TRANSPORTATION RESEARCH

RECORD

No. 1365

Highway Operations,
Capacity, and Traffic Control

Highway Capacity and Traffic Flow

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1992

Transportation Research Record 1365
Price: $26.00

Subscriber Category
IVA highway operations, capacity, and traffic control

TRB Publications Staff
Director of Reports and Editorial Services: Nancy A. Ackerman
Senior Editor: Naomi C. Kassabian
Associate Editor: Alison G. Tobias
Assistant Editors: Luanne Crayton, Norman Solomon, Susan E. G. Brown
Graphics Coordinator: Terri Wayne
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 and traffic flow.
p. cm.—(Transportation research record; ISSN 0361-1981; no. 1365)
"A peer-reviewed publication of the Transportation Research Board."
ISBN 0-309-05404-4
1. Highway capacity—Congresses. 2. Traffic flow—Congresses. I. National Research Council (U.S.). Transportation Research Board. II. Series: Transportation research record; 1365.
HE336.H48H53 1992
388.3'142—dc20 92-35073
CIP

Sponsorship of Transportation Research Record 1365

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

Facilities and Operations Section
Chairman: Jack L. Kay, JHK & Associates

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

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

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, 1991.

Transportation Research Record 1365

Contents

Foreword

The papers in 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 theoreticians and practitioners.

The area of highway capacity is receiving considerable attention as a result of the research effort leading toward the next edition of the *Highway Capacity Manual*, which will appear around 2000. The initial group of papers in this Record examine the issue of capacity at stop-controlled intersections, signalized intersections, freeways, rural highways, and urban interchanges.

Traffic flow theory, modeling, and control applications are also examined with papers related to car-following models, vehicle time headways, artificial intelligence techniques, and traffic flow simulation modeling.

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.

Empirical Method To Estimate the Capacity and Delay of the Minor Street Approach of a Two-Way Stop-Controlled Intersection

MICHAEL KYTE, B. KENT LALL, AND NASEER MAHFOOD

The results of a study of 12 single-lane approach, two-way stop-controlled intersection sites in the Pacific Northwest region of the United States are summarized. Traffic flow rate and delay data were collected for each site, and 15-min averages were prepared yielding a total of 107 data points. A capacity model was developed for the minor street approach proposing that capacity is a function of the flow rates and the speed on the major street. A delay model was developed proposing that delay increases exponentially as reserve capacity decreases. Although the data base assembled here is limited, both models appear promising. The results produced by the models indicate that the empirical model approach for unsignalized intersections may provide an alternative to the gap acceptance method currently used in the *Highway Capacity Manual*.

The standard U.S. procedure for evaluating the operation and performance of a two-way stop-controlled (TWSC) intersection is described in Chapter 10 of the *Highway Capacity Manual* (HCM) (*1*). This procedure is based on a method developed in Germany by Harders (*2,3*) and validated with a limited set of U.S. data (*4*). A number of problems have been identified with this procedure (*5–8*), three of the most important of which are (*a*) incorrect capacity estimates at both low and high ranges of major street flow rates, (*b*) difficulty in the estimation of the critical gap, and (*c*) lack of a useful measure of effectiveness.

The objective of the research described in this paper is to propose and test an empirically based method for the analysis of one set of traffic movements, the minor street approaches, at a TWSC intersection. Three topics are covered in pursuit of this objective: the data base developed for this study, the development of an empirically based capacity model, and the development of an empirically based delay model.

DATA BASE

Description of the Sites

Data were collected at 12 sites in Oregon, Washington, and Idaho over a period of 15 days. A total of 26.75 hr of intersection operations was observed. Each site had several common characteristics: single lanes on each approach and adequate sight distances for each minor street approach. Major street speeds at the sites varied from 25 to 55 mph. A wide range of traffic flows was observed at the sites. Observed major street flows ranged from 176 to 1,412 veh/hr; observed flows on the minor streets ranged from 56 to 732 veh/hr. Minor street delays ranged from 5.7 to 75.8 sec/veh.

M. Kyte and N. Mahfood, University of Idaho, Moscow, Idaho 83843. B. K. Lall, Portland State University, Portland, Oreg. 97203.

Data Collection and Reduction

Videotapes were made so that a permanent record was available of the traffic operations at each site. A field of view was established so that traffic flows could be clearly observed on each intersection approach and so that queue activity would be visible on one minor street approach.

Data were reduced from the videotapes using the Traffic Data Input Program (TDIP) software. A new version of this program (*9*) was written specifically for this study so that the characteristics of TWSC intersections could be directly accounted for.

As each videotape was observed, certain events were noted using the TDIP software. These events included the passage of each vehicle through the intersection and the times that vehicles on the minor street approach arrived at the end of the queue, arrived first in line at the stop line, and departed from the stop line.

Variables in the Data Base

TDIP produces two data files that were used to construct the traffic flow and delay data base. The first file consists of hourly flow rates for each 15-min period for each of the 12 vehicle movements through the intersection. The second file includes average time in queue, average time in service, and total delay for each 15-min period for vehicles on the minor street approach.

Table 1 gives the variables produced for each 15-min period of intersection operation. The data base includes 107 data points for each of the variables listed. The capacity of the minor street approach was calculated using Equation 1. The reserve capacity was calculated using Equation 2. The remainder of the variables in the table were directly available from the TDIP files.

TABLE 1 Data Base Variables

Category	Dimension	Variable	Description
Flow Rate	veh/hr	q_s q_o $q_{c,L,i}$ $q_{c,R,i}$	Subject approach flow rate Opposing approach flow rate Conflicting approach from the left flow rate Conflicting approach from the right flow rate
Capacity	veh/hr	Q_s $Q_{s,res}$	Subject approach capacity Subject approach reserve capacity
Delay	sec/veh	d_i	Subject approach total delay

$$Q_s = \frac{3{,}600}{d_s} \quad (1)$$

$$Q_{s,res} = Q_s - q_s \quad (2)$$

CAPACITY

Gap-Acceptance Method

The procedure used in the HCM for evaluating the operation and performance of TWSC intersections is based on gap acceptance theory developed by Harders (*2*,*3*). Harders's model for capacity of the minor street approach of a TWSC intersection is given in Equation 3.

$$Q_s = q_c \left(\frac{e^{-\beta}}{e^{\alpha} - 1} \right) \quad (3)$$

In Equation 3,

$$\alpha = \frac{q_c t_f}{3{,}600} \quad (4)$$

and

$$\beta = \frac{q_c t_g}{9{,}000} \quad (5)$$

where t_g is the critical gap and t_f is the follow-up gap.

Brilon (*2*) notes the following limitations for this equation: (*a*) the major street flow is assumed to be random with headways exponentially distributed, (*b*) all drivers have equal and constant critical gap and move-up times, and (*c*) there is a fixed relationship between critical gap and move-up time, namely $t_f = 0.6t_g$.

Figure 1 shows a plot of the minor street capacity as estimated by the HCM as a function of the major street flow for several values of the critical gap. Field measurements taken for this study are also shown. The plot shows that the HCM method tends to overestimate minor street capacity for low conflicting flows (below 600 vph) and underestimates minor street capacity for high conflicting flows (above 600 vph).

UK Empirical Method

The United Kingdom is the only country today that does not use the gap acceptance method for TWSC intersections. Kim-

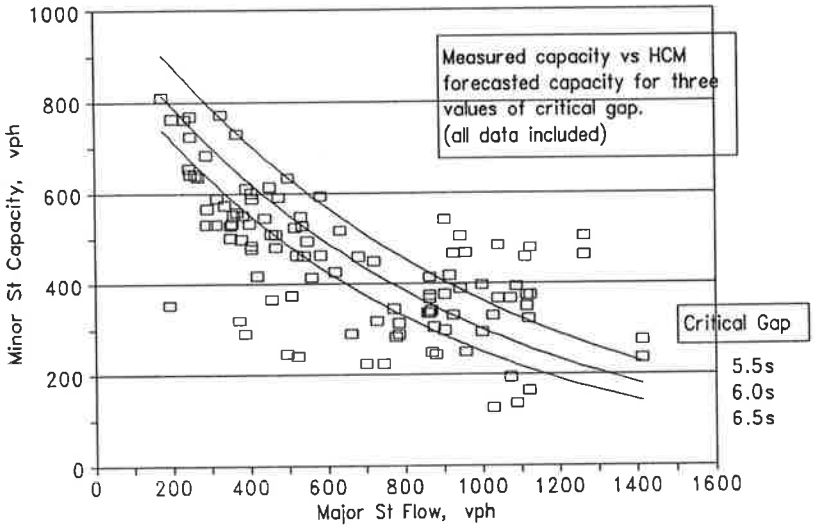


FIGURE 1 Measured capacity versus HCM forecast capacity.

ber and Coombe (*10*), Kimber (*11*), and Semmens (*12*), from the U.K.'s Transport Road Research Laboratory (TRRL), have developed empirically based capacity models based solely on traffic flow rates and site geometry. Kimber identified two potential problems with the gap-acceptance method (*11*, p. 101) that justify the empirical approach.

> Is simple gap acceptance a sufficient description of the vehicle-vehicle interaction process in all common circumstances, and are detailed assumptions of the theory adequate—for example, are the model parameters independent of the magnitude of the priority stream flow.

Kimber also notes that in observation of traffic flow at capacity conditions,

> . . . there were significant periods of priority reversal, during which non-priority vehicles edged into the priority streams, forcing their own gaps. . . . We therefore chose to develop empirical capacity models specified directly in terms of the traffic flows themselves, rather than to assume a priori the completeness of the gap acceptance description. This represents a different level of approach, rather like a thermodynamic description of the properties of a gas as contrasted to a kinetic theory description. (*11*, p. 102)

Kimber and Coombe describe the general equation for the capacity Q of a nonpriority movement:

$$Q = X\left(q_o - \sum_i Y\alpha_i q_i + Z\right) \tag{6}$$

where X, Y, and Z represent functions of the geometric parameters of the intersection.

Effect of Conflicting Flows

The gap acceptance method and the U.K. empirical method agree that the most important factor affecting minor street capacity is the flow rate on the conflicting approaches. This fact presents strong evidence, then, that any model developed here should relate the minor street capacity to the major street or conflicting flow rates. Furthermore, the assertions of Kimber and others from TRRL represent strong motivation to test the empirical method using the data base developed here.

A variety of linear functional forms were investigated using the basic format presented in Equation 7.

$$Q_s = \alpha_1 - \sum_i \alpha_i q_i \tag{7}$$

where q_i is the flow for the ith conflicting movement. Two of the models developed are given in Equations 8 and 9:

$$Q_s = 657.34 - 0.32q_{c,L} - 0.31q_{c,R} \qquad R^2 = 0.44 \tag{8}$$

$$Q_s = 657.44 - 0.31q_c \qquad R^2 = 0.44 \tag{9}$$

Figure 2 shows a plot of both Equation 9 and the actual data for capacity versus the conflicting flow rate.

Examination of Figure 2 shows that the model correctly represents the basic feature of the relationship: minor street capacity decreases as the major street flow increases. The wide dispersion of the data about the linear regression line and the magnitude of the R^2 parameter indicate that the relationship may be nonlinear, different functional forms may be evident for different ranges of conflicting flow rate, and additional variables are required to explain more of the variance.

Effect of Speed

The HCM provides critical gap estimates for two different speed ranges, 30 mph and 55 mph. The differences in the capacity curves for these values indicate at least a theoretical importance of the speed of traffic on the conflicting approaches for the minor street capacity. According to this formulation, the higher the speed, the lower will be the minor

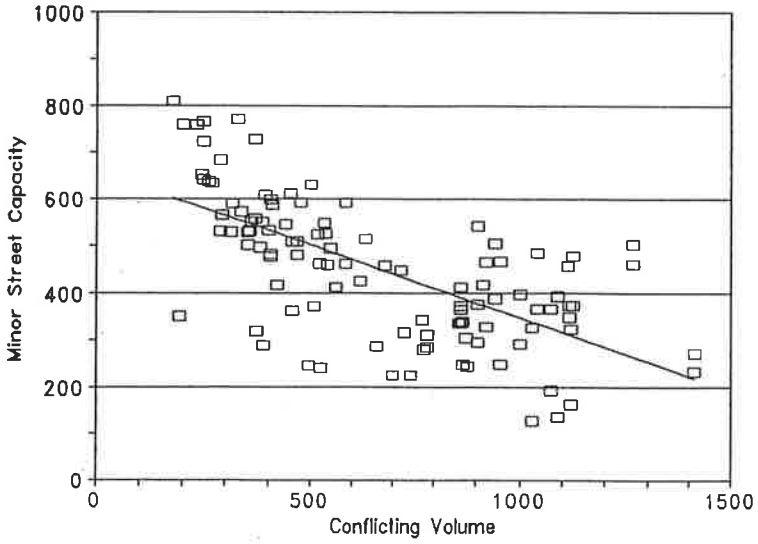


FIGURE 2 Linear capacity model versus measured data.

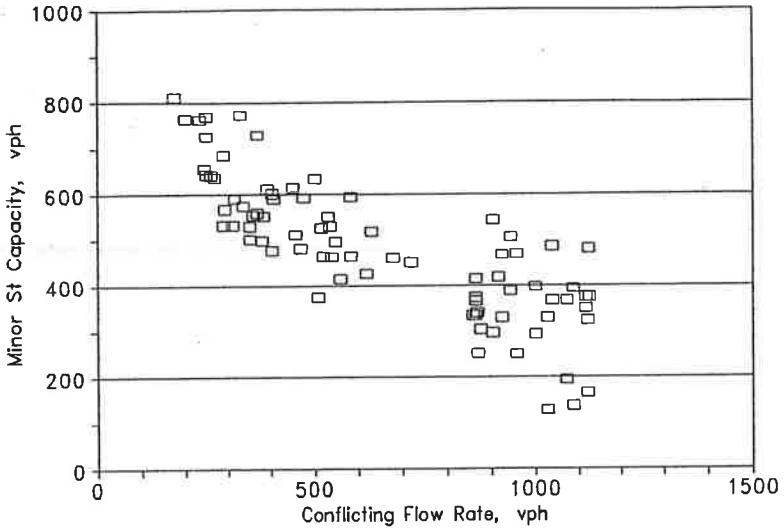
street capacity. This may seem intuitive, since drivers simply need more time to complete their maneuver if they have to make a judgment in higher-speed traffic than in lower-speed traffic.

Figures 3 and 4 show plots of the measured capacity data versus conflicting flow segregated by two speed limit ranges on the major street. Figure 3 includes data in the 25- to 35-mph range, whereas Figure 4 shows data in the 55-mph range.

Multiple regression models were developed to quantitatively determine the effect of major street speed on minor street capacity. The results of this analysis are given in Equations 10 and 11. For the lower speed ranges, typically found on urban arterials (25 to 35 mph), the capacity at low flow rates is about 200 to 250 vph higher than for major streets with higher speeds (55 mph). This difference narrows considerably as the major street flow rates increase. If linear best fit regression lines are drawn through each of these data sets, clear differences in both slope and intercept of the capacity-flow rate relationships appear in these figures. Figure 5 shows a plot of the capacity equation for each speed group.

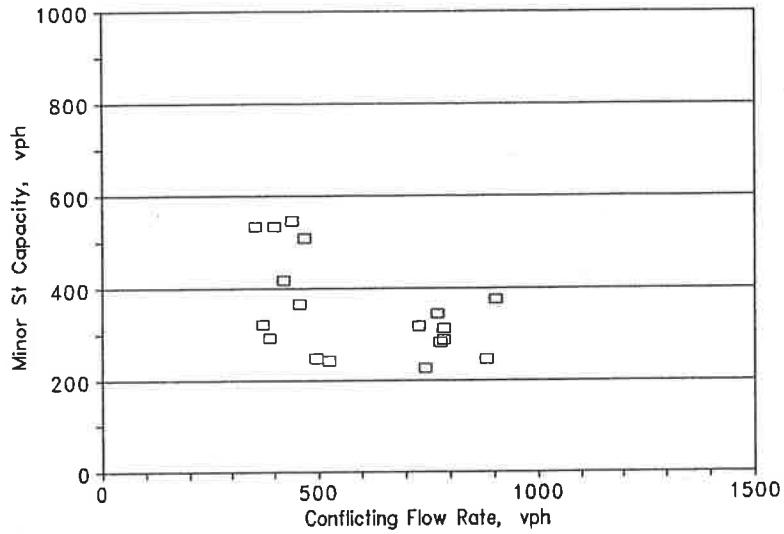
$$Q_s = 740.84 - 0.40q_c \qquad R^2 = 0.68 \tag{10}$$

$$Q_s = 523.99 - 0.29q_c \qquad R^2 = 0.26 \tag{11}$$



Speed Range: 25–35 mph

FIGURE 3 Minor street capacity versus conflicting flow, lower-speed range.



Speed Range: 55 mph

FIGURE 4 Minor street capacity versus conflicting flow, higher-speed range.

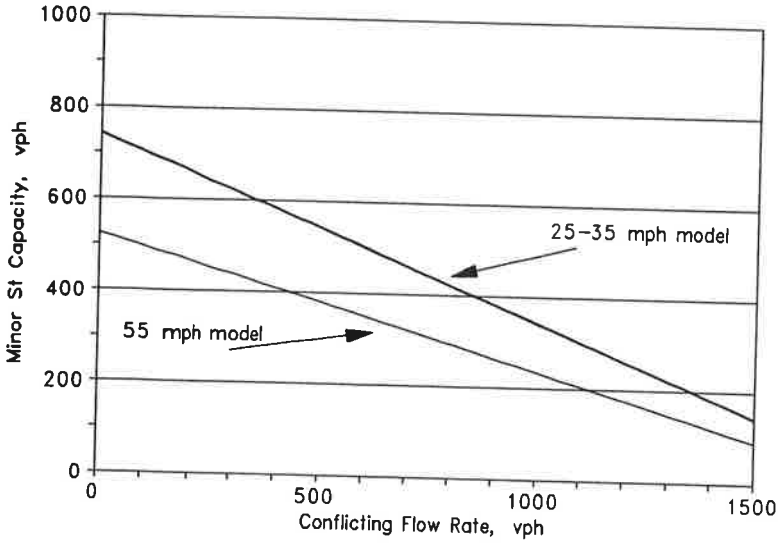


FIGURE 5 Minor street capacity versus conflicting flow, two speed ranges.

Effect of Other Flows

The effects of other flows (i.e., opposing flow and disaggregated conflicting flows) on the minor street capacity were also tested. Three of these models are given in Equations 12, 13, and 14.

$$Q_s = 684.40 - 0.38q_{c,L} - 0.22q_{c,R} - 0.35\,q_o \qquad R^2 = 0.47 \tag{12}$$

$$Q_s = 656.99 - 0.45q_{c,L,LT} - 0.30q_{c,L,TH} - 0.59q_{c,L,RT} - 0.36q_{c,R,LT} - 0.30q_{c,R,RT} \qquad R^2 = 0.42 \tag{13}$$

$$Q_s = 673.64 - 0.50q_{c,L,LT} - 0.27q_{c,L,TH} - 0.45q_{c,L,RT} - 0.31q_{c,R,TH} - 0.41q_o \qquad R^2 = 0.45 \tag{14}$$

Two conclusions can be drawn on the basis of a review of these equations.

1. The models presented in Equations 12, 13, and 14 have R^2 values between 0.4 and 0.5, indicating that improvements in model fit over the models given in Equations 8 through 11 are not gained by disaggregating conflicting flow rates or adding opposing flow rates. This may mean that other factors, such as major street speed or intersection geometry, have a more important effect on capacity than the disaggregated flow variables.
2. Conflicting flows from the left ($q_{c,L}$) have a more significant effect on capacity than conflicting flows from the right ($q_{c,R}$). This is expected, since all minor street movements (left, through, and right) are affected by the conflicting flow from the left, whereas only the left and through minor street movements are affected by the conflicting flow from the right.

Integration of Effects

From the analysis presented, it can be suggested that the capacity of the minor street approach is a function primarily of the conflicting flow rate and the major street speed. Opposing flow rates and disaggregated conflicting flow rates were not shown to significantly improve the capacity model. These factors are now integrated into a recommended capacity model.

The development of the capacity estimation procedure was accomplished incrementally. The steps are summarized below.

Step 1. The data were first segregated according to the speed group of the major street. Two groupings were considered: sites with speeds between 25 and 35 mph and those with speeds between 40 and 55 mph.

Step 2. Figure 6 shows a plot of minor street capacity versus conflicting flow rate for the 25- to 35-mph speed group and for conflicting flows of less than 600 vph. A linear model is fitted through these data, with one additional constraint: the saturation flow rate (when the conflicting and opposing flows are zero), which is just the y-intercept on the curve, is equal to approximately 900 vph. The selection of this constraint can be justified as follows. When the conflicting flow rate is zero, the capacity of the minor street approach is equal to the saturation flow rate of the minor street approach. That is, when there are no vehicles present on the other approaches, vehicles depart from the stop line as rapidly as safety and vehicle performance allow. Whereas this saturation rate has not been measured for TWSC intersections, it has been measured by Kyte (*13*) for all-way stop-controlled intersections. This saturation flow rate is approximately 900 vph.

Step 3. Figure 7 shows a plot of the 25- to 35-mph speed group data for the range of conflicting flows greater than 600 vph. A least squares regression line was fitted through these data.

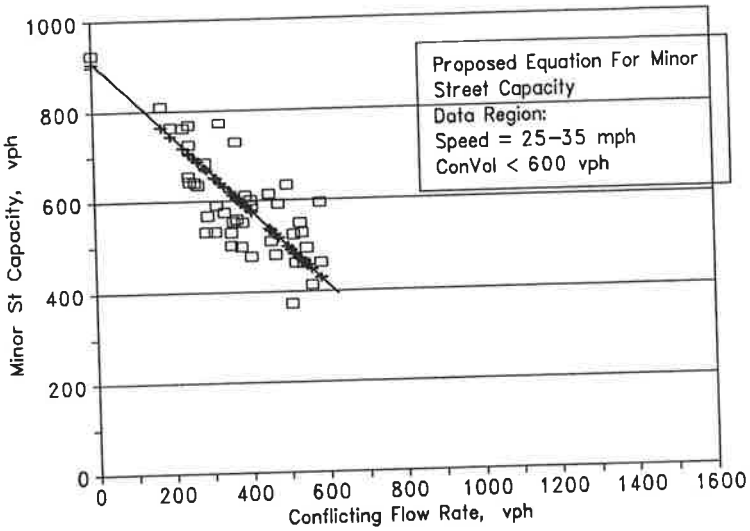


FIGURE 6 Proposed capacity model, lower conflicting flow ranges, lower-speed range.

