN NEAR CAPACITY TRAFFIC

	Volume (Veh.)	RMSE (Veh.)	Mean Absolute Error (%)
Case 1: Manual/VIDS (1 hour)			
Detector 1	1893/1838	5.68	2.85
Detector 2	1888/1850	8.89	4.78
Detector 3	2256/2205	8.04	3.91
Detector 4	2252/2246	7.48	3.18
Case 2: Loop/VIDS (1.5 hour)			
Detector 1	1754/1737	1.68	1.42
Detector 2	2293/2287	6.13	2.91

The current detection performance evaluation approach requires someone to validate all results manually. This is a time-consuming and error-prone process that can only be used in a limited number of cases. To eliminate this manual step, an automated scoring process is being developed to directly

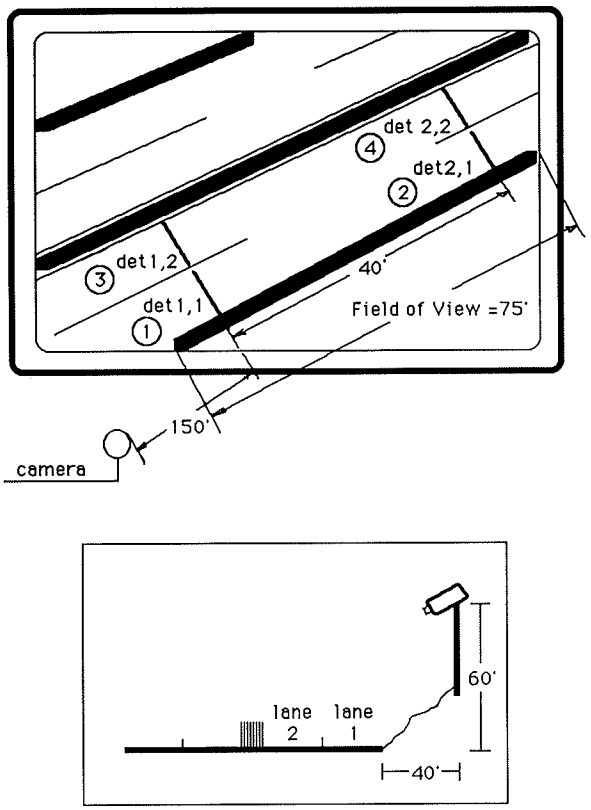


FIGURE 5 One hour evaluation geometry.

compare the VIDS system outputs to those from colocated magnetic loops. The MnDOT TMC has a number of surveillance cameras which can view positions on the freeway where loops are present. The counts from these loops are transmitted back to the TMC where they are logged every 30 sec by the TMC central computer. At this writing, a second camera site was selected (Case 2), and the accuracy of the loop-generated counts was confirmed by manually verifying the loop counts as opposed to counting vehicles from a videotape. This manual verification is necessary since some of the loops are nonfunctional. Subsequently, the VIDS system was run on this same videotaped sequence and the results to the loop counts were compared. The performance for this site is shown in Table 2 (Case 2). This second location is similar to the first location (2 lanes, camera mounted on side of road, Detector 1 in near lane, Detector 2 in far lane) except the camera pole is positioned closer to the roadway (10 ft) and the evaluation period increased to 1.5 hr. Also, the traffic composition contained a normal amount of truck traffic whereas Case 1 contains heavy truck traffic; as a result crosslane occlusion was reduced and system error was lower than expected. The system performance is further evidenced in Figure 6, which presents counts for comparison purposes measured from this second test site. In addition to loop (short-dashed line) and VIDS (long-dashed line) volumes, the manually derived volumes (solid line) that were used for loop verification are shown for both Detectors 1 and 2.

For further testing, the VIDS system is currently being interfaced to the TMC central computer to directly read the loop counts for real-time comparison with the VIDS system output. This will allow the system to automatically score its performance while running continuously and unattended all day, which will allow more extensive performance evaluations.

SPEED EVALUATION

The speed estimation algorithms have been evaluated off-line on the video disk sequences by percent errors and percent misses. The percent error measure is the comparison of the estimated speed to the actual speed; the actual speed is measured visually by counting the number of video frame-times it takes the vehicle to traverse the two detectors. The percent misses is the percent of vehicles that did not register a speed measurement (i.e., the algorithm was not able to estimate a speed). The evaluation was performed on 66 video disk sequences containing 391 vehicles. These sequences included vehicles travelling at speeds from 0 (stopped) to 70 mph in all weather conditions during both day and night. The overall error was 12 percent and the misses were 17 percent. It should be noted that these performance numbers are for instantaneous speeds and not average speeds; the average speed performance would improve with the amount of time the instantaneous speeds are averaged.