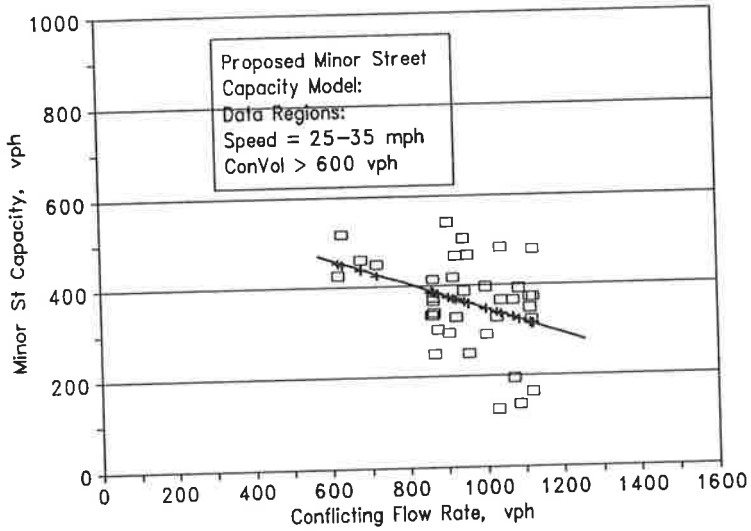


FIGURE 7 Proposed capacity model, higher conflicting flow range, lower-speed range.

Step 4. Figure 8 shows a plot of the 40- to 55-mph data. One equation was fitted through these data and is shown in the figure.

Step 5. Figure 9 shows the three equations together, two for the lower-speed data and one for the higher-speed data. Note, however, that there is an overlap at the higher conflicting flow ranges. To eliminate this overlap, the slopes of the equations were modified slightly; see Figure 10.

Step 6. The final models are shown as Equations 15, 16, and 17.

$$Q_s = 906 - 0.82q_c \quad (15)$$

$$Q_s = 623 - 0.28q_c \quad (16)$$

$$Q_s = 390 - 0.11q_c \quad (17)$$

Comparison of Proposed Method with HCM Procedure

The capacity estimation models proposed here for the two speed groups are compared with the HCM models for similar speed ranges for critical gaps of 6.0 and 7.5 sec, respectively (see Figures 11 and 12).

Figure 11 shows that there is reasonable agreement between the two models for the lower-speed range, with differences of no more than 100 vph, or less than 10 percent. However, the differences are considerable for the higher-speed range, as shown in Figure 12. At lower conflicting flow values, the HCM procedure forecasts capacity values up to 100 percent higher than the new procedures proposed here. The curves intersect at a conflicting flow value of 600 vph; above this value the HCM model forecasts values up to 100 to 150 vph lower than the new procedure.

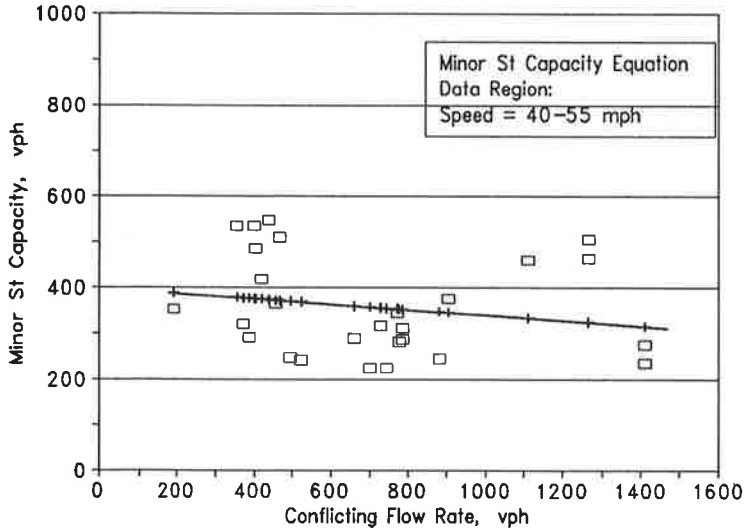


FIGURE 8 Proposed capacity model, higher-speed range.

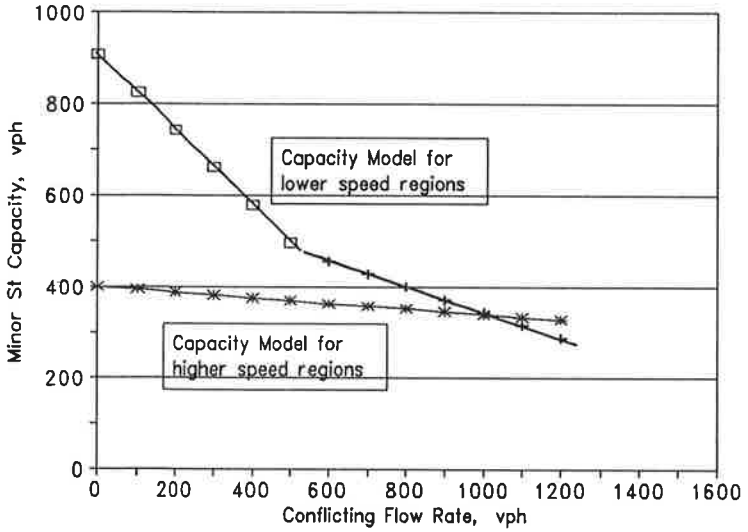


FIGURE 9 Proposed capacity models, preliminary.

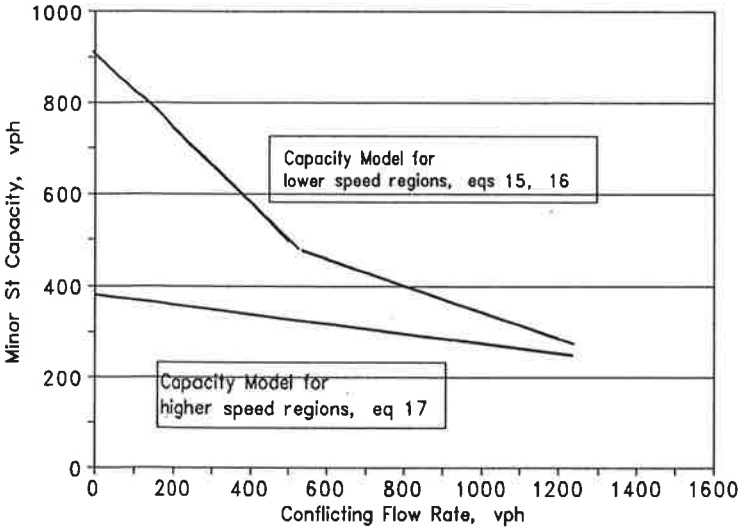


FIGURE 10 Proposed capacity models, final.

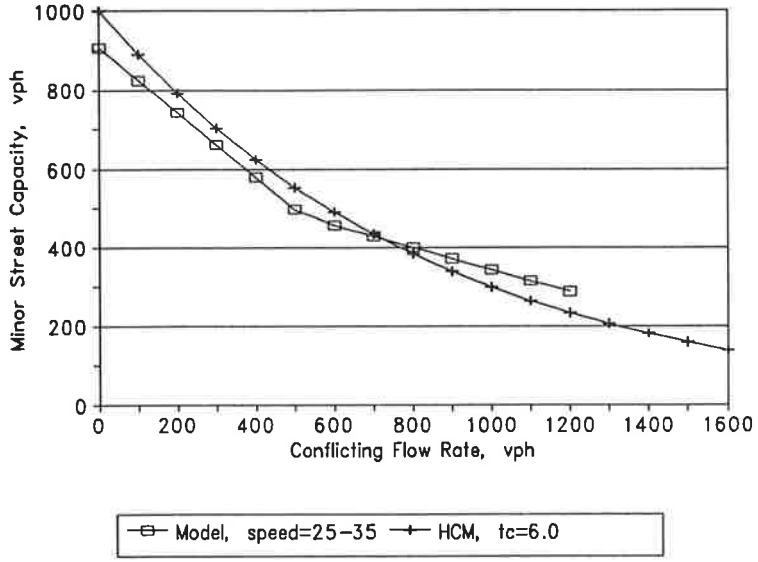


FIGURE 11 Proposed capacity model (lower-speed range) versus HCM model.

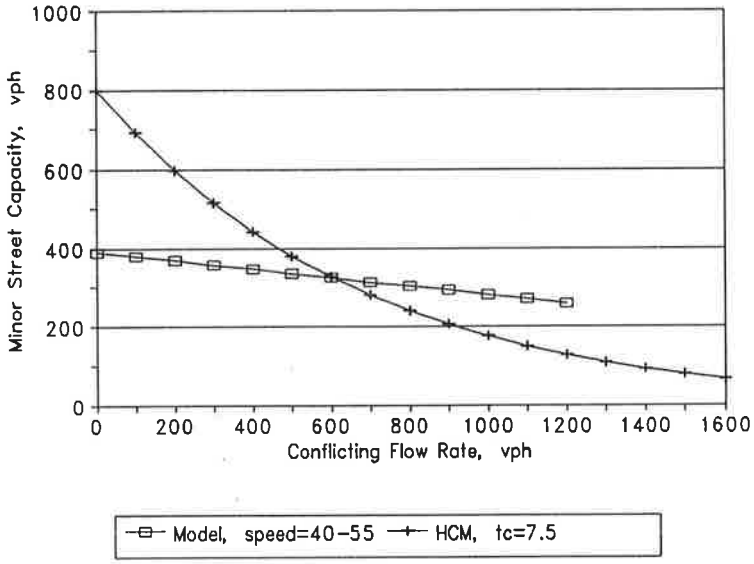


FIGURE 12 Proposed capacity model (higher-speed range) versus HCM model.

DELAY

HCM Method

Reserve capacity is the measure of effectiveness used in the HCM to determine the level of service for a TWSC intersection. Reserve capacity is defined as the unused capacity of a movement, or the difference between the actual capacity for a movement and the flow rate for the movement. The HCM establishes a level of service for each range of reserve capacity and a qualitative description of the delay likely to be experienced (see Table 2).

Reserve capacity, however, has not been a popular parameter with U.S. traffic engineers. It cannot be measured directly in the field, and it is not directly linked with quantifiable delay ranges. Brilon (*2*), however, has shown that reserve capacity is a useful measure and correlates well with the expected delay for a minor street. In fact, reserve capacity is somewhat analogous to the degree of saturation or volume/capacity ratio in that both parameters describe the amount of capacity remaining or available. An example of the relationship of minor street delay to reserve capacity and major street flow is given in Figure 13. The figure is from Brilon (*2*).

Effect of Reserve Capacity

Total delay is the sum of service delay and queue delay. An exponential model relating total delay to reserve capacity was

TABLE 2 Level of Service and Reserve Capacity

Reserve Capacity	Level of Service	Expected Delay to Minor Street Traffic
± 400 vph	A	Little or no delay
300-399 vph	B	Short traffic delays
200-299 vph	C	Average traffic delays
100-199 vph	D	Long traffic delays
0-99 vph	E	Very long traffic delays
< 0	F	-

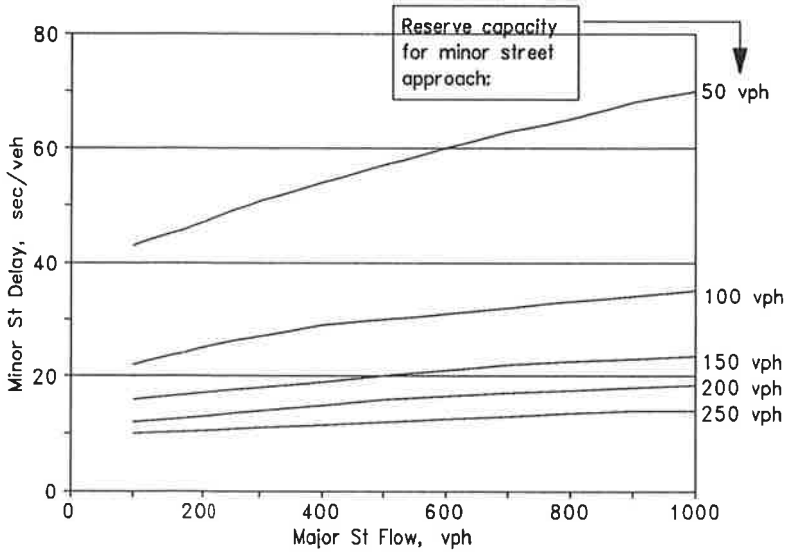


FIGURE 13 Delay as a function of reserve capacity and major street flow, Brilon model.

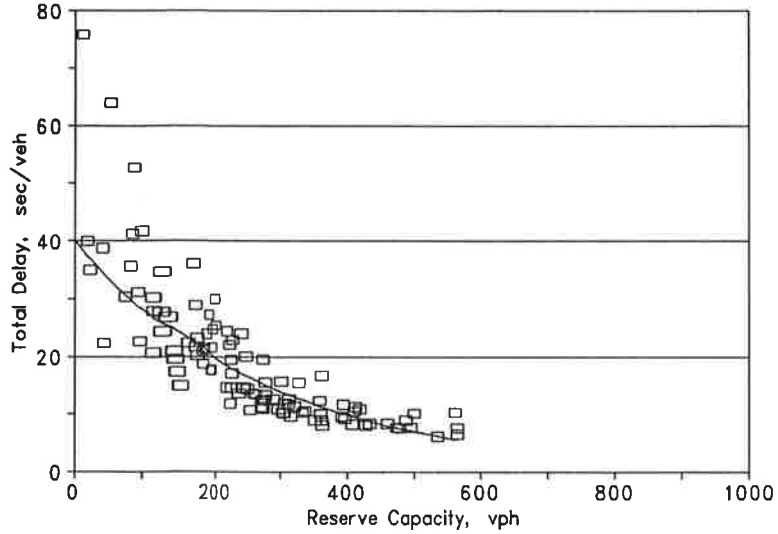


FIGURE 14 Total delay versus reserve capacity.

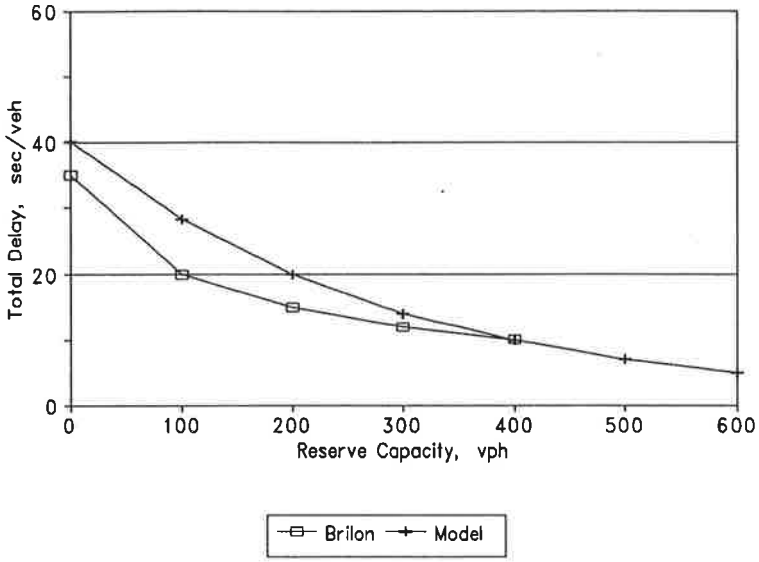


FIGURE 15 Proposed delay model versus Brilon model.

developed and is shown in Equation 18. The plot of Equation 18 is given in Figure 14.

$$d_t = 40.079e^{-0.0035q_{s,rcs}} \qquad R^2 = 0.78 \tag{18}$$

Comparison of Proposed Method with Brilon Model

The proposed delay estimation model is plotted against Brilon's delay model in Figure 15. The Brilon model forecasts somewhat lower delays (in the range of 5 to 10 sec lower) when the reserve capacity is less than 300 vph. The two models nearly coincide for higher values of reserve capacity.

FINDINGS AND CONCLUSIONS

The objective of the research described in this paper is to propose and test empirically based methods to forecast capacity and delay for the minor street approach of a TWSC intersection. This objective has been accomplished. The major findings of this research are summarized as follows.

Data Base

A data base has been assembled from 12 TWSC intersection sites in the Pacific Northwest region of the United States. The data base includes geometric and traffic characteristics from 26.75 hr of intersection operations. The data are summarized over 15-min periods; thus, 107 data points are included in the data base. A video camera was used to film each intersection, and computer software developed for this study was used to enter and reduce the data. Seven traffic variables were produced for each 15-min period of intersection operation: the capacity of the subject approach; flow rates on the subject approach, the opposing approach, and the conflicting approaches; and total delay, service delay, and queue delay on the subject approach.

Proposed Capacity and Delay Models

Proposed methods for estimating capacity and delay are presented. The flow rate on the conflicting approaches is the most important variable affecting capacity on the minor street. The functional form relating minor street capacity to conflicting flow is nonlinear and may depend on the range of conflicting flow rate. The speed on the major street affects the level and slope of the capacity relationship. Delay is affected primarily by reserve capacity. The delay model, an exponential form, provided an excellent fit to the data. The quality of the delay-reserve capacity model provides support to the earlier work of Brilon and others that suggests the importance of the reserve capacity parameter in determining intersection level of service.

Clearly the methods proposed here are only preliminary; both need to be validated with a larger data base. But the results show that the methods may represent a feasible alternative to the gap acceptance approach currently used in the HCM. In short, the empirical approach of directly relating capacity and delay to measured flow rates appears to be feasible and warrants further study.

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge the financial support for this study provided by TransNow, the Idaho Transportation Department, the University of Idaho, Portland State University, and Washington State University.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. W. Brilon. Recent Developments in Calculation Methods for Unsignalized Intersections in West Germany. In *Intersections Without Traffic Signals, Proceedings of an International Workshop*, Springer-Verlag, Berlin, 1988, pp. 111–153.

3. J. Harders. Die Leistungsfähigkeit nicht signalgeregelter städischer Verkerhsknoten (The capacity of unsignalized urban intersections). Schriftenreihe Straßenbau und Straßenverkehrstechnik, Heft 76, 1968.
4. J. Zegeer. Status of Unsignalized Intersection Capacity Research in the United States. In *Intersections Without Traffic Signals, Proceedings of an International Workshop*. Springer-Verlag, Berlin, 1988, pp. 35–47.
5. E. Ruehr. Comparison of U.S. and German Procedures for Unsignalized Intersection Analysis. Compendium of Papers at the Conference on Capacity Analysis Techniques for Two-Way Stop-Controlled Intersections, TRB Annual Meeting, 1991.
6. W. K. Kittelson and M. A. Vandehey. Delay Effects on Driver Gap Acceptance Characteristics at Two-Way Stop-Controlled Intersections. Compendium of Papers at the Conference on Capacity Analysis Techniques for Two-Way Stop-Controlled Intersections, TRB Annual Meeting, 1991.
7. M. Heffron and G. Bezkorovainy. A Methodology for Using Delay Study Data to Estimate Existing and Future Level of Service at Unsignalized Intersections. Compendium of Papers at the Conference on Capacity Analysis Techniques for Two-Way Stop-Controlled Intersections, TRB Annual Meeting, 1991.
8. M. Kyte, K. Lall, N. Mahfood, G. Panchavati, and C. Clemow. Development of Empirical Models To Forecast Delay and Capacity at Two-Way Stop-Controlled Intersections. Compendium of Papers at the Conference on Capacity Analysis Techniques for Two-Way Stop-Controlled Intersections, TRB Annual Meeting, 1991.
9. M. Kyte, A. Boesen, and B. Rindlisbacher. *TDIP—Traffic Data Input Program, Program Documentation and User's Manual*. Version 3.0. Department of Civil Engineering, University of Idaho, Moscow, Idaho, March 1991.
10. R. M. Kimber and R. D. Coombe. *The Traffic Capacity of Major/Minor Priority Junctions*. TRRL Supplementary Report 582. Transportation Road Research Laboratory, Crowthorne, Berkshire, U.K., 1980.
11. R. M. Kimber. Gap-Acceptance and Empiricism in Capacity Prediction. *Transportation Science*, Vol. 23, No. 2, May 1989, pp. 100–111.
12. M. C. Semmens. *The Capacity of Major/Minor Priority Junctions on High Speed Roads*. Working Paper TMN 157. Transportation Road Research Laboratory. Crowthorne, Berkshire, U.K., Oct. 1987.
13. M. Kyte. Estimating Capacity of an All-Way Stop-Controlled Intersection. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 70–81.

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

Synthesis of Recent Work on the Nature of Speed-Flow and Flow-Occupancy (or Density) Relationships on Freeways

FRED L. HALL, V. F. HURDLE, AND JAMES H. BANKS

Research published during the past 5 years has provided a revised picture of the relationships among the key traffic variables of speed, flow, and concentration. A review of the data from earlier studies with this revised picture in mind shows that many of these data are also compatible with the new picture.

The purpose of this paper is to pull together some of the ideas about speed-flow-concentration relationships that have appeared in the past decade. Because the results are not consistent with the commonly accepted depiction of these relationships, it is useful to trace the relationships back to their original development, more than 50 years ago, and to look at some of the old data from a new perspective based on the more recent data. Our conclusion is that little has changed except for the interpretation: freeway traffic may move a little faster and at somewhat smaller headways than in the past, but the old data are remarkably consistent with both the new data and the new interpretations.

The consequence is that a good case can be made for how we ought now to be interpreting the nature of these fundamental relationships, though some qualitative issues remain unresolved. It should also be made clear at the outset that our paper is empirical rather than mathematical. We believe that it is important to have a correct picture of the relationships before attempting to construct detailed mathematical models and that the classical picture found in most textbooks and in the *Highway Capacity Manual* (HCM) (*1*,*2*) is seriously defective. In this paper, we attempt to draw and defend a new picture that is in better agreement with recent empirical research.

The starting point for this discussion is the 1985 HCM (*1*). That publication raised several unanswered questions about the speed-flow relationship, in particular "the difficulty in firmly fixing the shape and location of the curve" beyond about 1,500 vehicles per hour (*1*, pp. 2–24). However, in Chapter 3, Basic Freeway Segments, definitive statements had to be and were made about the shape of these curves. In the 1965 HCM (*2*), the speed-flow curves were parabolic, with speeds decreasing with each increase in flow, even for very low flows; the 1985 HCM reduced the rate of the speed drop at low flows, but the basic shape remained the same: the slope of the curve is always negative and the right end of the curve is vertical.

There have been a number of papers in the past few years dealing with speed-flow-occupancy relationships on freeways. Our synthesis of that work is given added timeliness by the recently approved revised version of Chapter 7 of the HCM (*3*), which contains speed-flow relationships for multilane rural highways radically different from those that appear in Chapter 3. Not only do these new curves keep speeds constant until 75 percent of capacity, but their slope is quite modest even as the flow approaches capacity, with a total speed decrease of only 5 mph over the entire range of flows.

The first section presents our conclusions about the shapes of the speed-flow and flow-occupancy graphs. These conclusions are supported by citations from the literature of the past half dozen years. The second section is a review of older material, which on the face of it might appear to be in conflict with the recent results. One of the primary purposes of this section is to see whether some of the older data might also be consistent with our interpretation of the newer data, or whether freeway traffic behavior has clearly changed over the intervening 30 years.

DEPICTION OF THE FUNDAMENTAL CURVES

This section presents graphically our current understanding of the key bivariate relationships: speed-flow and flow-occupancy. The speed-occupancy relationship follows as a consequence of the other two. We do not discuss it here because of a lack of recent data on it and a lack of space. The two figures are drawn in general terms, without numerical values on the axes, but were constructed on the basis of published studies.

Each figure consists of three segments, represented by series of asterisks, squares, and triangles. The intention is that each of these points be taken as representing the average of some large number of observations. Thus stochastic variation has been suppressed. Our discussion of these figures is based on a hypothetical freeway like the one shown in Figure 1. The entire section has the same capacity, but two entrance ramps are followed by an exit. Clearly the flows will be greatest in Section CD, between the downstream entrance ramp and the exit. Section CD, therefore, is the bottleneck, and if a queue forms anywhere, it will be in Section BC, then spread into AB. It seems obvious that the speed within the queue will be slow and the density or occupancy high. Within the bottleneck

F. L. Hall, Department of Civil Engineering and Geography, McMaster University, Hamilton, Ontario, Canada L8S 4L7. V. F. Hurdle, Department of Civil Engineering, University of Toronto, Toronto, Ontario, Canada M5S 1A4. J. H. Banks, Department of Civil Engineering, San Diego State University, San Diego, Calif. 92128.

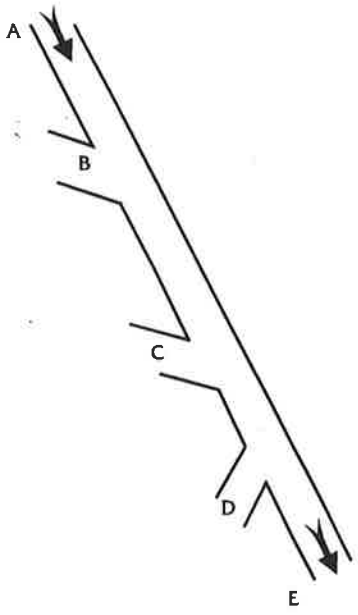


FIGURE 1 Representative freeway segment.

Section CD, however, one would suppose that the traffic might accelerate, such that there will be higher speeds at D than at C, and measurements confirm this commonsense conclusion (*4*).

Speed-Flow Relationships

Figure 2 shows the most likely interpretation we have found of the speed and flow data acquired from numerous freeway systems over the past 30 years, as applied to the hypothetical freeway in Figure 1. The nearly level upper line (indicated by * in the figure) represents traffic behavior in the absence of any queue. On the freeway in Figure 1, we would expect the conditions represented by this curve to occur everywhere until a queue formed in Sections BC and AB, but after that only in the part of AB upstream from the end of the queue and in DE, downstream from the bottleneck. The 1985 HCM (*1*) reduced the rate of the speed drop from that shown in the

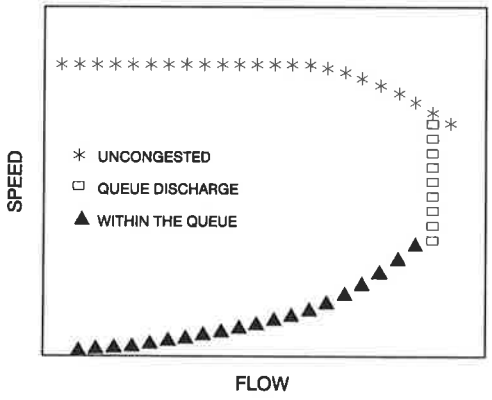


FIGURE 2 Generalized speed-flow relationships.

1965 HCM, and the 1990 revisions to Chapter 7 (*3*) enhance that tendency, keeping speeds constant until 75 percent of capacity is reached. Those results for multilane rural highways are consistent with comparable observations available for freeways. The only disagreement between the recent freeway studies and the Chapter 7 curve might be in the magnitude and shape of the speed decrease for higher flow rates. The revised Chapter 7 shows drops of only 5 mph, or roughly 10 percent of the free-flow speed. Hurdle and Datta (*5*), Persaud and Hurdle (*4*), and Hall and Hall (*6*), as reinterpreted by Hall and Agyemang-Duah (*7*), suggest about a 20 to 25 percent drop. Wemple et al. (*8*) found about a 10 to 15 percent drop. Banks (*9*) found a drop of less than 10 percent.

An alternative way of looking at these data is not to focus on the speed drop but to look at the speed at the endpoint of the curve. Several of the freeway studies cited find that point to be about 80 km/hr, or 50 mph. If speed at the right end of the curve is approximately the same for some large class of freeways, the magnitude of the speed drop is simply a function of the free-flow speed. Such a conclusion would go a long way toward explaining the differences between data from North American urban freeways and the high-speed German Autobahnen. Unfortunately, compatibility with Japanese data seems less likely.

Several studies, however, report different results. Persaud and Hurdle (*4*) cite a lower value, but the relevant figures in their paper suggest considerable scatter, with a range that includes 80 km/hr. Wemple et al. (*8*), on the other hand, report speeds of more than 60 mph at their I-680 site when the flow is more than 2,000 vehicles per hour per lane and show data for a site on I-880 that could be interpreted as indicating no decrease, or even an increase, in speed at very high flows. Hurdle and Datta (*5*), Persaud (*10*), and Hall and Hall (*6*) include similar evidence of circumstances in which speed seems to be completely independent of flow, but it will be presumed in this paper that this is not ordinary behavior.

For constructing Figure 2, a speed drop of 20 percent of free-flow speeds has been used, which would be consistent with a 100-km/hr or 60-mph free-flow speed and an endpoint speed of 80 km/hr or 50 mph. Despite the possible discrepancy between this value and the Chapter 7 curve, it is important to note that flows near capacity, when they occur before queue formation, occur at much higher speeds than are currently shown in the HCM and that none of the recent data indicate that this portion of the curve becomes vertical. The shape shown for the right end of the curve is arbitrary. The new Chapter 7 curves change slope in a more or less quadratic manner, but several data sources seem to imply a straight line.

The second segment to consider is the vertical band represented in Figure 2 by the squares. This is the behavior to be expected in bottleneck Section CD when there is a queue upstream. Everywhere within the bottleneck section, the mean flow rate will be the same, but the mean speed will be a function of the location of the observation point since drivers are accelerating from the slow speeds within the queue to their desired speeds for that section of roadway (*4*). Thus, the farther downstream from the front of the queue one measures, the higher the average speed will be. The vertical segment of the curve shown in Figure 2 is not really a speed-flow relationship at all, but a speed-location relationship plotted on a graph that has no location axis.

Although any one location is depicted in Figure 2 as a single point, both flow and speed are random variables, so actual data will contain a good deal of scatter in both directions. For example, Hurdle and Datta (*5*) reported standard deviations of 205 passenger car units per hour (for flows based on 2-min counts) and 11.0 km/hr for average queue discharge flow and speed of 1,984 passenger car units per hour per lane and 79.5 km/hr, at a point 2 km downstream from the head of a queue. Furthermore, the flow and speed random variables are not independent, so the queue discharge data at one location may exhibit a distinct slope. Data from several locations suggest that at a given location lower queue discharge flow will be accompanied by lower speeds. Hence, data from this part of the curve may look as if they are on the third part (the triangles). Figure 3 provides an example of both the scatter and the apparent slope. The data are from an Ontario freeway from two locations downstream of an entrance ramp, one 0.2 km (the filled squares) the other 2.5 km (the open squares) downstream. The data from 0.2 km downstream are clearly centered at an average speed well below that of the data from 2.5 km downstream, which are themselves still below the uncongested data. There is a trend in the data from 0.2 km downstream that might be thought to look like part of the lower branch of a two-regime speed-flow curve despite the fact that these data come from queue discharge flow. Koshi (*11*) has reported finding a decrease in queue discharge flows on Japanese freeways as queuing delays increased from 0 to about 10 min, but we have seen no evidence of such a trend in North American freeway data.

The final segment of the curve, the triangles, represents behavior within the queue. This segment of Figure 2 has been drawn on the basis of logical considerations rather than from data. There is very little modern data for this segment and hardly any below flows of 50 percent of capacity. Some even question the existence of a speed-flow relationship for these conditions. It is reasonable to suppose that there is a relationship of averages taken over periods long enough that the effect of stop-and-go conditions within the queue are smoothed out. The average time a vehicle spends in the queue (and thus its average speed) is determined by the number of vehicles to be "served" by the bottleneck before this vehicle, and by the "service rate" of the bottleneck. Thus, if more vehicles enter downstream of the point observed, both the flow and the average speed will necessarily decrease. In Figure 1, for example, the flow in Section AB will be lower than that in BC by the amount of the entrance ramp traffic at B, and average speeds will therefore be lower. This simple argument is persuasive, but it is not sufficient to define the curve. Furthermore, it is clear from the data that the scatter of data about this curve is far greater than for the other two portions of the relationship.

The logic of the curve begins with the conventional idea that the left end must be at the origin since zero speed and zero flow obviously occur together. The right end has been drawn more or less joining the bottom of the portion representing queue discharge flow because data we have seen tend to look like this, because flows in the queue cannot exceed those in the bottleneck, and because it seems illogical to suppose that vehicles being served moved slower than those waiting in line behind them. The last statement, however, is an oversimplification. Flows within Section BC of the Figure 1 freeway can approach those within the bottleneck only if the flow on the entrance ramp at C approaches zero, so the speed just upstream from the bottleneck will usually be less than that indicated by the right end of the triangles curve, which is an indication of what would happen if the ramp flow did approach zero. It follows, since the speed at the upstream end of Section CD cannot be much different than at the downstream end of BC, that the line of squares really should extend below the triangles curve. Like the right end of the plus-sign curve, the bottom end of the line of squares is not well defined. The lowest average speed one will observe in a bottleneck like CD is a function of the entrance ramp flow and the roadway geometry; the figure is only an indication of the possibilities. Furthermore, the right end of the triangles portion is unlikely to be observed in practice, so real data are likely to exhibit a gap, rather than a crossing of curves.

Flow-Occupancy Relationships

Figure 4 shows the flow-occupancy relationship. Although density was almost always the measure of traffic concentration used in the early days, and is still used extensively in theo-

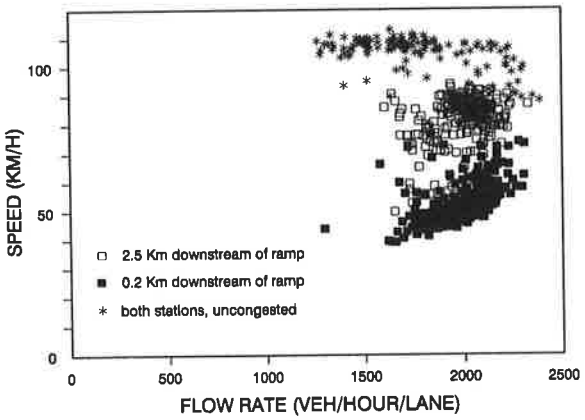


FIGURE 3 Ontario speed-flow data from within the bottleneck, 30-sec. intervals.

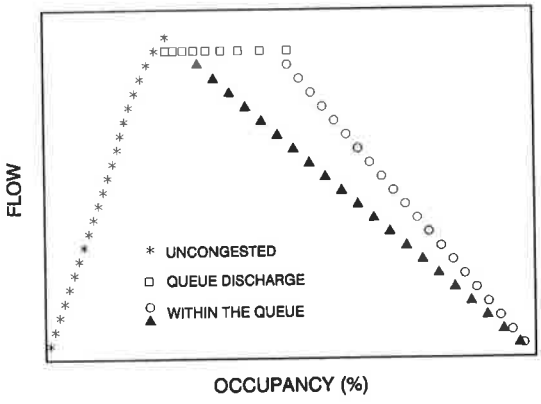


FIGURE 4 Generalized flow-occupancy relationships.

retical work, most present empirical studies have used occupancy because it is commonly measured by freeway management systems. The difference is not important for purposes of this discussion. Athol (*12*) stated that there is a linear relationship between occupancy and density, but more recent analyses [Koshi et al. (*13*) and Hall and Persaud (*14*)] indicate that the linear relationship holds over only a portion of the range of the variables, with the nonlinear aspect of the relationship between occupancy and density depending on the covariance between vehicle lengths and vehicle speeds. Given the small magnitude of the nonlinearity and the level of generality at which we will be discussing the relationships, what we say about occupancy can be applied to density as well.

The flow-occupancy graph in Figure 4 also has three segments. The location of the first segment, the asterisks, is well documented; the logic used to explain the squares in the previous section leaves little doubt about where they must lie in this representation; but we are unsure about the third segment, so have shown two versions. Looking at the three portions one at a time, we see that the constant speed portion of the asterisk curve in Figure 2, out to perhaps 75 percent of capacity, is reflected in a linear flow-occupancy relationship up to the same volume; beyond that point, the relationship curves off slightly to the right, reflecting the increase in occupancy that occurs for a given flow rate as speeds decrease.

The second segment of the diagram (the squares) represents queue discharge flows. As in the speed-flow diagram, the average rate of flow everywhere within the bottleneck must be the same and slightly less than the maximum flows observed during prequeue operations. The mean occupancy, on the other hand, will vary with location over a range that is not entirely well defined. At the left end, it seems obvious that the line of squares should meet the line of asterisks, since queue discharge speeds continue to increase until they regain normal uncongested speeds. On the right end, the obvious limitation is that the occupancy should not exceed that at the head of the queue, but just as with the speed-flow curves, this is not a clearly defined limitation since the occupancy in the queue clearly depends on the ramp flows. Thus, in a generalized diagram like Figure 4, the point at which the line of squares ends on the right is necessarily arbitrary. In a diagram showing the results of a specific experiment, there would, of course, be a rightmost point, but this would not mean that points further to the right could not occur if the traffic pattern changed.

The two possibilities for the segment representing the congested regime within the queue (the triangles and the circles) are based as much on logic, and even conjecture, as on data, though we did look at a good deal of data before drawing them. Just as was the case for the speed-flow relationship, the twin difficulties are that all of the studies we have seen show more scatter in these data than in the uncongested case and that empirical information about very low flows is scarce.

Banks (*15*) has suggested that this portion of the relationship is linear. The very extensive data set presented by Koshi et al. (*13*) shows slight but definite convexity, but that uses density rather than occupancy. The linear possibility being the simpler one, we have drawn straight lines, but without any intention of implying that this is necessarily correct.

The remaining question is where to place the upper end of this line. Three possibilities have been suggested. May and others working at the Chicago Freeway Surveillance and Control Center proposed that the congested segment start at the right-hand end of the queue discharge flow as indicated by the circles [May et al. (*16*), Figure 7, p. 53], and it would appear that Chicago continues to accept that depiction [McDermott (*17*), Figure 4, p. 338]. Koshi et al. (*13*) proposed a reverse lambda shape, which seems to imply that the line joins the plus-sign curve below the queue discharge curve. Both Hall et al. (*18*) and Banks (*15*) have suggested an inverted V, implying that the line should start near the peak of the prequeue flow in somewhat the manner of the line of triangles. The choice among these proposals should be based in part on logic and in part on the available data, but our reaction to the data we have seen is that it leaves as many questions as answers.

If one accepts the triangle line, the queue discharge curve (the line of squares) will stick out from the inverted V like a windblown flag—a situation that seems counterintuitive. However, as discussed in connection with Figure 2, these relationships in Figures 2 and 4 represent a wide variety of situations, not all of which are likely to happen at any one location. The "flag" represented by the squares is a natural consequence for the flow-occupancy graph of the vertical line that is the queue discharge operation, as shown in the speed-flow curve of Figure 2. As was discussed for Figure 2, the queue discharge data from any one location will show a range of flows and occupancies. Since these variables are not independent, a pattern will often be observable within them. A number of examples we have looked at tend to show a "drooping flag," which might easily be confused with the congested portion of this curve. Figure 5 shows an example of such a pattern from San Diego data downstream of a bottleneck. The downward trend in the queue discharge data is apparent in these data, but the figure also supports the constant volume segment shown in Figure 4 and the interpretation of it as queue discharge flow.

Summary

As was pointed out as long ago as 1958 (*16,19*), the nature of the data acquired from a freeway depends on where the

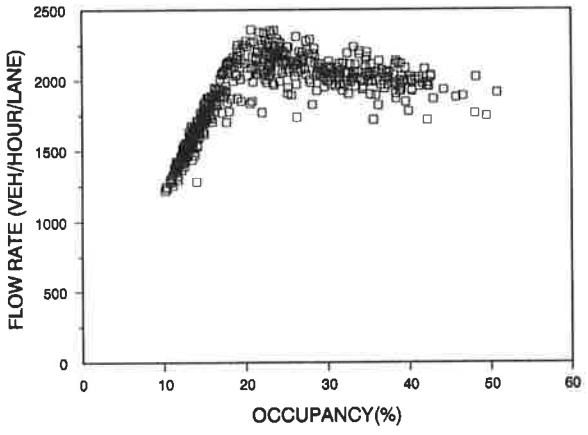


FIGURE 5 San Diego flow-occupancy data from the bottleneck.

measurements are taken with respect to bottlenecks and queues. May (*20*, p. 288) provides a helpful discussion and set of diagrams explaining the general nature of this phenomenon. In brief, upstream of a limiting bottleneck (e.g. a lane drop or closure), there will be congested flow, but capacity operations will probably not occur. Within the bottleneck section, capacity flow will occur, but there will be no congested operations. If capacity increases downstream of the bottleneck section (for example, the dropped lane has been restored or reopened), at such a location there will be neither congested operation nor capacity operation. One important consequence is that it is impossible to obtain data covering the full range of possible operations at any one station. Indeed, the data needed to create even the one segment for queue discharge operations must necessarily come from a series of locations within the bottleneck.

The consequences of this dependence on location for the specific case of flow-occupancy data are shown in Figure 6. The freeway segment shown in Figure 1 forms the basis of the diagram, with conditions as they would be observed during a peak traffic period. The triangle at the foreground of the picture represents the flow-occupancy inverted V proposed by Banks (*15*) and Hall et al. (*18*). There is a queue upstream of Ramp C, extending back beyond Ramp B, but not as far as Location A. Thus for some distance downstream of A, traffic is still uncongested. At some intermediate location between A and B, drivers encounter the back end of the queue and experience an abrupt change of conditions, to heavily congested. When they pass Ramp B, conditions remain congested, but average flow rates increase (and occupancies decrease, because speeds also increase). Upon arrival at Ramp C, drivers are able to accelerate through the bottleneck. Hence the early part of the data in Section CD is the "flag" of Figure 4, and the later part is back on the uncongested surface (if Section CD is sufficiently long). Past Ramp D, volume decreases, due to vehicles exiting. In the absence of incidents somewhere along the road, one should not expect to see Section CD behavior anywhere upstream of C or BC behavior downstream of C.

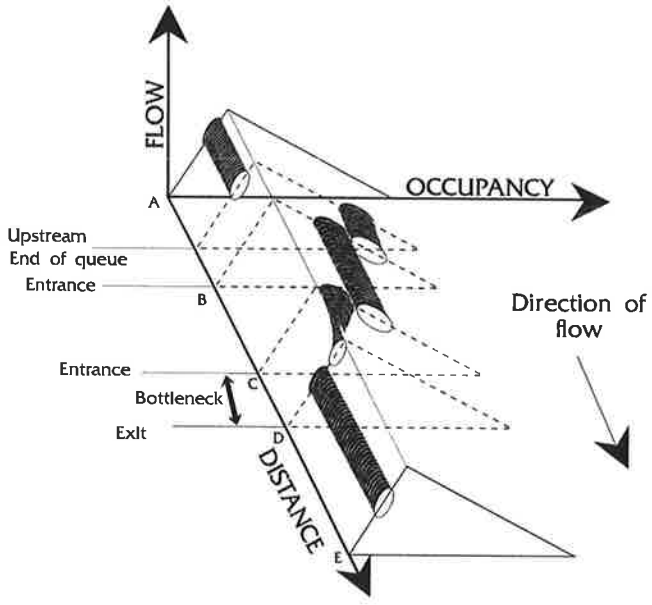


FIGURE 6 Expected approximate location of flow-occupancy data in the vicinity of a bottleneck during congested conditions.

REVIEW OF EARLIER WORK

In reviewing the historical roots of the existing HCM depictions of speed-flow-occupancy relationships, we consider the possibility that much of the earlier data might be consistent with the representations given in the preceding section and that the conventional interpretations arise because of particular historical events. In particular, we suggest that Greenshields's seminal work in 1935 (*21*) has had an unduly dominant influence on all subsequent interpretations of such data.

Consider the details of Greenshields's paper. The data were collected on one lane representing one direction of a two-lane two-way rural road. The seminal graphs of speed versus density and speed versus flow are based on seven data points, six of which are at densities below 60 veh/mi and come from one highway; the seventh is at a density of 150 veh/mi and is taken from a different highway. The straight-line relationship for speed versus density assumes away a lot of missing data between 60 and 150 veh/mi and in turn determines the parabolic shape of the speed-flow curve. Despite the presence of such heroic assumptions, it is possible that these visual representations, with their accompanying elegant mathematics, have had a determining impact on the perception of relationships within freeway data.

Besides the Greenshields model, another possible influence inclining people to U-shaped flow-concentration models may well have been the fact that that was the shape derived by Gazis et al. (*22*) when they originally linked car-following models and these macroscopic flow models. This additional influence would have been reinforced by the simplicity of single-regime models and the flexibility of the nonlinear models as developed by Gazis et al. (*23*).

The generalized flow-occupancy relationship presented in Figure 4 (using the circles for the third segment) is similar to one put forth by May et al. (*16*). It is important to note the differences in interpretation of their flow-occupancy diagram. Their interpretation was that speeds remain constant up to capacity. Ours, on the other hand, has a slight decrease in speeds approaching maximum flow. We agree completely with them that there is a zone of constant volume. However, we have identified this region as queue discharge flow, whereas May et al. say that it "represents impending poor operations" (*16*, p. 52). Another similarity is the use of a linear relationship within congested operations, although they change the slope of that line at a density of 100 veh/mi.

Despite efforts such as that by May et al. to produce a newer picture of these relationships, the curves appearing in the 1965 manual are still clearly based on Greenshields's parabola. Two out of the three empirical studies summarized in Figure 3.37 of the 1965 HCM gave clear evidence of nearly constant speeds out to volumes of 1,000 or 1,200 vph/lane, yet the very next page of the 1965 HCM puts forth the parabolic shape as "typical" for freeways and expressways.