The majority of the speed estimation errors were caused by having an insufficient number of samples of vehicles moving at high speed. Since the spacing of the two detectors used for all of these test sequences was 10 ft, a vehicle travelling at 60 mph (88 ft/sec) traverses a 10 ft trap in approximately 4 video frame-times; this could result in an average estimation error of 25 percent (a 1 frame-time estimation error). This error can be decreased by increasing the spacing between detectors. For a 40 ft detector spacing, error rates of 6–7 percent have been measured for vehicles moving approximately 60 mph. In fact, the system's dynamic detector placement capability allows this spacing to be automatically adjusted as the speed estimate changes. The speed estimation misses primarily occurred in sequences with extremely heavy fog (too much noise in the video signal), stopped vehicles (the vehicle reached the first detector but not the second), and sequences where one of the detectors was in a fixed shadow (from a building) and the other in the sun (this caused poor signal correlation). The speed estimation algorithms were only recently improved to deal with these problems and integrated into the real-time system for purposes of performing more extensive on-line evaluation. Preliminary test results obtained suggest accuracies of 94–96 percent. Finally, the signal correlation technique described herein does not require presence and passage signal extraction and in preliminary testing resulted in speed measurement accuracy of 90 percent or higher.

CONCLUSIONS

Despite major worldwide efforts to develop a machine vision system for traffic surveillance and control, a real-time device having the capabilities and performance required for practical applications has been elusive. The system presented here may not have the requirements of a commercial product but, at least in terms of detection performance, it is compatible with existing devices such as loops. Speed measurements, developed only recently, are already very satisfactory and should

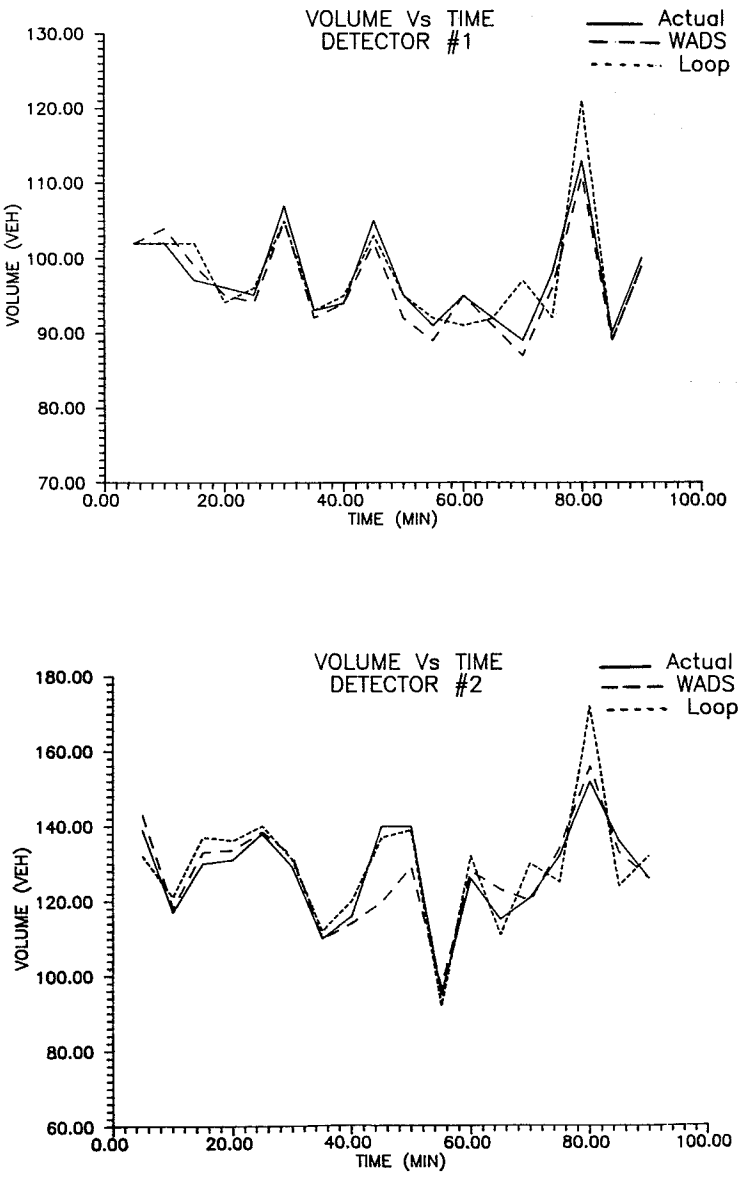


FIGURE 6 Volume comparisons for one-hour period for data derived manually, from loop detectors, and from the WADS system.

improve with further experimentation. It is worth noting that the detection test results presented here refer to individual vehicles rather than averages. Naturally, averaging the measurements over time would result in lower errors, but this method would only mask true system performance and therefore it was not attempted here.

Live demonstrations of the system in its present form to professional engineers and potential users generated favorable comments and lead to the conclusion that the research team should proceed with prototype development and demonstration projects. At least partial funding for prototype development is already available and is expected to be supplemented by early 1989. At this time the prototype development phase should begin; in the mean time, demonstration projects in several states and cities are being considered.