Soon after the appearance of the 1965 HCM, Drake et al. (*24*) published a thorough empirical testing of a number of

mathematical expressions for these three relationships. One of their two subjective choices of best models was a three-regime linear model—and our proposed diagrams also reflect three distinct "regimes" of behavior. May and his students have continued to work on the question of the best type of mathematical model to fit empirical data [Ceder (*25*), Ceder and May (*26*), and Easa and May (*27*)]. The models still do not fit the data well in the vicinity of capacity, and each data set requires different parameters for the mathematical model, the extreme instance of this occurring for 2 days of data from the same location.

Important insights into the task of estimating the relationships among these data were presented by Duncan (*28,29*). In his 1976 paper, he dealt with studies in which speed and flow data were generated and density data calculated from the relationship volume = speed × density. He noted that the consequence of taking a relatively good-fitting speed-density function and transforming it (via the same equation) to a speed-flow function was a function that did not fit well with the original speed-flow data. In his 1979 paper, Duncan used real data, rather than randomly generated data, to enlarge on the difficulties: minor changes in the nature of the speed-density function resulted in major changes in the speed-flow function. The data in his paper would also be consistent with Figure 2, although that is not the nature of the function that Duncan shows.

Overall, then, it can be seen that previous studies do not provide strong support for the parabolic speed-flow curve, which nevertheless continues to influence the shape of the HCM (and other) representations of this relationship [see, for example, May (*20*, p. 288) or McShane and Roess (*30*, p. 286)]. On the other hand, it is possible that many of the data in the earlier studies are consistent with Figures 2 and 4. Indeed, as early as 1961, the Chicago group identified many of the same features of these relationships. Hence, it may be that driver behavior has not changed in any fundamental way during the past 30 years. This, however, is a stronger conclusion than is warranted from the available data. Suffice to say that the earlier studies do not contradict our proposed model, and provide some support for it.

SUMMARY AND CONCLUSIONS

In all likelihood, no one location will be able to provide data for the full range of operations; hence it will be impossible to identify the shape of the speed-flow-occupancy relationships on the basis of data from one location. Consequently, the curves in Figures 2 and 4 are composite curves, drawing together information from numerous locations. It would in fact be a mistake to attempt to construct separate curves for specific locations, since any particular location can be expected to be missing important parts of the overall relationship.

The location of the congested branch of the curve is still a problem. Two possibilities were offered in Figure 4. Perhaps more important, however, the location of the other two segments of the curves seems to be clear. Uncongested operations on freeways are consistent with the speed-flow figure for multilane rural highways approved last year for a revised Chapter 7 of the HCM. The remaining questions are the flow at which speeds begin to decline, the exact shape and magnitude of the decline, and the value of capacity. One feasible interpretation of the data is that the speed at the right-hand end of this segment of the curve is constant, at least for similar types of freeways and similar driver populations. (Japanese data might show results different from these North American studies, for example). Should this interpretation be correct, the speed drop depends on free-flow speeds, which are probably affected by such things as posted speed limits and level of enforcement. The presence of queue discharge flow has been clearly demonstrated, confirming (with a different interpretation) the zone of constant flow identified 30 years ago in the Chicago studies.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. *Special Report 87: Highway Capacity Manual.* HRB, National Research Council, Washington, D.C., 1965.
3. Committee A3A10, Subcommittee on Multilane Highways. Chapter 7: Capacity and Level of Service Procedures for Multilane Rural and Suburban Highways. 1990.
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, pp. 191–198.
5. 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.
6. F. L. Hall and L. M. Hall. Capacity and Speed Flow Analysis of the QEW in Ontario. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 108–118.
7. F. L. Hall and K. Agyemang-Duah. Freeway Capacity Drop and the Definition of Capacity. Presented at the 70th Annual Meeting of the Transportation Research Board, Washington, D.C., 1991.
8. E. A. Wemple, A. M. Morris, and A. D. May. Freeway Capacity and Level of Service Concepts. In *Highway Capacity and Level of Service, Proc. of the International Symposium on Highway Capacity* (U. Brannolte, ed.), Karlsruhe, Germany, 1991, pp. 439–455.
9. J. H. Banks. Flow Processes at a Freeway Bottleneck. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.
10. B. N. Persaud. Study of a Freeway Bottleneck To Explore Some Unresolved Traffic Flow Issues. Ph.D. dissertation. University of Toronto, Toronto, Ontario, Canada, 1986.
11. M. Koshi. Japanese Country Report—Highway Capacity Research Activities in Japan. Presented at the International Symposium on Highway Capacity, Karlsruhe, Germany, 1991.
12. P. Athol. Interdependence of Certain Operational Characteristics Within a Moving Traffic Stream. In *Highway Research Record 72*, HRB, National Research Council, Washington, D.C., 1965, pp. 58–87.
13. M. Koshi, M. Iwasaki, and I. Okhura. Some Findings and an Overview on Vehicular Flow Characteristics. *Proc., 8th International Symposium on Transportation and Traffic Theory*, 1983, pp. 403–426.
14. F. L. Hall and B. N. Persaud. An Evaluation of Speed Estimates Made with Single-Detector Data from Freeway Traffic Management Systems. In *Transportation Research Record 1232*, TRB, National Research Council, Washington, D.C., 1989, pp. 9–16.
15. J. H. Banks. Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 53–60.
16. A. D. May, Jr., P. Athol, W. Parker, and J. B. Rudden. Development and Evaluation of Congress Street Expressway Pilot

Detection System. In *Highway Research Record 21*, HRB, National Research Council, Washington, D.C., 1963, pp. 48–70.
17. J. M. McDermott. Freeway Surveillance and Control in Chicago Area. *Transportation Engineering Journal, ASCE*, Vol. 106, 1980, pp. 333–348.
18. F. L. Hall, B. L. Allen, and M. A. Gunter. Empirical Analysis of Freeway Flow-Density Relationships. *Transportation Research A*, Vol. 20A, 1986, pp. 197–210.
19. L. C. Edie and R. S. Foote. Traffic Flow in Tunnels. *HRB Proc.*, Vol. 37, 1958, pp. 334–344.
20. A. D. May. *Traffic Flow Fundamentals*. Prentice-Hall, 1990.
21. B. D. Greenshields. A Study of Traffic Capacity. *HRB Proc.*, Vol. 14, 1935, pp. 448–477.
22. D. C. Gazis, R. Herman, and R. B. Potts. Car Following Theory of Steady-State Flow. *Operations Research*, Vol. 7, 1959, pp. 499–505.
23. D. C. Gazis, R. Herman, and R. W. Rothery. Non-Linear Follow-the-Leader Models of Traffic Flow. *Operations Research*, Vol. 9, 1961, pp. 209–229.
24. J. S. Drake, J. L. Schofer, and A. D. May. A Statistical Analysis of Speed Density Hypotheses. In *Highway Research Record 154*, HRB, National Research Council, Washington, D.C., 1967, pp. 53–87.
25. A. Ceder. Investigation of Two-Regime Traffic Flow Models at the Micro- and Macroscopic Levels. Ph.D. thesis. University of California at Berkeley, Berkeley, 1975.
26. 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.
27. S. M. Easa and A. D. May. Generalized Procedures for Estimating Single- and Two-Regime Traffic Flow Models. In *Transportation Research Record 772*, TRB, National Research Council, Washington, D.C., 1980, pp. 24–37.
28. N. C. Duncan. A Note on Speed/Flow/Concentration Relations. *Traffic Engineering and Control*, 1976, pp. 34–35.
29. N. C. Duncan. A Further Look at Speed/Flow/Concentrations. *Traffic Engineering and Control*, 1979, pp. 482–483.
30. W. R. McShane and R. P. Roess. *Traffic Engineering*. Prentice Hall, 1990.

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

Capacity of Two-Lane, Two-Way Rural Highways: The New Approach

PLANKO ROZIC

The development of methodology analysis of traffic flow on two-lane, two-way highways is described. The extensive data were collated through 7 days of measuring on the primary road section Zagreb-Velika Gorica (Zagreb Airport). On the basis of overtaking, the new vehicle classification from the aspect of position in traffic flow is defined. By concentrating vehicles into two groups, vehicle file and vehicles involved in overtaking, the four possible combinations of traffic flow condition are introduced. Traffic flows were divided on a time basis, on a functional basis, and from the aspect of car-following. On those bases 24 relationship curves between fundamental traffic stream variables were developed. Different divisions of traffic flows proved that divisions on a time basis do not yield real values of the capacity and have to be rejected. Combinations of traffic flow condition offer a basis for calculation of capacity and levels of service. Classification of flow from the aspect of car-following gives the capacity of a traffic lane and ideal capacity of a highway. It was proved that calculation of capacity depends, in addition to already known elements, on the method of measuring and dividing the traffic flow. Computer programs for graphical presentation of vehicle trajectories in space-time charts, as well as for numerical analyses and graphical depiction of different combinations of traffic flow condition, were developed.

Since the publication of the 1965 *Highway Capacity Manual* (HCM) (*1*), there has been an increasing need for research in highway capacity, which resulted in issuing the new 1985 HCM (*2*). Because of uncritical adoption of the experience gained in the United States, some European countries have embarked on a detailed study of highway capacity.

The 1965 HCM (*1*) defined the ideal capacity of two-lane rural highways as 2,000 pcph, whereas the 1985 HCM (*2*) defined the ideal capacity as 2,800 pcph. Observations on two-lane rural highways undertaken in some European countries have been reported at even higher volumes (*3*). The first researches, undertaken by Zemljic (*4*), Kuzovic et al. (*5*), and Kuzovic (*6*), revealed a discrepancy between American practice at that time and our own.

Most local as well as foreign researches have been based on a search for local factors affecting highway capacity. Contrary to being designed as a critique of the HCM and to seek a range of adjustment factors for our own conditions, the aim of the research was to analyze the possibilities and factors contributing to a high flow rate through practical measurement and to assess the capacity and the ideal capacity of two-lane, two-way rural highways on the basis of a new methodological approach. This methodology was basically developed by the author (*7*) and subsequently fully established and implemented by him (*8*).

Croatian Road Organization, 41000 Zagreb, Voncinina 3, Croatia.

SELECTION OF A TEST SECTION

Basic data on traffic volumes on major two-lane highways in Croatia are being collated by automatic traffic counters (*9*,*10*). Analysis of the AADT and hourly volumes for all locations with automatic counters has shown that the highway with the heaviest volume is the primary road section M-12/2, Zagreb-Velika Gorica (Zagreb Airport). A ratio between hourly volumes and the AADT is presented in Figure 1 (*9*). The Zagreb-Velika Gorica highway meets all highway geometric features for ideal conditions (*2*).

Through in situ observation, a test section length of 960 m was determined, along which there are no obstructions or other restrictions and which is 25 percent over the minimum length for passing sight distance for a design speed of 100 kph (*11*). The test section is out of influence of intersection (i.e., the highway is an uninterrupted flow facility). Such a length of test section, combined with the given roadway elements, permits unrestricted overtaking for vehicles in both directions.

MEASURING OF TRAFFIC FLOW

Good reviews of the definition of traffic stream variables and factors affecting them, techniques of measurement, and application of data are presented in May (*12*), Pignataro (*13*), Gerlough and Huber (*14*), Drew (*15*), and Wattleworth (*16*). French and Solomon (*17*) describe the technology that is being used for traffic data collection.

Measurement of traffic flow on the basis of measurement at a point and measurement of travel time was carried out. Travel time was measured by the license plate method. Observers were positioned at entry and exit of the test section, for each direction separately. One observer recorded the last three digits of the vehicle license plate number, the second observer recorded the time (hour, minute, and second) the vehicle passed through that particular point on the highway, and a third recorded the type of vehicle. Travel times are transformed into individual vehicle speeds.

If Berry and Green's (*18*) concept of a sample is abandoned and all vehicles are included, one would arrive at the travel time, and hence the speed, of all vehicles. On the basis of those speeds it is possible to calculate the space mean speed, which includes all variations in the speed of individual vehicles.

Since vehicle license plate numbers and corresponding times of passage of vehicles are recorded at entry and exit and thus converted into two measurements at a point, it was decided to measure rates of flow at the exits. Through analysis of the sequence of vehicles and their time of entry into and exit from

Road No. M-12/2 Zagreb-Velika Gorica

Traffic counter 39 Velika Gorica

24-hour volume (Both directions)		Volume in selected highest hours as a percentage of AADT							
AADT	PEAK DAY	MAX	10th	20th	30th	40th	50th	100th	200th
19819	29089	2133 10.8%	1984 10.0%	1939 9.8%	1909 9.6%	1884 9.5%	1850 9.3%	1729 8.7%	1615 8.2%

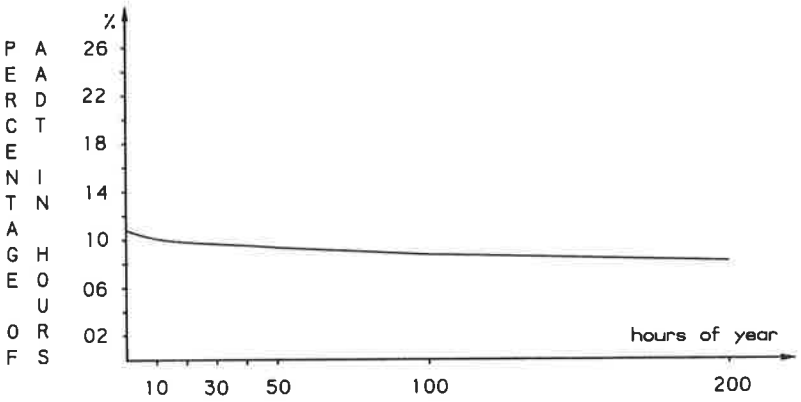


FIGURE 1 Relation of hourly volumes and AADT on the Zagreb–Velika Gorica highway.

the section, the following information was obtained: vehicle type; average vehicle speed; which vehicle overtook, at what speed, and what it overtook; which vehicle kept in its own lane; and headway at entry and exit of test section.

Rate of flow measured at exit of the section can be analyzed in the context of this information. Measurement was carried out through 7 days for a total duration of 14 hr 5 min. All days on which measurement was carried out were sunny, the roadway was dry and in relatively good condition, and it was possible to compare obtained measurement results by days. Vehicle flow was divided according to type of vehicle, as follows (*19*): passenger car, bus, light truck (up to 3 t), medium truck (two-axle, up to 19 t), heavy truck (three-axle, exceeding 19 t), trailer/semi-trailer, and tractor.

A total of 21,559 vehicles were recorded on the test section. Traffic in the direction Zagreb-Velika Gorica (Direction 1) accounts for 62 percent of overall traffic on the average. Passenger cars constitute 89 percent of all vehicles in Direction 1 and 81 percent in Direction 2, which from that aspect brings the traffic flow closer to the ideal.

CLASSIFICATION OF POSITION OF VEHICLES AND TRAFFIC FLOW CONDITIONS ON TWO-LANE, TWO-WAY HIGHWAYS

Vehicle Classification from the Aspect of Position in Traffic Flow

Identification of all vehicles at entry into and exit from the test section makes it possible to establish the sequence of vehicles in both cross sections. In this way five vehicle categories may be defined from the aspect of vehicle position in traffic flow, as shown in Figure 2:

1. Vehicles that keep in their own lane and are not involved in overtaking and are not positioned between two adjacent maneuvers of overtaking vehicles traveling in the same direction (two vehicles, one of which is completing an overtaking maneuver and the second of which is beginning the maneuver);
2. Vehicles passive in overtaking—those keeping in their own lane, not being overtaken, and finding themselves between two adjacent overtaking maneuvers;
3. Vehicles active in overtaking—those overtaking;
4. Vehicles active in overtaking—those overtaking and themselves being overtaken; and
5. Vehicles passive in overtaking—those keeping in their own lane and being overtaken.

Classification of Traffic Flow Conditions

A concentration of vehicles of Category 1 results in a group of vehicles annotated as "vehicle file." Vehicle file is understood here to be a moving line of vehicles regardless of headway between them. It must be differentiated from the car-following or platoon mode. "Percent time delay" is defined as the average percentage of time that all vehicles are delayed while traveling in platoons due to inability to pass (*2*). On

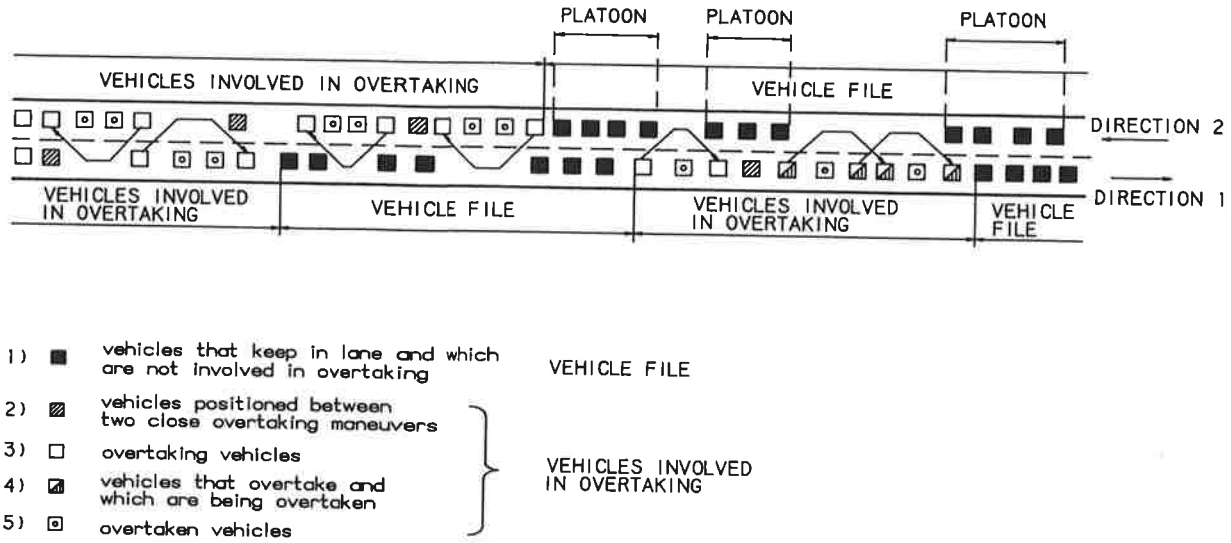


FIGURE 2 Definition of vehicle position within traffic flow and traffic flow conditions: vehicle file and vehicles involved in overtaking.

the basis of extensive field measurement (*7,8*) it could be stated that numerous vehicles were forced to keep in their own lane and at the same time they were not involved in platoons. From the aspect of car-following, Ovuworie et al. (*20*) divided vehicles into three categories:

1. Free-moving (independent) vehicles—those moving without being influenced by other vehicles;
2. Partially independent vehicles—those either joining or leaving the platoon; and
3. Vehicles in platoon—those following a lead vehicle.

Vehicle file may consist of a number of vehicles of all three categories. The definition of platoon in the 1985 HCM (*2*, Chapter 8) corresponds to the definition of vehicle file in this paper.

Analogous to vehicle file, all vehicles between two vehicle file (Vehicle Categories 2 through 5) form a group of vehicles annotated as "vehicles involved in overtaking." Concepts of vehicle file and vehicles involved in overtaking are presented in Figure 2.

Such an approach to the analysis affords the possibility for new combinations of traffic flow condition on a two-lane, two-way highway, which are presented in Table 1.

When both directions contain vehicles involved in overtaking (Combination 1 in Table 1), overtaking can be mutually executed, in which case flow rate results are shown for both directions together. Through an increase of rate of flow, headways, after a certain level of flow rate, acquire such values that safe overtaking is no longer possible. At that point the traffic flow on a two-lane, two-way highway is transformed into two one-way traffic flows (i.e., into vehicle file in both directions concurrently). According to the 1985 HCM (*2*), this combination begins at level of service D (i.e., at flow rate above 1,200 pcph, total in both directions). From Figure 3 it can be seen that real grouping of such a combination begins at approximately 800 vph.

Vehicle Classification from the Aspect of Car-Following

From the aspect of car-following where a stimulus-response link exists between the lead and trailing vehicles, traffic flows may be divided as follows:

- Platoon with mixed vehicles, located within a combination of traffic flow condition: vehicle file-vehicle file; and
- Platoon with passenger cars only, located within a platoon with mixed vehicles.

Platoon with Mixed Vehicles

A platoon is by its very nature a subgroup of a vehicle file. Extreme cases can occur, where a whole vehicle file is at one and the same time a platoon, and also where there is not one platoon within a vehicle file. Within a platoon it is not even theoretically possible that overtaking vehicles from opposing directions could appear, since they automatically break the stimulus-response link between vehicles in the platoon. Combination 4 comprising a vehicle file in both directions simultaneously thus becomes the first prerequisite for the appearance of a platoon. The second prerequisite is that vehicles must follow a lead vehicle.

There is no unique definition of platoon. In addition to the 1965 HCM (*1*), Cunagin and Chang (*21*) and Radwan and Kalevela (*22*) did not consider headways exceeding 9 sec. Miller (*23*) separated platoons if the headway was longer than 8 sec. Edie et al. (*24*) adopted the criterion of headway from 4 to 5 sec, depending on the speed of vehicles. Keller (*25*) was even more rigid, with a headway of less than 2 sec. The 1985 HCM (*2*) defines vehicles following at headway less than 5 sec. On the basis of research by Chrissikopoulos et al. (*26*) the criterion was adopted whereby a short platoon is one with two to four vehicles, and a long platoon is one with five or

TABLE 1 Combinations of Traffic Flow Condition on Two-Lane, Two-Way Highways

Ord. No.	Direction 1	Direction 2	Descriptions of conditions of traffic flow progress
1	VEHICLES INVOLVED IN OVERTAKING	VEHICLES INVOLVED IN OVERTAKING	Headways and corresponding spacing in both directions ensure safe overtaking for vehicles travelling in either direction. Flow rates are considerably below capacity with traffic flow speeds and densities that correspond to the conditions of noncongestion traffic.
2	VEHICLE FILE	VEHICLES INVOLVED IN OVERTAKING	Headways and spacings in direction 1 allow vehicles from direction 2 to cross to the lane used by vehicles in direction 1 and to execute an overtaking maneuver. In direction 1, density and rates of flow are slightly higher than in the combination No.1. In direction 2, rate of flow, speed and density retain the characteristics from the combination of traffic flow condition No. 1.
3	VEHICLES INVOLVED IN OVERTAKING	VEHICLE FILE	Situation is now the reverse of combination No. 2, i.e. conditions from direction 1 have been transposed into direction 2, and vice versa.
4	VEHICLE FILE	VEHICLE FILE	Headways and spacings in both directions have acquired such values that safe overtaking is not possible for vehicles travelling in either direction. Traffic flow has been transformed into two one-way, opposite flows. Rates of flow are within a wider area of capacity with the accompanied corresponding speeds and densities.

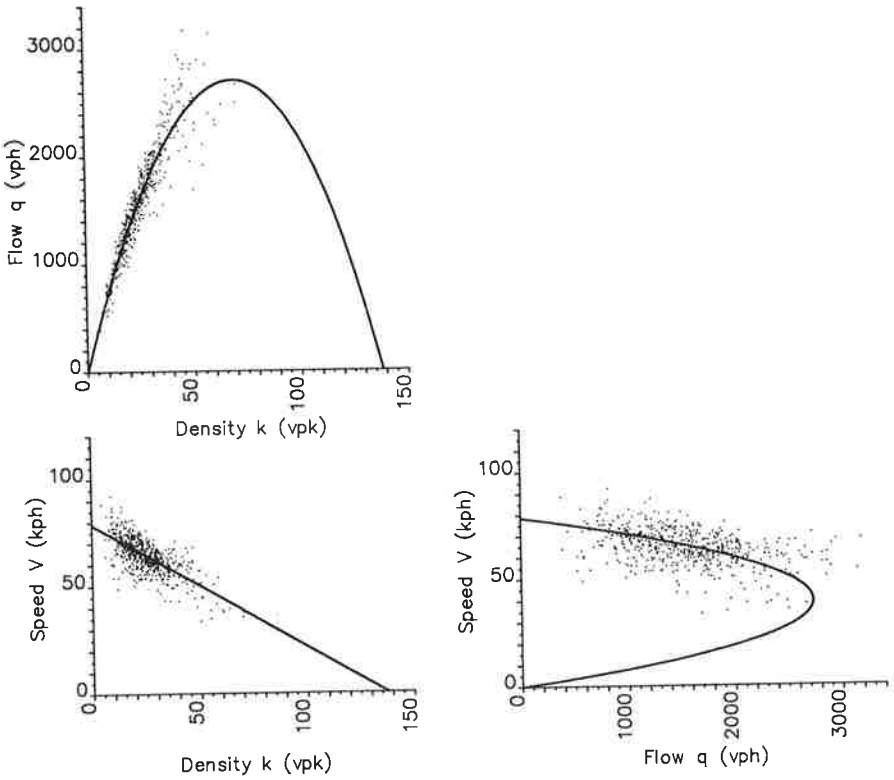


FIGURE 3 Division of traffic flow on a functional basis: all combinations of traffic flow condition taken together—both directions together.

more vehicles. In this research the headway of less than or equal to 9 sec at both entry and exit of section was adopted.

Platoon with Passenger Cars Only

Platoons consisting of passenger cars only are formed in this research in such a way that all platoons having a passenger car as a lead vehicle are separated. Among such platoons a further distinction was made whereby all consecutive platoons that comprised five or more passenger cars were separated. A platoon of passenger cars ends with the first replacement of a passenger car by some other type of vehicle.

Both the 1965 HCM (*1*) and the 1985 HCM (*2*) define the ideal capacity with an idealized flow in which there are only passenger cars. Before rates of flow achieve maximum level, flows start to run in the combination of traffic flow Condition 4 (i.e., of the concurrent vehicle file in both directions). At near-to-capacity flow rates of two-lane, two-way highways, long platoons are formed in both directions. If platoons are formed of passenger cars only, a macroscopic evaluation will give a maximum flow rate (i.e., the ideal capacity of two-lane, two-way highways).

DATA PROCESSING

The software for the graphical presentation of the vehicle trajectories, different divisions of the same traffic flow, and the calculation of flow rate, density, space mean speed, and time mean speed, together with a statistical analysis of mean speeds, was developed and applied to traffic flows.

For all 7 days, seven separate diagrams of a total length of 51 m were made. The chart was produced using a CALCOMP 1075 A plotter. A sample was extracted from the data base on the seventh day of measuring, which is suitable for explanation of all elements in the data base and is presented in the space-time chart in Figure 4. The graphical presentation contains a separate space-time chart for each direction. The diagram abscissa is the time axis with its smallest division being

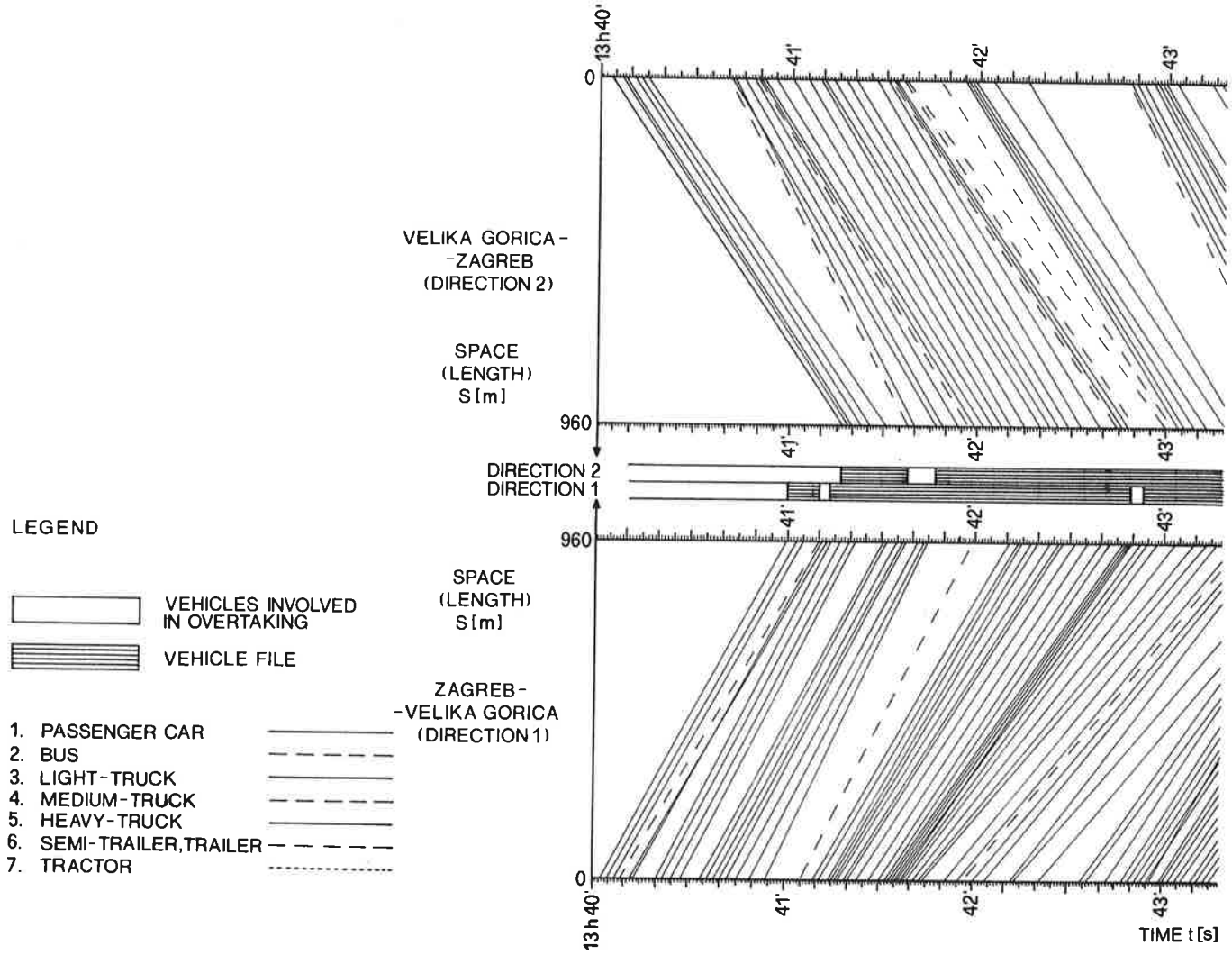


FIGURE 4 Vehicle trajectories recorded on the section of primary road M-12/2 Zagreb–Velika Gorica, Monday, 04/05/1982.

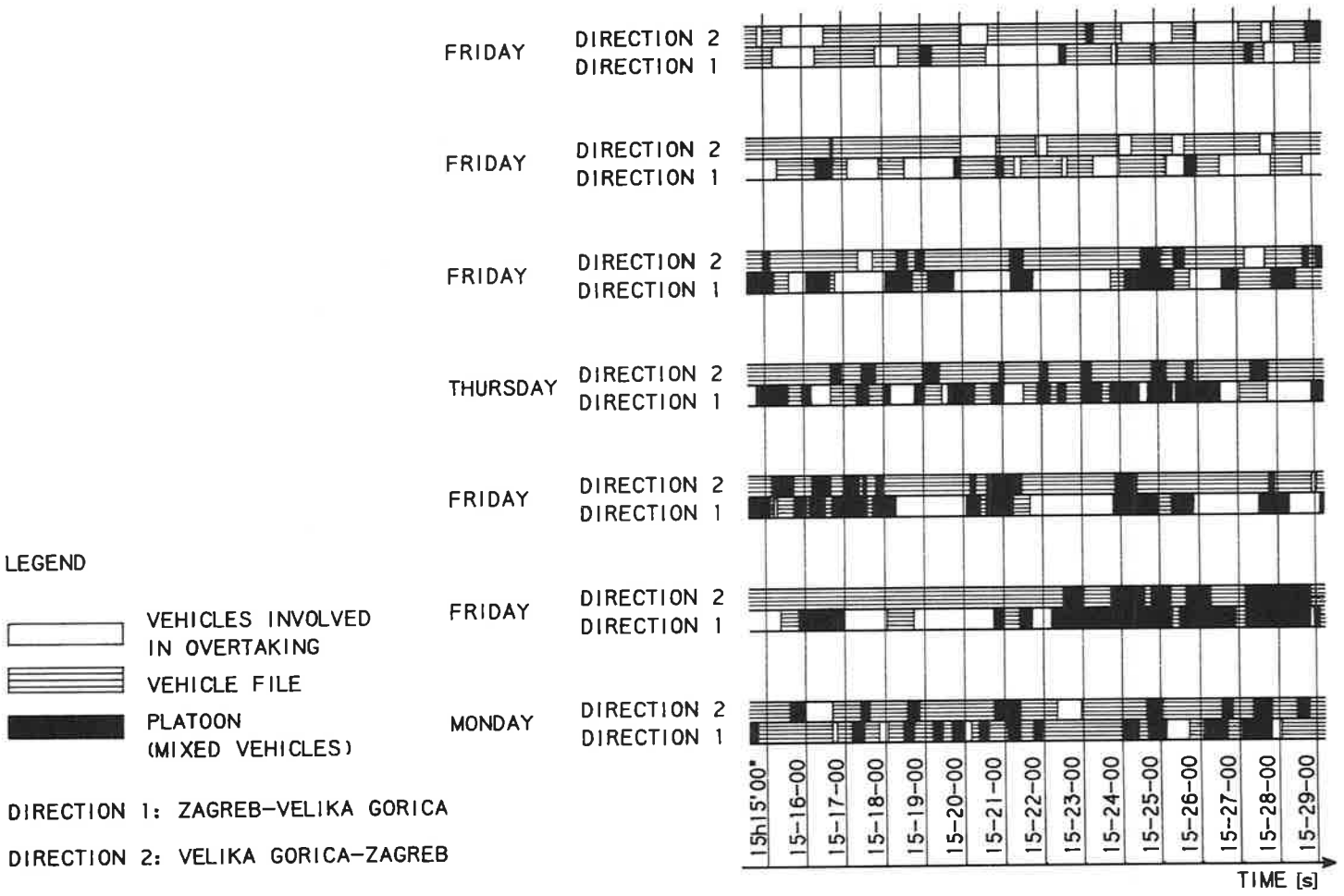


FIGURE 5 Combinations of traffic flow condition by days of measurement on the section of primary road M-12/2 Zagreb–Velika Gorica.

1 sec, whereas the ordinate shows the space (length of test section is 960 m). Although the space is common to both directions (in the sense of test section length), charts are separated to facilitate easier understanding. The vehicle trajectories depicted in different forms and colors describe vehicle movement in space and time. The trajectories are straight, since the measurement method disregarded variations in vehicle speed within the test section (i.e., the average speed is presented). Combinations of traffic flow condition are presented on the central axis between two separated space-time charts. Duration of vehicle file is measured at the exit from the test section, where rate of flow is also measured.

Combinations of traffic flow condition are partly presented graphically in toto in Figure 5 for all days on which measuring was carried out by transposing the central axes of individual space-time diagrams. Platoons are presented graphically in Figure 5 by shaded areas within vehicle file.

MODELING OF TRAFFIC STREAM VARIABLES

Some of the better-known traffic stream models are the Greenshields linear model (*27*), the Greenberg logarithmic model (*28*), the Underwood exponential model (*29*), the Drake bell-shaped model (*30*), the Drew parabolic model (*15*), and the Edie multiregime model (*31*). Analyses that Drake et al. (*30*) and Radwan and Kalevela (*22*) carried out by means of the listed models showed that no unique model exists that is capable of satisfying all boundary conditions while giving realistic values of the fundamental traffic stream variables, where the same ones differ even on the same types of highway.

The Greenberg model (*28*) can probably be applied to vehicle files on two-lane, two-way highways (Combination 4). Contrary to the Greenberg model (*28*), it is assumed that the Underwood model (*29*)—applicable to traffic flow with low density and high speed—should satisfy those conditions on two-lane, two-way highways. Those conditions consist of Combination 1 and, possibly, of a part of the area covered by Flow Combinations 2 and 3.

Because of the need for comparison between different approaches and divisions of a traffic flow, and because of its simplicity, the Greenshields linear model (*27*)—giving a continual curve (i.e., meeting all three boundary conditions)—was selected. Relationship curves were generated in the following manner:

- A linear relationship (straight line) of speed and density was defined by means of the least squares method (*32*), and
- Relationships between the flow rate and density and between flow rate and speed (parabola) were defined through inclusion of the equation of straight line into a fundamental equation of the traffic flow (i.e., $q = k \times V$).

Analyses of relationship between fundamental traffic stream variables were produced separately by directions and for both directions together. For purposes of analysis, 24 different models were generated. The most indicative were selected and are presented in this paper.

ANALYSIS OF TRAFFIC FLOW WITHIN THE TEST SECTION

Classification of Traffic Flow

On the basis of the newly established classification of traffic flows on two-lane, two-way highways, traffic flow was divided into subflows on the basis of different approaches: on a time basis, on a functional basis, and from the aspect of the car-following.

Three methods of dividing the same traffic flow yielded a different number of subflows, which are presented in Table 2. The number of vehicles participating in individual subflows is also of interest, as can be seen in Table 3.

Division of Traffic Flow on a Time Basis

The time basis for division of traffic flows is 5-, 15-, and 60-min intervals. The number of subflows is a function of the time basis upon which the flow is divided. Flows were analyzed separately for Direction 1, Direction 2, and for both directions together.

When division is on a time basis, the sum of vehicles in subflows is always equal, although it differs in particular intervals within a division—which is the result of traffic variations in space and time.

The capacity levels obtained on the basis of both the 5- and 15-min intervals amount to 2,400 vph, total in both directions, and 2,300 vph, total in both directions on the basis of 60-min intervals. The comparison of models (curves) for 5-, 15-, and 60-min intervals for each direction separately and for both directions together indicates that differing capacity values are the result of the size of the time base upon which the traffic flow is divided. The differences widen with an increase in traffic volume.

TABLE 2 Number of Subflows Obtained on the Basis of Different Methods of Dividing the Same Traffic Flow

Division of traffic flows	Direction 1	Direction 2	Direction 1+2
Five-minute intervals[a]	169	169	169
Fifteen-minute intervals[a]	153	153	153
Sixty-minute intervals[a]	84	84	84
Overtaking-overtaking[b]	7	7	7
Overtaking-file[b]	150	30	180
File-overtaking[b]	30	150	180
File-file[b]	356	356	356
Platoon (mixed vehicles)[c]	519	353	–[d]
Platoon (passenger cars only)[c]	299	114	–[d]

[a]division of traffic flows on a time basis
[b]division of traffic flows on a functional basis
[c]division of traffic flows from the aspect of car-following
[d]data not applicable

TABLE 3 Number of Vehicles in Individual Subflows and Their Participation in Total Flow and Some Subflows

Division of traffic flows	Direction 1	Direction 2	Direction 1 + 2
1 Total number of vehicles (total flow)	13351	8208	21559
2 Five-minute intervals[a]	13351	8208	21559
Participation in total flow (2 / 1)	100%	100%	100%
3 Fifteen-minute intervals[a]	13351	8208	21559
Participation in total flow (3 / 1)	100%	100%	100%
4 Sixty-minute intervals[a]	13351	8208	21559
Participation in total flow (4 / 1)	100%	100%	100%
5 Overtaking - overtaking (O - O)[b]	285	158	443
Participation in total flow (5 / 1)	2%	2%	2%
6 Overtaking-file (O - F)[b]	4003	601	4604
Participation in total flow (6 / 1)	30%	7%	21%
7 File-overtaking (F - O)[b]	774	1823	2597
Participation in total flow (7 / 1)	6%	22%	12%
8 File-file (F-F)[b]	8289	5626	13915
Participation in total flow (8 / 1)	62%	69%	65%
9 (O - O)+(O - F)+(F - O)+ (F - F)[b]	13351	8208	21559
Participation in total flow (9 / 1)	100%	100%	100%
10 Platoon (mixed vehicles)[c]	6557	3915	—[d]
Participation in total flow (10 / 1)	49%	48%	
Participation in file-file (10 / 8)	79%	70%	
11 Platoon (passenger cars only)[c]	2820	829	—[d]
Participation in total flow (11 / 1)	21%	10%	
Participation in file-file (11 / 8)	34%	15%	
Participation in platoon (mixed vehicles) (11 / 10)	43%	21%	

[a]division of traffic flows on a time basis
[b]division of traffic flows on a functional basis
[c]division of traffic flows from the aspect of car-following
[d]data not applicable

Division of Traffic Flow on a Functional Basis

Division of traffic flow on a functional basis involves four combinations of traffic flow condition (see Table 1). All four combinations were analyzed for Direction 1, Direction 2, and both directions together.

Functional division of the traffic flow indicates that Combination 1 (overtaking-overtaking) participated with only 2 percent of vehicles. The other extreme is that Combination 4 (file-file) participated in traffic flow from 62 to 69 percent of all vehicles.

The overtaking-overtaking combination of traffic flow condition has not been processed as a single category, since with only seven combinations (see Table 2) it does not constitute a sufficiently large sample to permit statistical analysis. Combinations 2 and 3 (file-overtaking and overtaking-file) were dealt with as a single category for both directions together, because the highway meets the requirements for ideal conditions, and in this case the direction from which a vehicle overtakes loses its importance. The result arrived at was a capacity of 2,700 vph, total in both directions. Combination 4 (file-file) produces a highway capacity of 2,700 vph, total in both directions. All four combinations of traffic flow conditions make up the parts of the same traffic flow and as such are mutually exclusive. For this reason all combinations were also processed together as a single category, the result being a capacity of 2,700 vph, total in both directions, as shown in Figure 3. From a statistical point of view the most important combination in this research is the file-file combination of traffic flow condition, which is proved by the coinciding results obtained from processing of the file-file combination and all combinations of traffic flow conditions together.

Division of Traffic Flow from the Aspects of Car-Following

Platoons with mixed vehicles and platoons with passenger cars only have been analyzed simply by direction since they appear only after traffic flows on two-lane, two-way highways become separated into two opposite single-way flows (two vehicle files). In addition, despite the high flow rates, no platoons of significant duration appeared in both directions simultaneously to make possible a simultaneous comparison in both directions together. From 519 platoons in Direction 1, 299 (or 58 percent) began as platoons with passenger cars, whereas in Direction 2, 114 (or 32 percent) of 353 were platoons with passenger cars.

The results of analysis of platoons with mixed vehicles showed a capacity of 2,150 vphpl in Direction 1 and 2,000 vphpl in

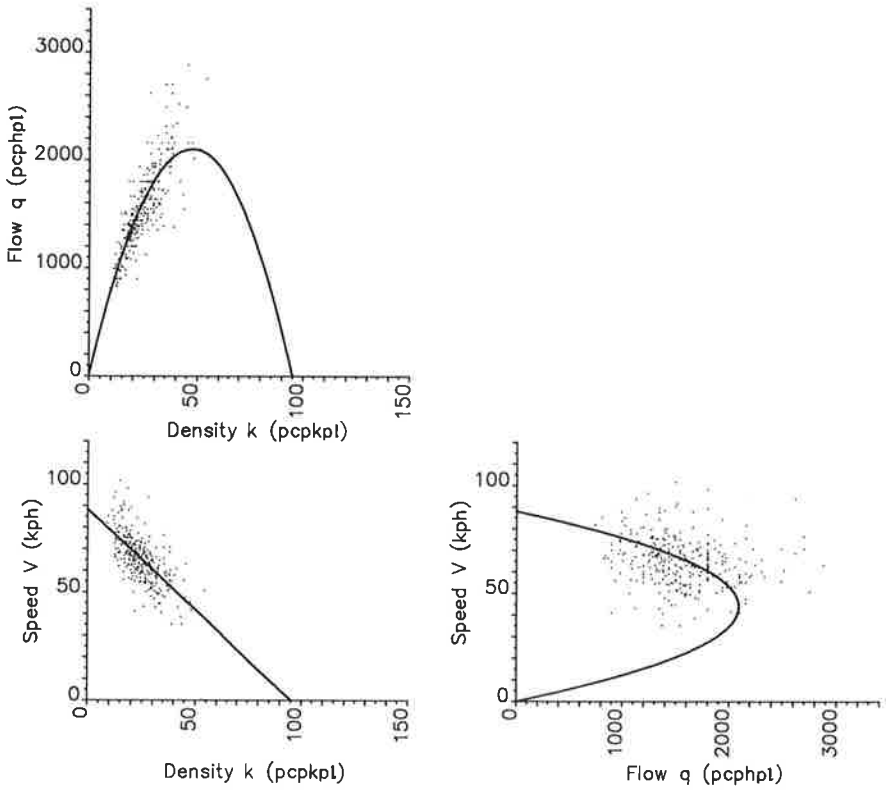


FIGURE 6 Division of traffic flow from the aspect of car-following: platoon (passenger cars only) Direction 1.