The serious consideration of the system in these locations at this early stage of development is primarily owed to the system's expected impact in traffic surveillance and control. Indeed, the major advantages of this machine vision system lie in the multispot, multilane, wireless detection capabilities. Along with recent advances in image understanding, the system should essentially be transformed to an "electronic eye" for computerized surveillance and control or for automating time consuming and expensive functions (performance evaluation, derivation of measures of effectiveness, etc.). Finally, an imaging detection system can measure traffic variables that cannot easily or accurately be measured by conventional detection devices. For example, queue length and size can be extracted by VIDS without much difficulty; measurement of these parameters requires many loop detectors and cannot be obtained if the queue extends beyond the last detector. Similarly, at this time, density can only be approximated from occupancy; this variable can be measured and more accurately by VIDS. In short, the system should provide the machine

vision link needed to take advantage of recent technological innovations in microprocessors, artificial intelligence, and telecommunications. Clearly, the research effort described here suggests that off-the-shelf technology and equipment for developing a cost-effective video detection system are available today.

ACKNOWLEDGMENTS

The authors would like to acknowledge the cooperation of the technical staff of MnDOT's Traffic Management Center; especially valuable was the assistance of D. Stregge, G. Carlson, R. Lau, C. Harens and G. Bodoczy.

REFERENCES

1. P. Michalopoulos, R. Fitch, and M. Geokezas. *Development of a Visible/Infrared Vehicle Detection System: Feasibility Study*. Department of Civil and Mineral Engineering, University of Minnesota, 1986.
2. R. M. Inigo. Traffic Monitoring and Control Using Machine Vision: A Survey. *IEEE Transactions on Industrial Electronics*, Vol. IE-32, No. 3, 1985, pp. 177–185.
3. E. E. Hilbert et al. *Wide Area Detection System—Conceptual Design Study*. JPL Interim Report, FHWA-RD-77-86. FHWA, U.S. Department of Transportation, 1978.
4. A. P. Schlutmeyer et al. *Wide Area Detection System (WADS)*. JPL publication 83-15, FHWA, U.S. Department of Transportation, 1982.
5. E. E. Hilbert et al. A Sensor for Control Arterials and Networks. *IEEE Transactions of Vehicular Technology*, Vol. VT-29, 1980, pp. 208–215.
6. Sperry Corporation. *Wide Area Detection System (WADS)—Applications*. Interim Reports, FHWA, U.S. Department of Transportation, 1984 and 1985.
7. P. Michalopoulos, R. Fitch, and B. Wolf. *Development of Wide Area Vehicle Detection System (WADS): Image Recognition Algorithms*. Center for Transportation Studies, Department of Civil and Mineral Engineering, University of Minnesota, FHWA, U.S. Department of Transportation, 1986.
8. R. C. Waterfall and K. W. Dickinson. Image Processing Applied to Traffic: 2. Practical Experience. *Traffic Engineering and Control*, 1984, pp. 60–67.
9. A. Houghton, G. S. Hobson, L. Seed, and R. C. Tozer. Automatic Monitoring of Vehicles at Road Junctions. *Traffic Engineering and Control*, Vol. 28, No. 10, 1987, pp. 541–543.
10. N. Hoose and L. G. Willumsen. *Real-Time Vehicle Tracking Using Advanced Image Processing Techniques*. Technical Report. University College, University of London, England, 1987.
11. S. Deucher, J. M. Blosseville, and F. Lenoir. Traffic Spatial Measurements Using Video Image Processing. *SPIE Proc., Intelligent Robots and Computer Vision*, Vol. 848, 1987.
12. W. Maes, A. Barbe, and A. Vits. Development of a Detection System Based on the Use of a CCD Camera. In *Proceedings, Seminar on Micro-Electronics for Road and Traffic Management*, Organization for Economic Cooperation and Development, Paris, France, 1984, pp. 78–86.
13. S. Takaba et al. Measurement of Traffic Flow Using Real-Time Processing of Moving Pictures. In *Proceedings, 32nd Conference on Vehicular Technology*, IEEE, New York, 1982, pp. 488–494.
14. S. Takaba and N. Koishi. Real Time Simulation of Road Traffic Flow. Summer Computer Conference, Vancouver, Canada, 1983, pp. 585–590.
15. S. Takaba. Measurement of Flow of Vehicles and Pedestrians Using Real Time Processing of Moving Pictures. In *Proceedings, International Symposium of Processing and Applications*, Institute of Industrial Science, University of Tokyo, Japan, 1984.
16. S. Takaba, T. Sekine, and R. Hwaug. A Traffic Flow Measuring System Using a Solid-State Image Sensor. In *Proceedings, International Conference on Road Traffic Data Collection*, IEEE, London, England, 1984.
17. K. Shigeta, N. Ooyama, and S. Matsumoto. A New Traffic Flow Sensor Using Photoelectric Converting Elements—Operating Principle and Practical Experiments. *Mechanical Engineering Laboratory*, Technical Reports, Namiki Sakura-mura, Japan, Vol. 35, No. 1, 1981, pp. 21–48.
18. N. Ooyama and K. Shigeta. Area Measurement of Traffic Flow Using Photoelectric Elements. *Proc., 4th IFAC/Int. Fed. for Info. Processing/Int. Fed. Operational Res. Soc. Conference*, Baden-Baden, FRG, 1983, pp. 281–287.
19. J. S. Dods. The Australian Road Research Board Video Based Vehicle Presence Detector. In *Proc., International Conference on Road Traffic Data Collection*, London, England, 1984, pp. 96–100.

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