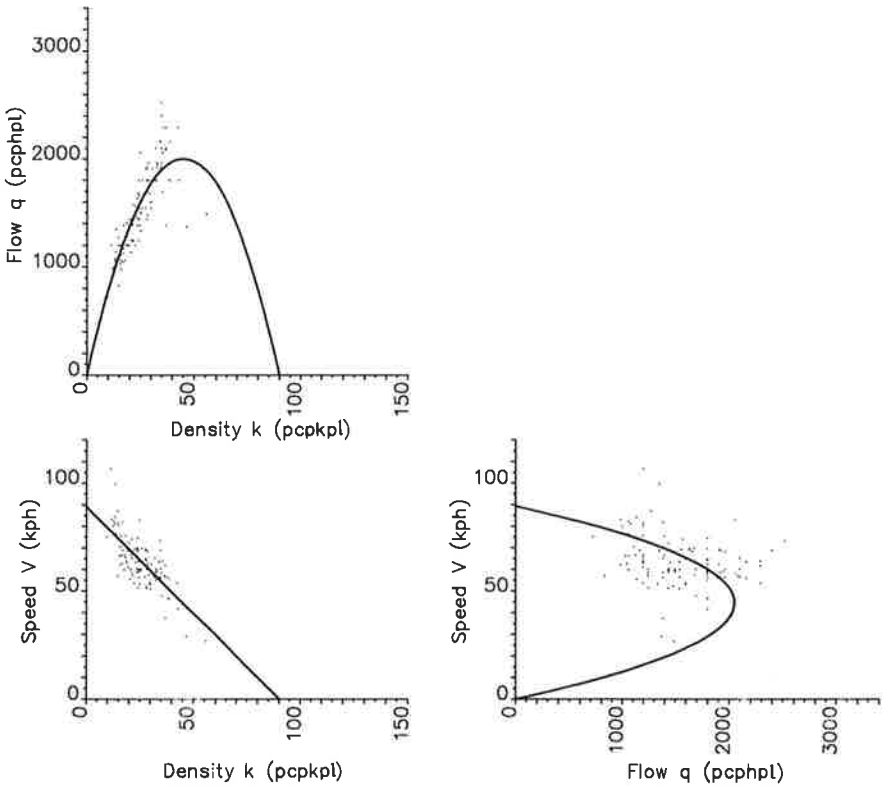


FIGURE 7 Division of traffic flow from the aspect of car-following: platoon (passenger cars only) Direction 2.

Direction 2. Platoons with passenger cars only produced a capacity of 2,100 pcphpl in Direction 1 and 2,050 pcphpl in Direction 2, as shown in Figures 6 and 7.

ANALYSIS OF RESULTS

The results lead to the conclusion that the capacity of a two-lane rural highway with a realistic composition of traffic flow in roadway conditions approaching the ideal can be expected to be about 2,700 vph, total in both directions. Results correspond to more recent measurements obtained in Finland (*33*).

Since it became apparent during the course of defining the vehicle file that there may be gaps that can be filled with other vehicles, it is realistic to expect that such classifications can yield values exceeding 3,000 vph, total in both directions. Those gaps are filled if all vehicles in the vehicle file link up in a dynamic way (i.e., if, in the ultimate case, the entire vehicle file becomes a long platoon). In that way the traffic flows on a two-lane, two-way highway convert into two platoons moving in opposite directions, when the capacity of a single traffic lane is defined, and the capacity of the highway as a whole is the sum of the capacities of two traffic lanes. In this particular research such an ideal capacity can be assessed at 4,000 pcph, total in both directions. This means that the capacity of 2,800 pcph may be considered as conservative. It confirmed Yagar's concept that flows approaching 4,000 pcph are possible (*34*).

Numerous studies have been made in platoon research in various conditions. Some of them analyze platoons in one cross section. The space-time charts developed in this research proved that in platoon analysis at least two cross sections should be established. This means that a headway of less than 5 sec, as a criterion for definition of a platoon and percent time delay, cannot be used as surrogate measure in field studies. Two cross sections analyses also make it possible to define a free-moving vehicle and thus to be able to calculate free-flow speed in actual conditions.

CONCLUSION

The depiction of vehicle trajectories in space-time charts as well as numerical and graphical presentation of three different divisions of the same traffic flow enable us to undertake detailed analysis of traffic flow within various traffic, roadway, and environmental conditions. The methodology of highway capacity assessment presented is indifferent to the traffic volume. It requires a large number of people and is therefore costly. With the electronic equipment available today (e.g., video cameras), following the philosophy of this methodology makes its realization more accurate, easier, and much less expensive.

Divisions of traffic flows on a time basis do not yield real values of maximum flow rates (i.e., the capacities of two-lane, two-way highways), and therefore these divisions must be rejected. The sample in question was sufficiently large, which means that neither further measuring nor extension of the sample would produce significantly different curves.

Classification of traffic flows by combinations of traffic flow condition offers a basis for calculation of capacity and levels of service since all combinations together cover the entire area of the curve designating the noncongested flow.

Classification of traffic flows from the aspect of car-following gives the capacity of a traffic lane and the ideal capacity of two-lane, two-way highways. Ideal capacity occurs during the appearance of platoons with passenger cars only, in both directions at the same time. This being an event that rarely occurs in practice over a longer period of time, an assessment must be accepted.

Different forms of parabola with different vertices of the parabola show that the basic diagram of the traffic flow does not only depict characteristics of place and time of research and driver population, but also the method of measurement and division of the traffic flow.

Computer programs developed step by step in accordance with work progression become uneconomical. Production of a program package for computer processing with a possibility for the iterant method of analysis—which is a task for the future—will make it possible to establish with certainty the real values of all parameters and criteria influencing the values of fundamental variables and, in parallel with that, the highway capacity. Consequently, the values obtained through application of the Greenshields model have to be taken as an assessment and not as being absolutely correct.

ACKNOWLEDGMENTS

The valuable suggestions and support of Vlasto Zemljic during the research are greatly appreciated. Many of my students of the Faculty of Transportation Engineering, University of Zagreb, have contributed to extensive traffic flow measurements and preliminary evaluation of collated data. I am indebted to all of them.

REFERENCES

1. *Special Report 87: Highway Capacity Manual*. HRB, National Research Council, Washington, D.C., 1965.
2. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
3. *Two-Lane Rural Roads: Design and Traffic Flow*. Organization for Economic Cooperation and Development, Paris, 1972.
4. V. Zemljic. *Driving Speed in a Platoon and Its Influence on Capacity and Safety* (in Slovenian). Ph.D. thesis, Ljubljana, Slovenia, 1974.
5. L. J. Kuzovic et al. *Defining Speed of a Traffic Flow at Capacity Depending on Influencing Factors and Quantification of Those Factors in Our Roadway and Traffic Conditions* (in Serbian). Institute of the Traffic Faculty, Belgrade, 1980.
6. L. J. Kuzovic. *Defining the Basic Parameters Determining Highway Capacity and Their Quantification for Roadway and Traffic Conditions in Yugoslavia* (in Serbian). Institute of the Traffic Faculty, Belgrade, 1979.
7. P. Rozic. *Research into Parameters of Traffic Flow on Two-Lane Highways* (in Croatian). Master's thesis. University of Zagreb, Zagreb, Croatia, 1987.
8. P. Rozic. *Capacity on Two-Lane, Two-Way Highways in Our Conditions* (in Croatian). Ph.D. thesis. Edvard Kardelj University of Ljubljana, Ljubljana, Slovenia, 1989.
9. P. Rozic and M. Herak. *Processing of Data Obtained Through*

Traffic Counting Utilizing Automatic Traffic Counters (in Croatian). Croatian Road Organization, Zagreb, Croatia, 1980.
10. F. Mihoci and P. Rozic. Organization of the Collation and Processing of Data on Traffic Using the Highways of SR Croatia (in Croatian). *Contemporary Traffic*, Vol. 6, No. 6, Zagreb, Croatia, 1984, pp. 404–408.
11. *Regulations Manual on Conditions To Be Met from the Aspect of Traffic Safety by Rural Public Roads and Their Elements* (in Serbian). Federation of Associations for the Highways of Yugoslavia, Belgrade, 1981.
12. A. D. May. *Traffic Flow Fundamentals*. Prentice Hall, Englewood Cliffs, N.J., 1990.
13. L. J. Pignataro. *Traffic Engineering Theory and Practice*. Prentice-Hall, Englewood Cliffs, N.J., 1975.
14. D. L. Gerlough and M. J. Huber. *Special Report 165: Traffic Flow Theory*. TRB, National Research Council, Washington, D.C., 1975.
15. D. R. Drew. *Traffic Flow Theory and Control*. McGraw-Hill, New York, 1968.
16. J. A. Wattleworth. Traffic Flow Theory. In *Transportation and Traffic Engineering Handbook*. Institute of Traffic Engineers, Prentice-Hall, Englewood Cliffs, N.J., 1976, pp. 258–308.
17. A. French and D. Solomon. *NCHRP Report 130: Traffic Data Collection and Analysis: Methods and Procedures*. TRB, National Research Council, Washington, D.C., 1986.
18. D. Berry and F. M. Green. Evaluation Techniques for Measuring Overall Speeds in Urban Areas. *HRB Proc.*, Vol. 29, 1949, pp. 311–318.
19. *Internationaler Nutzfahrzeug-Katalog*. Vogt Schild Ltd., Solothurn 1, 1979.
20. G. C. Ovuworie, J. Darzentas, and M. R. C. McDowell. Free Movers, Followers and Others: A Reconsideration of Headway Distributions. *Traffic Engineering and Control*, Vol. 21, No. 8/9, 1980, pp. 425–428.
21. W. D. Cunagin and E. C. Chang. Effects of Trucks in Freeway Vehicle Headways Under Off-Peak Flow Conditions. In *Transportation Research Record 869*, TRB, National Research Council, Washington, D.C., 1982.
22. S A. E. Radwan and A. F. Kalevela. Investigation of the Effect of Change in Vehicular Characteristics on Highway Capacity and Level of Service. In *Transportation Research Record 1005*, TRB, National Research Council, Washington, D.C., 1985, pp. 65–71.
23. A. J. Miller. A Queuing Model for Road Traffic. *Journal of Royal Statistics Society*, Vol. B23, 1961, pp. 64–75.
24. L. C. Edie, R. S. Foote, R. Herman, and R. Rothery. Analysis of Single Lane Traffic Flow. *Traffic Engineering*, Jan. 1963, pp. 21–27.
25. H. Keller. Effects of a General Speed Limit on Platoon of Vehicles. *Traffic Engineering and Control*, Vol. 17, No. 7, July 1976, pp. 300–303.
26. V. Chrissikopoulos, J. Darzentas, and M. R. C. McDowell. Aspects of Headway Distributions and Platooning on Major Roads. *Traffic Engineering and Control*, Vol. 21, No. 5, May 1982, pp. 268–271.
27. B. D. Greenshields. A Study in Highway Capacity. *HRB Proc.*, Vol. 14, 1934, pp. 448–477.
28. H. Greenberg. Analysis of Traffic Flow. *Operations Research*, Vol. 7, No. 1, 1959, pp. 79–85.
29. R. T. Underwood. *Speed, Volume and Density Relationships, Quality and Theory of Traffic Flow*. Bureau of Highway Traffic, Yale University, New Haven, Conn., 1961, pp. 141–187.
30. J. Drake, J. Schofer, and A. D. May, Jr. A Statistical Analysis of Speed-Density Hypothesis. *Proc., Third International Symposium on Theory in Traffic Flow*, American Elsevier, New York, 1967, pp. 112–117.
31. L. C. Edie. Car-Following and Steady-State Theory for Non-Congested Traffic. *Operations Research*, Vol. 9, No. 1, 1961, pp. 66–76.
32. J. R. Benjamin and C. A. Cornell. *Probability, Statistics and Decision for Civil Engineers*. McGraw-Hill, New York, 1970.
33. M. Pursula and A. Enberg. Characteristics and Level of Service Estimation of Traffic Flow on Two-Lane Rural Roads in Finland. Presented at the 70th Annual Meeting of the Transportation Research Board, Washington, D.C., 1991.
34. S. Yagar. *Capacities for Two-Lane Highways*. Australian Road Research, Vol. 13, No. 1, Vermont South, 1983, pp. 3–9.

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

Study of Headway and Lost Time at Single-Point Urban Interchanges

JAMES A. BONNESON

The results of a recent study of the headway and lost time at three single-point urban interchanges (SPUIs) are summarized. The data base, containing more than 38,000 headway observations, was collected primarily in the Tampa, Florida, area. The data were used to calculate the minimum discharge headway and start-up lost time for the SPUI's three basic movements: cross road left-turn, off-ramp left-turn, and cross road through. It was found that traditional procedures for estimating the minimum discharge headway may be biased toward values higher than ultimately achieved by the traffic queue. Moreover, the degree of bias varied widely among the movements and sites studied because of unequal numbers of observations. As a result, initial attempts at a cause-and-effect analysis were clouded by a high degree of variability in the data. In recognition of the aforementioned bias, alternative statistical analysis techniques and regression models were used to identify significant effects and to calibrate predictive models of minimum discharge headway and start-up lost time. The results indicate that the minimum discharge headway of the SPUI's two left-turn movements are significantly lower than its through movements and lower than values traditionally used for protected left-turn movements under "ideal" conditions. In fact, the calibrated models predict minimum discharge headways that are generally lower, and start-up lost times that are higher, than those calculated by traditional procedures. Left-turn headway was also found to vary with turn radius.

Within the past two decades a new type of interchange, the single-point urban interchange (SPUI), has emerged in response to increasing urban traffic demands. In some ways this new interchange (shown in Figure 1) is similar to a diamond interchange, and in other ways it is similar to a high-type at-grade signalized intersection. The most distinctive feature of this interchange is the convergence of all through and left-turn movements into a single, signalized conflict area. The advantage of this feature is that all movements can be served by a single signal with, at most, one stop required to clear the intersection. In contrast, a diamond interchange requires two separate signalized junctions (one for each on/off ramp), which presents the possibility of being stopped twice while traveling through the interchange.

Messer et al. (*1*) indicate that the SPUI design was first considered in the United States in the mid-1960s and that the first SPUIs were constructed in the early 1970s. Messer et al. (*1*) also cite evidence that several European countries also constructed SPUIs in the early 1970s. To date, there are only about 40 operational SPUIs in the United States.

This paper summarizes a portion of the results of a larger study of the operational efficiency of the SPUI (*2*). In particular, this paper describes the statistical analysis of headway and lost time data collected at three SPUIs. The analysis identifies and quantifies factors affecting the discharge headway and lost time of SPUI traffic queues; it also compares the SPUI data with similar data collected at two at-grade intersections (AGIs).

Civil Engineering Department, University of Nebraska-Lincoln, Lincoln, Nebr. 68588-0531.

BACKGROUND—CHARACTERISTICS OF SIGNALIZED JUNCTIONS

Discharge Headway

Soon after the start of the signal phase, a traffic queue begins to discharge past the stop line. The time headways between vehicles are initially large but eventually they converge to a minimum discharge headway (H). The basic trend of convergence to a relatively constant, minimum headway is recognized in the 1985 *Highway Capacity Manual* (HCM) (*3*, Chapter 2). Calculation of the minimum discharge headway is typically accomplished by averaging the headways of those queue positions that are relatively constant. In this regard, the 1985 HCM suggests that a constant headway is reached by the fifth queue position. Thus, the minimum discharge headway is calculated by averaging headways for the fifth and subsequent queue positions.

Another commonly used term to describe the service time of a traffic queue is saturation flow rate. Saturation flow rate is calculated as $3{,}600/H$ and has units of vehicles per hour green per lane (vphgpl). Under ideal operating conditions (i.e., 12-ft lanes, all through vehicles, all passenger cars, level grade, no parking, and no pedestrian activity), the 1985 HCM (*3*, Chapter 9) recommends the use of 1,800 vphgpl for the saturation flow rate of a traffic lane at a signalized intersection, unless local field measurements prove otherwise. This value corresponds to a minimum discharge headway of 2.0 sec/veh.

In recognition of the difference between through and turning driver behavior, the 1985 HCM (*3*, Chapter 9) recommends a saturation flow rate for turn movements that is lower than that used for through movements. In particular, the 1985 HCM recommends that the saturation flow rates of right- and left-turn movements in exclusive lanes with protected phases be 85 and 95 percent, respectively, of the saturation flow rate for through movements. Similarly, the 1985 HCM recommends that dual-lane right- and left-turn movements be 75 and 92 percent, respectively, of the through saturation flow rate.

A recent study of the effect of radius on the saturation flow rate of turning vehicles was conducted by Kimber et al. (*4*).

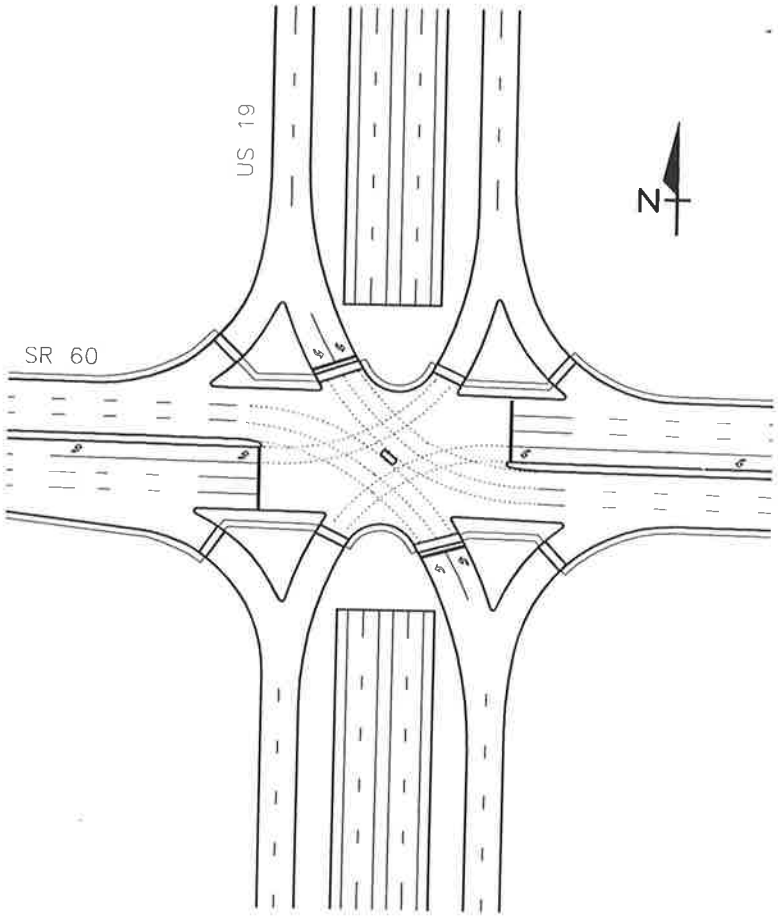


FIGURE 1 Typical geometric configuration of the SPUI.

On the basis of their research, Kimber et al. recommended the following relationship between radius and saturation flow rate for a left-turn movement:

$$S_t = \frac{2{,}080}{1 + \frac{4.92}{R}} \quad (1)$$

where S_t is saturation flow rate of a turn movement (vphgpl) and R is radius of curvature (feet).

Start-Up Lost Time

The first few vehicles in a traffic queue have headways in excess of the minimum discharge headway as their drivers accelerate to a desired speed. This excess time is commonly referred to as lost time because it represents time that is inefficiently used by the discharging traffic queue. Start-up lost time can be calculated by adding the individual lost time for these first few starting vehicles. The equation for start-up lost time is

$$K_s = \sum_{n=1}^{N} (h_n - H) \quad (2)$$

where

K_s = start-up lost time (sec/phase),
h_n = headway of the nth queued vehicle (sec),
H = minimum discharge headway (sec/veh), and
N = number of queue positions having headways larger than H.

Equation 2 indicates that the magnitude of start-up lost time is directly dependent on the value used for minimum discharge headway (H).

As discussed previously, the 1985 HCM (*3*, Chapter 9) recommends that the minimum discharge headway be calculated as the average of the headways for the fifth through last queued vehicles. This approach implies that the first four vehicles incur all of the start-up lost time (i.e., $N = 4$). The 1985 HCM (*3*, Chapter 2) indicates that start-up lost time is generally about 2.0 sec/phase.

EXPERIMENTAL DESIGN

Headway and lost time for the three basic movement types found at the SPUI were examined. The basic movement types included the cross road left-turn movement, cross road through movement, and off-ramp left-turn movement.

(*3*, Chapter 9). However, the results are not totally satisfying because there are no obvious trends where trends are expected and, in one case, the trend found is contrary to recommended practice.

Potential Bias in Headway and Lost Time Estimation

A closer examination of the data was undertaken to explore the causes of the wide variability in the tabulated results. This examination focused on the possibility that the method used to calculate minimum discharge headway had some inherent biases. In particular, headways averaged for each queue position indicated a trend toward decreasing values through the first 8 to 10 queue positions at most locations. This trend suggests that a minimum headway may not be achieved by the fifth queue position, as implied by the 1985 HCM procedure. Obviously, if the queue has not reached a minimum value by the fifth queue position, the minimum headway calculated using the HCM procedure would be biased toward a larger value than that ultimately achieved by the traffic queue.

This bias is further magnified when frequency of occurrence is considered. In this regard, the number of observed headways generally decreases with increasing queue position. As a result, the minimum headway calculated using the HCM procedure would be weighted toward the values observed in the lower queue positions. In effect, if headways for the lower queue positions are consistently larger than those of higher positions and if they are also observed with the greatest frequency, then the bias in the estimated average minimum headway will be even greater due to the unequal frequency of observations.

This bias will also affect the estimation of start-up lost time for two reasons. First, the minimum headway is needed to calculate start-up lost time (see Equation 2). A minimum headway biased toward a larger value will result in an estimated start-up lost time that is smaller than actually incurred. Second, the number of queue positions included in the sum (N in Equation 2) would need to include all of the positions that are incurring some added lost time because of start-up effects. If too few queue positions are included, the estimated start-up lost time would be biased toward a smaller value than actually incurred.

The magnitude of the potential bias resulting from the use of the 1985 HCM procedure is given in Table 3. The data in Column 1 represent an estimate of the minimum discharge headway with most of the bias by queue position (as magnified by unequal frequency) removed. This was accomplished by first averaging the headways for queue positions 13 and higher for the through movements. Positions 10 and higher were used for the left-turn movements because they tended to reach a minimum headway sooner than the through movements. These averages-by-queue-position were then averaged to yield the values given in Column 1. To add stability to the estimates, only queue positions having 20 or more observations were considered. Because of these restrictions, only 4 of the 5 through lanes and 6 of the 17 left-turn lanes had enough observations to calculate a minimum discharge headway by this procedure.

A comparison of Columns 1 and 2 in Table 3 (shown in Column 3) indicates that the 1985 HCM procedure always overestimates the minimum discharge headway. Moreover, the overestimation appears to be greater for the left-turn (0.13) than the through (0.06) movements. The error over all movements and sites ranges from 0.02 to 0.20 sec/veh.

In some cases, this bias may not be large enough to compromise the results of a capacity analysis; however, it does tend to cloud any statistical analysis of cause and effect by introducing added variability in the data set. As a result, the true effect of a treatment or factor (e.g., lane width, percent trucks, etc.) may be obscured by data from sites having different amounts of bias by queue position and a different frequency of observations at each position.

Factors Affecting Discharge Headway

Several precautions were taken to eliminate effects that might confound the analysis of discharge headway. To minimize differences in driver acceleration for the first few queue positions, the ANOVA tests only considered headways for the fifth and higher queued vehicles. In addition, the ANOVA tests included queue position as a blocking factor to preclude any bias that might be introduced by the different queue lengths found at each site. By "blocking" on queue position, all of the ANOVA comparisons are made on a queue-position-by-queue-position basis, thereby eliminating any bias by queue position. The data set used in the ANOVA analysis consisted of the individual passenger car headways recorded at each site for each movement studied, not the averages in Table 1.

One of the most interesting findings of the ANOVA tests is the significantly smaller headways of the SPUI left-turn movements compared with the SPUI through movements ($p = 0.001$). This trend is contrary to the relationship suggested by the 1985 HCM (*3*, Chapter 9). Possible reasons for the lower left-turn headways at SPUIs are (*a*) a heightened awareness of left-turn (relative to through) drivers, (*b*) the ability of left-turning drivers to see preceding drivers in the queue complete the turn and thus anticipate the correct turn path, and (*c*) the provision of lane markings through the interchange for the left-turn (but not the through) movements at SPUIs. The trend toward smaller left-turn headways was also observed by Poppe et al. (*7*) at two of the three SPUIs they studied.

The ANOVA tests also indicated that the through movement headways at the SPUIs are significantly larger than those at the AGIs ($p = 0.001$). An explanation for the larger through movement headways found at the SPUIs may be the extra caution exercised by drivers when entering the rather lengthy conflict area associated with the SPUI. Time of day was not found to have a significant effect on discharge headway.

The effect of traffic pressure was examined by considering both traffic volume per cycle per lane and queue length per cycle. Both measures were found to be significant; however, lane volume per cycle accounted for considerably more of the total sum of squared deviations from the mean ($SS_v = 26.9$, $p_v = 0.001$; $SS_q = 15.0$, $p_q = 0.001$). As a result, lane volume per cycle was determined to be the strongest measure of traffic pressure.

Increased traffic pressure was found to decrease headways at all queue positions. This trend is shown in Figure 2 for queue position five of the through movement at the Belcher

TABLE 3 Comparison of Different Methods of Estimating Minimum Discharge Headway

Movement	Site	Turn Radius, R (ft)	Traffic Pressure, v (vpcpl)	Minimum Discharge Headway, sec/veh				
				Average by Posn. (1)	5th to Last (2)	Diff. (2)-(1) (3)	Equs. 4 & 6 (4)	Diff. (4)-(1) (5)
Thru	SR 60		10.3[a]	2.06[b]	2.12[c]	0.06	2.00[d]	-0.06
	SR 686		10.6	1.95	1.97	0.02	1.99	0.04
	SR 694		9.7	2.00	2.13	0.13	2.00	0.00
	Belcher		14.8	1.82	1.84	0.02	1.73	-0.09
			Average:	1.96	2.02	**0.06**	1.93	**-0.03**
			Standard Deviation:	0.10	0.14	**0.07[e]**	0.13	**0.06[e]**
Left	SR 60	180	8.3	1.69	1.82	0.13	1.79	0.10
		180	8.4	1.76	1.87	0.11	1.79	0.03
	SR 686	260	13.3	1.70	1.78	0.08	1.70	0.00
		275	8.5	1.82	1.97	0.15	1.76	-0.06
	SR 694	280	6.7	1.85	1.97	0.12	1.78	-0.07
		230	9.1	1.66	1.86	0.20	1.76	0.10
			Average:	1.75	1.88	**0.13**	1.76	**0.01**
			Standard Deviation:	0.08	0.08	**0.14[e]**	0.03	**0.07[e]**

Notes:

a - Average traffic pressure for study period.

b - Average minimum discharge headway calculated by first averaging the observed headways for each queue position. Then, the average of these averages-by-queue-position was calculated and is shown above. For through movements, this procedure considered average headways for queue positions 13 and higher (H_{13}, H_{14}, H_{15}, etc.). For left-turn movements, queue positions 10 and higher were considered because of a tendency to reach a minimum value at lower queue positions. Only those queue positions having 20 or more observations were considered. The number of queue positions included in the averages shown ranges from 3 to 14.

c - Minimum discharge headway based on the 1985 HCM procedure (i.e., the average of all observed headways for the fifth through last queue positions).

d - Minimum discharge headway predicted by Equations 4 and 6.

e - Standard error, calculated as: $s = \sqrt{(\Sigma diff^2)/n}$.

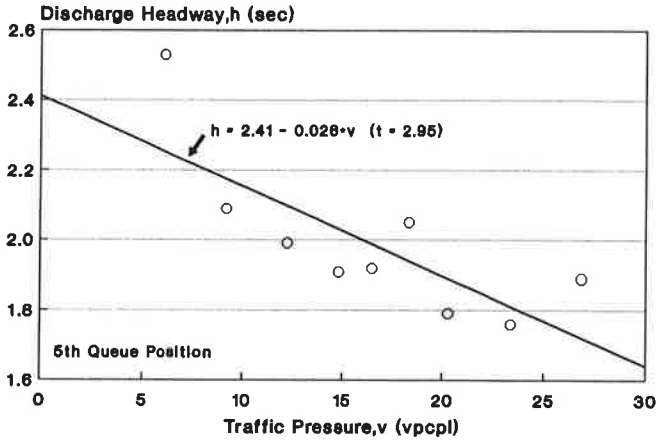


FIGURE 2 Discharge headway as a function of traffic pressure.

AGI. The data points in this figure represent the average of 15 observations each; however, the equation shown was determined from a regression analysis of the individual observations. As indicated by the *t*-statistic, the trend was highly significant ($p = 0.004$). This finding is consistent with that of Stokes et al. (*8*) and Lee and Chen (*9*).

Other factors considered in the ANOVA tests include lane width and dual-versus-single-lane left-turn operation. Lane widths ranged from 10 to 18 ft with a median of 12 ft. The wider lane widths were found on the single-lane off-ramp left-turn bays. On the basis of this analysis, it was found that neither lane width nor number of left-turn lanes has a significant effect on headway at the five study sites ($p_w = 0.110$ and $p_n = 0.303$, respectively).

Headway Model Development

As shown elsewhere (*2*), a linear speed-based acceleration model can be used to describe the dynamics of a starting traffic

queue. This model was extended to the discharge headway process, which led to the development of a discharge headway model that is sensitive to driver perception-reaction time, queue position, and vehicle speed.

Calibration of the discharge headway model was based on a least-squares regression of discharge headways averaged by site and queue position. Because the number of headways recorded varied widely among these factors, the headway data were averaged to remove the bias that an unequal sample size would have on model parameters. As a result, the statistics used to assess model fit to the data (i.e., standard deviation and R^2) do not reflect the total variability in individual driver headways. Rather, these statistics indicate the ability of the model to predict the average discharge headway by queue position.

The calibrated discharge headway model is

$$h_n = \tau' * N_1 + T' + \frac{d'}{V_{max}} + b_3 * \left(\frac{V_{sl(n)} - V_{sl(n-1)}}{A_{max}} \right) + b_4 * v + b_5 * \text{AGI} \qquad (3)$$

The model parameters are as follows:

τ' = regressed additional response time of the first queued driver (sec),
T' = regressed driver starting response time (sec),
d' = regressed distance between vehicles in a stopped queue (ft),
$V_{sl(n)}$ = stop line speed of the nth queued vehicle (fps),
V_{max} = common desired speed of queued traffic (fps), and
A_{max} = maximum acceleration (fpss).

The model variables are as follows:

h_n = headway of the nth queued vehicle (sec);
n = queue position, $n = 1, 2, 3, \ldots$;
v = traffic pressure (veh/cycle/lane);
N_1 = indicator variable (1 for first queue position; 0 for all others); and
AGI = indicator variable (1 for AGI, 0 for SPUI).

Regression results for through movements are as follows:

Parameter	Parameter Value	t-statistic
$\tau'(b_0)$	1.03	17.7
$T'(b_1)$	1.57	4.6
$d'(b_2)$	25.25	1.7
b_3	0.357	8.2
b_4	−0.0086	1.6
b_5	−0.23	4.3

Variable	Minimum Value	Maximum Value
n	1	18
v	0.0	16.8
h	1.6	3.8
Observations:	164	
Std. Deviation:	0.16	
R^2:	0.88	

Regression results for left-turn movements are as follows:

Parameter	Parameter Value	t-statistic
$\tau'(b_0)$	0.76	14.5
$T'(b_1)$	1.58	20.5
$d'(b_2)$	9.82	4.8
b_3	0.538	17.9
b_4	−0.0121	5.2
b_5	0.00	

Variable	Minimum Value	Maximum Value
n	1	24
v	0.0	18.3
h	1.3	4.0
Observations:	215	
Std. Deviation:	0.14	
R^2:	0.94	

In general, the parameter values for b_0, b_1, and b_2 are consistent with the definitions of the theoretical model parameters to which they correspond (i.e., τ, T, and d, respectively). Since these values do not, however, represent actual physical measurements, a prime symbol (′) has been added to each model parameter to denote that its value was established using regression analysis and that the relationship between this value and the theoretical definition may be distorted.

The statistical analysis revealed that traffic pressure, as measured by lane volume per cycle, was significant in reducing discharge headway. In using this component of the headway model, a one-to-one relationship between the duration of the volume average and the predicted average headway must be maintained. In other words, a lane volume representing a 1-hr average should be used in the regression model to predict discharge headways during the same 1-hr period.

In general, the significance of the calibrated parameter values combined with the theoretical basis of the discharge headway model suggests that the model adequately describes the headway process of queued vehicles. This ability is shown in Figure 3, which compares the calibrated model with the data for two SPUI left-turn movements.

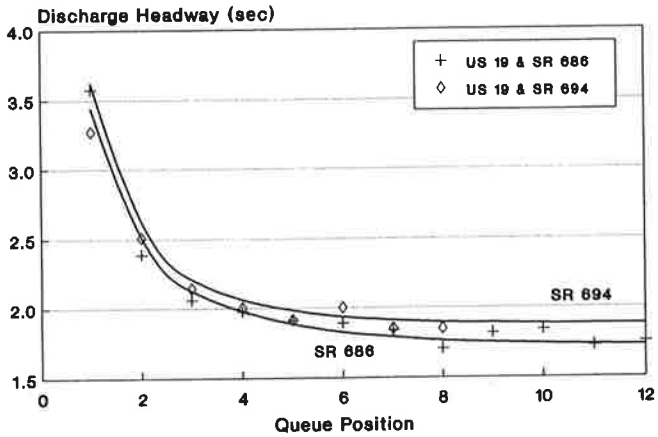


FIGURE 3 Discharge headway as a function of queue position.

Figure 3 indicates that the predicted discharge headway does reach a relatively constant, minimum value after the eighth or ninth queue position for these particular movements. This trend suggests that the procedure recommended by the 1985 HCM for estimating the minimum discharge headway, if applied to these movements, would include queue positions that have not reached a minimum headway. As a result, the HCM procedure would result in estimates that are biased toward values larger than the minimum headways ultimately achieved.

Minimum Discharge Headway Model

As discussed previously, the minimum discharge headway is the constant headway reached by the traffic queue. Examination of Equation 3 indicates that a constant headway is not reached until the traffic queue reaches its common desired speed (V_{max}), at which point the difference in stop line speeds (fourth term) equals zero. Furthermore, a study of the stop line speeds of the through movements indicated that V_{max} was relatively constant at 49.0 fps (*2*). As a result, Equation 3 can be simplified into the following model of minimum discharge headway for through movements:

$$H_{th} = 2.09 - 0.0086 * v_{th} - 0.23 * AGI \quad (4)$$

where

H_{th} = minimum discharge headway for through movements (sec/veh),
v_{th} = traffic pressure (veh/cycle/lane), and
AGI = 1 if the movement is at an AGI and 0 if it is at a SPUI.

Equation 4 implies that an AGI with a nominal through volume of 5.0 veh/cycle/lane should have an ideal minimum discharge headway of 1.81 sec/veh. This value suggests that the ideal minimum headway at these sites may be smaller than the 2.0 sec/veh recommended by the 1985 HCM (*3*, Chapter 9). Equation 4 also implies that the SPUI through movement headway is 0.23 sec/veh longer than that of the AGI. This trend is consistent with the findings from the ANOVA analysis discussed previously.

The relationship between turn radius and V_{max} for the left-turn movements studied indicated a statistically significant trend toward higher speeds on larger-radius turns (*2*). This relationship was quantified as

$$V_{max} = 8.85 * R^{0.245} \qquad R^2 = 0.81 \quad (5)$$

where R is the radius of curvature (ft).

Combining this relationship with Equation 3 and simplifying results in the following minimum discharge headway model for left-turn movements:

$$H_{lt} = 1.58 + \frac{1.11}{R^{0.245}} - 0.0121 * v_{lt} \quad (6)$$

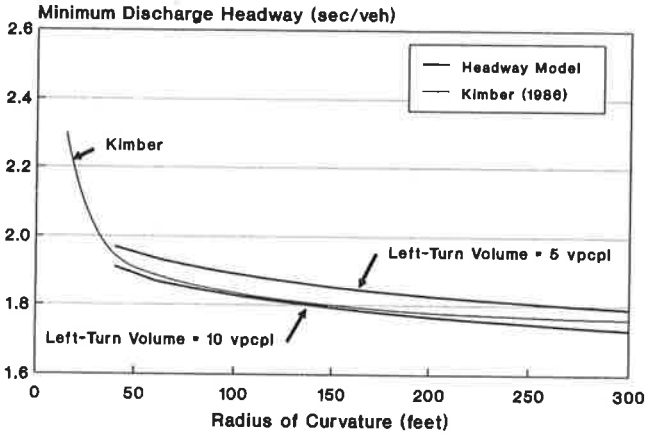


FIGURE 4 Minimum discharge headway of a left-turn movement as a function of curve radius.

where

H_{lt} = minimum discharge headway for a left-turn movement (sec/veh),
v_{lt} = traffic pressure (veh/cycle/lane), and
R = radius of curvature (ft).

The relationship between radius and the minimum discharge headway for left-turn movements predicted by Equation 6 is shown in Figure 4. Figure 4 also shows the relationship between radius and headway predicted by Equation 1 (with proper conversion from saturation flow rate). For purposes of comparison, the headway model from this research is shown with a traffic pressure of 5 and 10 vehicles per cycle per lane. In general, the agreement appears to be good for the range of radii studied (i.e., 60 to 280 ft).

The ability of Equations 4 and 6 to predict the minimum discharge headway is shown in Table 3. A comparison of the predicted values in Column 4 with the average values in Column 1 indicates that the model is a better predictor than the 1985 HCM procedure. This ability is quantified in terms of the average difference and the standard error given in Column 5. These statistics indicate that the model predicts minimum headways closer to those in Column 1 (model: −0.03 throughs and 0.01 lefts versus HCM: 0.06 throughs and 0.13 lefts) and with less standard error (model: 0.06 throughs and 0.07 lefts versus HCM: 0.07 throughs and 0.14 lefts). Moreover, the model does not appear to have any bias toward overestimation, as does the 1985 HCM procedure.

Start-Up Lost Time Model

The theoretical start-up lost time can be calculated from the calibrated discharge headway model (Equation 3) and Equation 2 as

$$K_{s(th)} = 1.03 + 0.357 * \frac{V_{max}}{A_{max}} \quad (7)$$

$$K_{s(lt)} = 0.76 + 0.538 * \frac{V_{max}}{A_{max}} \quad (8)$$

where

$K_{s(th)}$ = start-up lost time for a through movement (sec/phase),
$K_{s(lt)}$ = start-up lost time for a left-turn movement (sec/phase), and
A_{max} = maximum acceleration (equal to 6.63 fpss).

Driver acceleration data were also collected to validate the linear, speed-based acceleration model. Analysis of these data indicated that the acceleration profile of both the through and left-turn movements supported the use of a single value of A_{max} (*2*). This suggests that drivers initially accelerate in a similar manner, regardless of whether they are turning or traveling straight. This value of A_{max} was 6.63 fpss.

Equation 7 suggests that the start-up lost time for a through movement with V_{max} of 49 fps and A_{max} of 6.63 fpss is about 3.67 sec. This value is larger than the 2.0 sec suggested by the 1985 HCM (by about 80 percent) because it includes the lost time incurred by queue positions five and higher.

Combining Equation 5 with Equation 8 yields the following equation for estimating the start-up lost time for left-turn movements:

$$K_{s(lt)} = 0.76 + 0.718 * R^{0.245} \quad (9)$$

where $K_{s(lt)}$ is start-up lost time for a left-turn movement (sec/phase) and R is the radius of curvature (ft). These start-up lost time relationships are not based directly on the data in Table 2. They were developed using the minimum discharge headway relationship in Equation 3. As a result, they should be used in conjunction with Equations 4 and 6.

CONCLUSIONS AND RECOMMENDATIONS

This research found that traditional methods for estimating the average minimum discharge headway and start-up lost time may be biased toward values higher than ultimately achieved by the traffic queue. Moreover, the degree of bias varied widely among the movements and sites studied. As a result, initial attempts at a cause-and-effect analysis were clouded by a high degree of variability in the data.

In recognition of the aforementioned bias, alternative statistical analysis techniques and regression models were used to identify significant effects and to calibrate predictive models of minimum discharge headway and start-up lost time. On the basis of this research, it is recommended that statistical analyses of cause and effect in headway data use ANOVA techniques that account for unbalanced data (e.g., SAS System's general linear model) (*6*). It is also recommended that the ANOVA include queue position as a blocking factor.

Examination of the headway model components indicates that the minimum discharge headway may not always be reached by the fifth queue position. This suggests that traditional procedures for calculating the minimum discharge headway may not yield the value ultimately achieved by the traffic queue. It is recommended that minimum discharge headway be calculated by first averaging headway observations by queue position and then averaging these averages-by-queue-position. Only those queue positions that appear to have stabilized at a constant value should be included in the overall average. Alternatively, the regression modeling approach described in this paper could be used.

The headway model calibration suggests that through movement headways at SPUIs are larger than those at AGIs (0.23 sec/veh larger for the SPUIs and AGIs studied). The SPUI through movements were also found to have minimum headways larger than those of the left-turn movements. The latter trend is supported by the results of another study (*7*); however, further studies are needed to fully verify these findings because they are so contrary to conventional trends found at AGIs.

Left-turn movement headways were found to vary with turn path radii. The larger radii of the SPUI left-turn paths resulted in minimum headways that are about 0.12 sec shorter (based on Figure 4) than those for the AGI left-turn paths. A comparison of this trend in left-turn headway with the results of other research (*4*) suggests that Equations 6 and 9 can be extended to other SPUIs and AGIs.

Traffic pressure, as measured by traffic volume per cycle, had a statistically significant effect on discharge headway. An increase in traffic pressure resulted in a decrease in discharge headway. This trend has been noted in other studies (*8,9*); however, all of these studies (including this study) are based on a small number of sites. In recognition of the potential significance and magnitude of this effect, it is recommended that traffic pressure be more fully examined in any future studies of discharge headway.

ACKNOWLEDGMENTS

The author would like to recognize the agencies responsible for sponsoring the research project that provided the data for this study. In particular, the data were collected as part of a study titled Single Point Urban Interchange Design and Operations conducted by the Texas Transportation Institute and sponsored by the National Cooperative Highway Research Program, Washington, D.C. The author is particularly grateful to the faculty members at Texas A&M University who served on his committee: Carroll Messer (chairman), Daniel Fambro, Raymond Krammes, and Martin Wortman.

REFERENCES

1. C. J. Messer, J. A. Bonneson, S. D. Anderson, and W. F. McFarland. *NCHRP Report 345: Single-Point Urban Interchange Design and Operations Analysis*. TRB, National Research Council, Washington, D.C., 1992.
2. J. A. Bonneson. *Operational Characteristics of the Single-Point Urban Interchange*. Ph.D. dissertation. Texas A&M University, 1990.
3. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
4. R. M. Kimber, M. McDonald, and N. B. Hounsell. *The Prediction of Saturation Flows for Road Junctions Controlled by Traffic Signals*. TRRL Research Report RR67. Transport and Road Research Laboratory, Department of Transport, Berkshire, England, 1986.

5. *SAS/STAT User's Guide*. Release 6.03 Edition. SAS Institute Inc., Cary, N.C., 1988.
6. R. J. Freund, R. C. Littell, and P. C. Spector. *SAS Systems for Linear Models*. SAS Institute Inc., Cary, N.C., 1986, pp. 101–108.
7. M. J. Poppe, A. E. Radwan, and J. S. Matthias. Some Traffic Parameters for the Evaluation of the Single-Point Diamond Interchange. In *Transportation Research Record 1303*, TRB, National Research Council, Washington, D.C., 1991, pp. 113–124.
8. R. W. Stokes, C. J. Messer, and V. G. Stover. Saturation Flows of Exclusive Double Left-Turn Lanes. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986, pp. 86–95.
9. J. Lee and R. L. Chen. Entering Headway at Signalized Intersections in a Small Metropolitan Area. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986, pp. 117–126.

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

Potential Accuracy of a Planning Application for the HCM Signalized Intersection Operational Procedure

MARK R. VIRKLER AND CHIHNG-CHIR CHEN

The *Highway Capacity Manual* signalized intersection planning procedure uses limited data to identify overcapacity situations. However, the planning procedure lacks an indication for level of service. The signalized intersection operational procedure requires a large amount of data but identifies flow to capacity (v/c) ratios, delay, and level of service. A planning application of the operational procedure, using the same inputs as the present planning procedure, has been suggested. The application would require a large number of default values for inputs along with a method to develop a surrogate signal timing plan. The potential accuracy of such a planning application of the operational procedure was examined through applications to morning and evening peak-period data from 40 intersections in Missouri. The default values for adjustment factors performed well. Whereas the default values generally led to an underestimation of capacity, there was still a strong relationship for the v/c and level of service results derived from the default versus the actual adjustment factors. The surrogate signal timing algorithm performed adequately. There was a reasonably consistent relationship for the results generated by the surrogate signal timings compared with the results from the actual signal timings. A planning application of the operational procedure would be a valuable asset for planning and design analyses of intersections similar to those studied here. The application should encourage agencies to calibrate typical values for such variables as saturation flow rate, peak-hour factor, percent trucks, and pedestrian volumes. For consistency, the application should use a signal timing algorithm that at least approximates the best level of service to be expected from the intersection. The application's estimates of v/c, delay, and level of service would be valuable additions to the planning and design processes.

The *Highway Capacity Manual* (HCM) planning procedure for signalized intersections (*1*) has been criticized for lacking an indication for level of service. It has been suggested that the HCM operational procedure, which does predict level of service, could be modified to be used with only planning-level information. The purpose of this study was to examine how accurately a planning application of the operational procedure could predict the outcome of a more data-intensive operational analysis for a variety of signalized intersections.

BACKGROUND

The HCM uses three levels of analysis for traffic facilities: planning, design, and operational. Planning procedures use limited information at the earliest stages of planning to provide rough estimates of the number of lanes required. Design procedures more accurately estimate the needed number of lanes through the use of detailed data on expected traffic volumes and characteristics. Operational procedures are the most detailed and flexible of the analysis approaches. Known or projected traffic demands and characteristics are compared with known or projected highway characteristics to estimate the expected level of service.

The HCM contains a planning procedure and an operational procedure for the analysis of signalized intersections. The two procedures approach the analysis of signalized intersections in vastly different ways.

Operational Procedure for Signalized Intersections

Operational analysis requires detailed data on roughly 20 types of information relating to prevailing traffic, roadway, and signalization conditions. The procedure considers service flow rates on intersection approaches, the signalization plan, quality of signal progression, geometric design, and the resulting delay. Level of service is determined by the average delay.

As shown in Figure 1, the operational methodology can be applied to solve for a variety of variables.

1. Level of service can be determined from the details of traffic demand, geometrics, and signalization (Figure 1a).
2. Allowable service flow rates (allowable demand) can be determined from geometric and signalization conditions (also through the sequence in Figure 1a).
3. A reasonable signal timing (for an assumed phase plan) can be determined from information on flows and geometrics (Figure 1b).
4. The number and directional designations of lanes can be determined for a desired level of service and the details of flows and signalization. This is the design application (Figure 1c).

The operational procedure was developed to solve directly for level of service (the first application). This use also yields flow to capacity (v/c) ratios for each lane group, the delay for each lane group, the critical v/c for the intersection, and the average delay for the intersection. The other applications can require more than one pass through the procedure.

Department of Civil Engineering, University of Missouri-Columbia, Columbia, Mo. 65211.

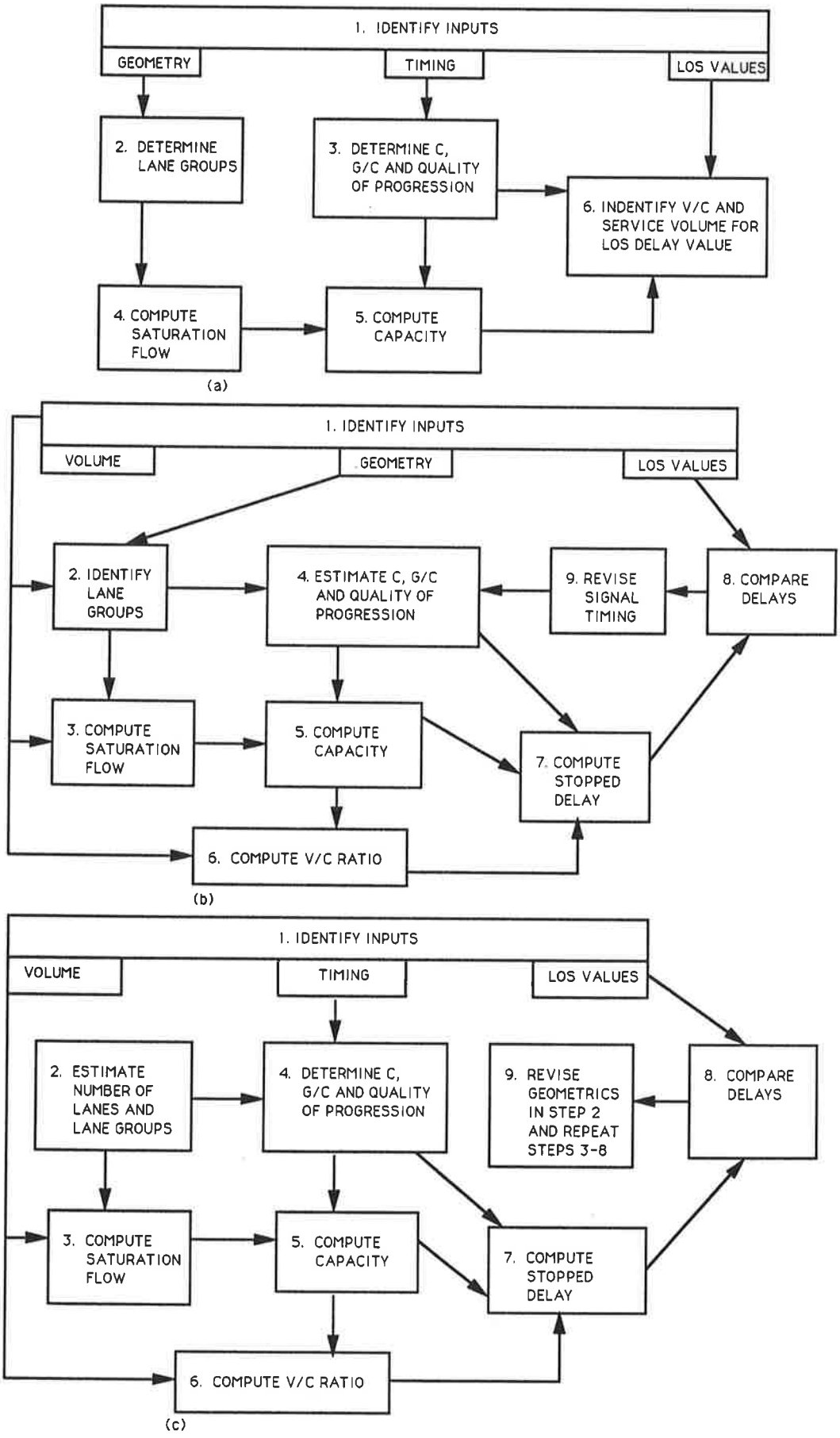


FIGURE 1 Alternative computation using operational analysis (*1*): (*a*), determining v/c ratios and service flow rates; (*b*), determining signal timing; and (*c*), determining number of lanes.

Planning Procedure for Signalized Intersections

The signalized intersection planning procedure is generally used when the detailed information required to estimate delay is not available. The planning procedure uses only the hourly traffic volumes and the number and directional designation of lanes. The method provides a determination that the intersection will be under, near, or over capacity. Since delay is not estimated, no level of service determination is made. Calculations are simple and are typically performed manually.

Highway Capacity Software

The operational procedure is generally performed on a microcomputer. The Highway Capacity Software (HCS) was developed for FHWA (*2*) and, as with a variety of similar software packages, can provide for data inputs and outputs within a matter of minutes if the analyst has the traffic, geometric, and signal timing data in hand. The HCS outputs v/c, delay, and level of service through a procedure similar to Figure 1a. Many of the HCS data inputs have default values that can be used if the actual values are unavailable. Some of the default values are recommended by the HCM. Some default values differ from or are in addition to the HCM recommendations. Default values are not available for traffic volumes, lane directional designation, and traffic signal timing plans.

It has been suggested that, since default values are available for most of the inputs, the addition of an algorithm to generate a reasonable signal timing plan would allow an HCS-type package to serve as a planning application of the operational procedure. The analyst would input only the traffic volumes, the number of lanes, and the directional designation of lanes. The software could then provide the expected level of service, critical v/c ratio, and associated measures. In other words, the same limited input data could be used to provide information that is more useful.

TABLE 1 Input Data for Operational Analysis

Type of Condition	Parameter	Study Default Value
Geometric	Area type	non-CBD *
	Number of lanes	_____
	Lane width	12 ft. *
	Approach grades	0% *
	Existence of exclusive LT or RT lanes	-----
	Parking allowed (yes or no)	no
Traffic	Volumes by movement	_____
	Peak hour factor	0.9
	Percent heavy vehicles	2%
	Conflicting pedestrian flow rate	50 peds./hr. (low)
	Number of local buses stopping	0 buses/hr.
	Number of parking maneuvers	0 man./hr.
	Quality of progression	Type 3
Signal	Cycle length (40 to 120 sec.)	_____
	Green times for each phase	_____
	Actuated vs. Pretimed	_____
	Pedestrian push button	no *
	Minimum pedestrian green	(not used)
	Phase plan	_____

* Indicates a default value used in this study which differs from the HCM default value or for which there is no HCM recommended default value.

RESEARCH APPROACH

Traffic data from signalized intersections in one large city and three smaller cities were provided by the Missouri Highway and Transportation Department (MHTD). The large-city data, for suburban St. Louis (population 450,000; 19 intersections) were from widely dispersed locations. The three smaller cities, all in central Missouri, were Columbia (population 68,000, nine intersections), Jefferson City (population 35,000, six intersections), and Sedalia (population 20,000, six intersections).

MHTD provided a recent turning movement count (actual 15-min turning movements as opposed to approach or demand volumes), a phasing timing sheet, and an intersection sketch for each location. The geometric and signal timing information were generally complete. Bus stops and on-street parking were generally not present.

The intent of the analysis was to determine how well the default data input values and a default signal timing algorithm could serve in a planning application of the operational procedure. Operational analysis calculations for v/c, delay, and level of service were performed to compare the use of default adjustment factors with the actual adjustment factors and the use of the signal timing algorithm with the actual timing.

Default Values Used in Planning Application of Operational Analysis

The default values used include some suggested by the HCM and some deemed appropriate after review of the intersections under study. The default values for geometric and traffic data are given in Table 1.

Signal Timing Rules

The default timings were based solely on peak-hour volumes and the number and designation of lanes. The HCM planning procedure was used to generate volume per lane. The rules used for the traffic signal timing are as follows:

1. If the left-turn volume on either direction of a street exceeded 100 vph, then left turns were protected. The ring concept (*3*) was then used for phasing.
2. The assumed saturation flow rate, including a consideration for a typical peak-hour factor, was 1,600 vphgpl.
3. Cycle length was found by setting the critical v/c ratio equal to 0.9, subject to the constraint that the cycle length must be between 40 and 120 sec. Green time was allocated in proportion to the volumes on the critical movements.
4. Streets with a single lane approach received a single phase.
5. Lost time equaled 3 sec per phase.

DATA

Tables 2 through 7 summarize the results of the operational analysis applications for each intersection. The results include the critical v/c for the intersection and the indicated level of service. The HCM delay equation is not recommended for a v/c ratio more than 1.2. If any lane group had a v/c ratio greater than 1.2, no intersection delay measure was calculated by the HCS.

All pretimed signals were analyzed in four ways:

1. Existing geometric and traffic conditions and existing signal timing (actual adjustment factors/actual timing or AAF/AT)—all adjustments for volume and saturation flow rate represent the appropriate HCM factors for the geometric and traffic demand conditions present. The existing signal timing plan was also used in the analysis.
2. Default geometric and traffic conditions and existing signal timing (default adjustment factors/actual timing of DAF/AT)—all adjustment factors were derived from Table 1.
3. Existing geometric and traffic conditions and signal timing from timing algorithm (actual adjustment factors/default timing or AAF/DT)—HCM adjustment factors were used for adjusting for geometrics and traffic demand characteristics. The algorithm described in the previous section was used to generate the signal timing plan.
4. Default geometric and traffic conditions and signal timing from timing algorithm (default adjustment factors/default timing or DAF/DT).

The average signal timings of the actuated signals were estimated by a method modeled after that recommended by Chapter 9, Appendix I of the HCM. The actual and default signal timings therefore did not differ for the actuated signals. The only actuated signal comparison is between actual geometric and traffic demand characteristics and default geometric and traffic demand characteristics (Tables 6 and 7).

ANALYSIS

Tables 8 through 11 compare the level of service derived for the actual geometric, traffic demand, and signal timing data with level of service derived from the other three approaches. Tables 8 and 11 indicate that the use of default geometric and traffic variables generally led to an equal or poorer level of service than that derived from the analysis of actual conditions. On the other hand, use of the traffic signal algorithm often led to a better level of service than that derived from the actual timing (see Table 9). The results were mixed when default adjustment factors and the default signal timings were both applied (see Table 10). In general, the smaller city intersections in mid-Missouri had indications of better performance with the use of all defaults. Intersections in suburban St. Louis had poorer levels of service with all defaults than with the actual geometric, traffic, and signalization conditions.

One unexpected result was the high number of instances when the critical v/c for the intersection exceeded unity (see Table 12). The traffic volumes used were actual throughput volumes rather than approach or demand volumes. If the operational procedure was completely correct, none of the intersections should have v/c ratios greater than 1.0. The likely reasons for this inconsistency include the following:

1. The actual saturation flow rates were higher than those estimated in the procedure. The default value of 1,800 pas-

TABLE 2 Critical v/c and LOS for St. Louis A.M. (SLAM) Peak

PRETIMED SIGNALS								
Intersection #	Critical v/c				Intersection LOS			
	AAF/AT	DAF/AT	AAF/DT	DAF/DT	AAF/AT	DAF/AT	AAF/DT	DAF/DT
SLAM01	0.422	0.431	0.517	0.548	B	B	B	B
SLAM04	1.225	1.439	1.179	1.295	*	*	*	*
SLAM05	1.198	1.458	0.803	0.844	*	*	*	*
SLAM06	0.819	1.017	0.856	1.089	C	*	D	*
SLAM07	0.915	0.924	0.858	0.870	*	*	C	*
SLAM09	0.852	0.878	0.902	0.938	B	B	B	*
SLAM10	1.225	1.257	1.151	1.191	*	*	F	*
SLAM11	1.126	1.202	1.207	1.091	*	*	D	*
SLAM12	1.182	1.399	1.149	1.375	*	*	*	*
SLAM13	0.681	0.681	0.970	1.153	*	*	D	*
SLAM14	1.165	1.148	1.192	1.183	*	*	*	*
SLAM15	0.481	0.596	0.652	0.740	*	*	B	C
SLAM16	0.726	0.832	0.927	1.052	*	*	C	E
SLAM17	0.505	0.590	0.596	0.696	D	*	B	B
SLAM18	0.498	0.627	0.541	0.679	B	B	B	B
SLAM19	0.940	0.943	0.967	0.971	*	*	E	E

AAF/AT: Actual adjustment factors/actual timing
DAF/AT: Default adjustment factors/actual timing
AAF/DT: Actual adjustment factors/default timing
DAF/DT: Default adjustment factors/default timing

TABLE 3 Critical v/c and LOS for St. Louis P.M. (SLPM) Peak

PRETIMED SIGNALS								
Intersection #	Critical v/c				Intersection LOS			
	AAF/AT	DAF/AT	AAF/DT	DAF/DT	AAF/AT	DAF/AT	AAF/DT	DAF/DT
SLPM01	0.550	0.602	0.818	0.904	C	C	C	C
SLPM04	-	-	-	-	-	-	-	-
SLPM05	1.309	1.588	0.954	1.104	*	*	*	*
SLPM06	0.784	1.031	1.010	1.298	C	*	D	*
SLPM07	0.694	0.759	0.802	0.869	C	C	B	*
SLPM09	0.902	0.940	1.020	1.058	*	*	D	*
SLPM10	1.688	1.812	1.444	1.450	*	*	*	*
SLPM11	1.309	1.452	1.271	1.380	*	*	*	*
SLPM12	1.007	0.998	0.976	1.050	*	*	*	*
SLPM13	0.650	0.669	0.878	0.865	*	*	E	*
SLPM14	-	-	-	-	-	-	-	-
SLPM15	0.730	0.855	1.112	1.249	*	*	F	*
SLPM16	0.800	0.927	0.785	1.110	*	*	D	F
SLPM17	0.392	0.456	0.413	0.479	C	*	D	*
SLPM18	0.473	0.583	0.501	0.614	B	B	B	B
SLPM19	0.936	1.022	1.213	1.325	*	*	*	*

AAF/AT: Actual adjustment factors/actual timing.
DAF/AT: Default adjustment factors/actual timing.
AAF/DT: Actual adjustment factors/default timing.
DAF/DT: Default adjustment factors/default timing.

TABLE 4 Critical v/c and LOS for Mid-Missouri A.M. (MMAM) Peak

PRETIMED SIGNALS								
Intersection #	Critical v/c				Intersection LOS			
	AAF/AT	DAF/AT	AAF/DT	DAF/DT	AAF/AT	DAF/AT	AAF/DT	DAF/DT
MMAM01	0.601	0.598	0.691	0.695	B	B	B	B
MMAM02	0.830	0.727	0.734	0.590	D	D	B	B
MMAM03	0.947	1.134	0.960	1.154	*	*	D	*
MMAM04	0.648	0.686	0.742	0.785	B	B	B	B
MMAM06	0.830	0.903	0.639	0.598	*	*	E	*
MMAM07	0.771	0.785	0.701	0.722	*	*	B	B
MMAM08	0.389	0.390	0.696	0.714	*	*	B	B
MMAM09	0.705	0.688	0.716	0.698	D	D	B	B
MMAM14	0.484	0.559	0.474	0.560	B	B	B	B
MMAM16	0.491	0.524	0.645	0.687	C	C	B	B
MMAM17	1.606	1.295	1.516	1.223	*	*	*	*
MMAM18	0.473	0.453	0.641	0.630	B	B	B	B
MMAM19	0.829	0.647	0.892	0.696	*	C	C	B
MMAM20	0.698	0.670	0.842	0.874	B	C	C	C
MMAM21	0.764	0.711	0.518	0.488	*	*	B	B

AAF/AT: Actual adjustment factors/actual timing.
DAF/AT: Default adjustment factors/actual timing.
AAF/DT: Actual adjustment factors/default timing.
DAF/DT: Default adjustment factors/default timing.

TABLE 5 Critical v/c and LOS for Mid-Missouri P.M. (MMPM) Peak

PRETIMED SIGNALS								
Intersection #	Critical v/c				Intersection LOS			
	AAF/AT	DAF/AT	AAF/DT	DAF/DT	AAF/AT	DAF/AT	AAF/DT	DAF/DT
MMPM01	1.448	1.361	0.915	1.039	*	*	B	D
MMPM02	0.814	0.892	0.778	0.801	E	E	C	D
MMPM03	0.778	1.017	0.827	1.083	B	D	C	*
MMPM04	0.787	0.785	0.967	0.965	C	C	C	*
MM0M06	0.871	0.881	0.879	0.838	C	*	E	*
MMPM07	0.539	0.598	0.725	0.796	*	*	B	B
MMPM08	0.438	0.539	0.785	0.850	*	*	B	B
MMPM09	0.675	0.802	0.748	0.882	C	*	A	B
MMPM14	0.601	0.723	0.710	0.831	C	C	B	B
MMPM16	0.475	0.575	0.564	0.687	C	C	B	B
MMPM17	1.630	1.553	1.594	1.518	*	*	*	*
MMPM18	0.594	0.606	0.883	1.616	B	B	C	C
MMPM19	1.197	0.778	1.067	0.934	*	*	E	C
MMPM20	0.724	0.950	0.876	1.085	C	*	C	*
MMPM21	0.757	0.828	0.708	0.763	*	*	E	E

AAF/AT: Actual adjustment factors/actual timing.
DAF/AT: Default adjustment factors/actual timing.
AAF/DT: Actual adjustment factors/default timing.
DAF/DT: Default adjustment factors/default timing.

TABLE 6 Critical v/c and LOS for St. Louis A.M. and P.M. Peaks

ACTUATED SIGNALS				
Intersection #	Critical v/c		Intersection LOS	
	AAF/AT	DAF/AT	AAF/AT	DAF/AT
SLAM02	1.076	1.133	F	*
SLAM03	0.929	0.996	D	E
SLAM08	0.507	0.532	A	A
SLPM02	1.239	1.411	*	*
SLPM03	1.072	1.110	E	F
SLPM08	0.804	0.879	B	*

AAF/AT: Actual adjustment factors/actual timing.
DAF/AT: Default adjustment factors/actual timing.

TABLE 7 Critical v/c and LOS for Mid-Missouri A.M. and P.M. Peaks

ACTUATED SIGNALS				
Intersection #	Critical v/c		Intersection LOS	
	AAF/AT	DAF/AT	AAF/AT	DAF/AT
MMAM05	1.005	0.835	*	*
MM1M10	0.630	0.693	B	B
MMAM11	0.452	0.467	B	B
MMAM12	0.642	0.639	B	B
MMAM13	0.450	0.450	B	B
MMAM15	0.418	0.426	A	A
MMPM05	0.697	0.757	B	C
MMPM10	0.917	0.999	D	D
MMPM11	1.040	1.080	C	D
MMPM12	0.642	0.602	B	B
MMPM13	0.525	0.601	B	B
MMPM15	0.514	0.615	B	B

AAF/AT: Actual adjustment factors/actual timing.
DAF/AT: Default adjustment factors/actual timing.

TABLE 8 Accuracy of Level of Service Prediction Using Default Adjustment Factors and Actual Timing

Actual Adjustment Factors/ Actual Timing	Prediction for St. Louis A.M.						
	A	B	C	D	E	F	*
A							
B		3					
C							1
D							1
E							
F							
*							11

Actual Adjustment Factors/ Actual Timing	Prediction for St. Louis P.M.						
	A	B	C	D	E	F	*
A							
B		1					
C			2				2
D							
E							
F							
*							9

Actual Adjustment Factors/ Actual Timing	Prediction for Mid Missouri A.M.						
	A	B	C	D	E	F	*
A							
B		4	1				
C			1				1
D				2			
E							
F							
*							6

Actual Adjustment Factors/ Actual Timing	Prediction for Mid-Missouri P.M.						
	A	B	C	D	E	F	*
A							
B		1		1			
C			3				3
D							
E					1		
F							
*							6

senger cars per hour of green per lane was used in all of the analyses. The HCM recommends that agencies calibrate saturation flow rates appropriate for the intersections within their jurisdictions.

2. Right-turns-on-red may have lessened the demand for green time in right-turn-only lanes or in shared lanes with right turns. The HCM does not include a procedure for dealing with right-turns-on-red. The HCS allows the analyst to subtract right-turn-on-red volumes from the traffic demand. No such adjustments were made in this study even though several lane groups with right turns were identified as critical.

Another unexpected result was the high number of situations where at least one lane group had a v/c greater than 1.2 (and hence no indication for level of service). In several cases realistic estimates of saturation flow rates or right-turns-on-red would probably have eliminated this problem. In some cases a left-turn lane group with low demand but an even lower capacity had a very high v/c ratio. Many of these intersections may in fact have been operating reasonably well.

Table 13 gives linear regression equations derived by predicting actual v/c ratios by each of the three approaches. If a set of predicted v/c ratios had been perfect, the regression equation would have a slope of one and an intercept of zero.

Default Adjustment Factors

Parts A and D of Table 13 indicate that the v/c ratios predicted by using default values for the HCM adjustment factors are closely related to the v/c ratios derived from the actual adjustment factors (the correlation was significant at the 1 percent level in all cases). In Part D each signal's timing was derived from the raw traffic data and the HCM Chapter 9, Appendix II method for actuated signals. For the data of Part A, the actual signal timing was used. For both Parts A and D, only the adjustment factor values differed.

In St. Louis the use of default adjustment factors led to an average 10.7 percent overestimation of v/c. In mid-Missouri the average overestimation of v/c was only 3.9 percent. The average peak-hour factor in St. Louis was 0.93 and in mid-Missouri was 0.87. If the average peak-hour factors for these two areas had been used, the St. Louis data would have an average v/c overestimation of 7.1 percent, and the mid-Missouri data would have an average v/c overestimation of 7.8 percent. These overestimations would be primarily due to lane width and pedestrian flows. In many instances lanes, particularly left-turn lanes, had less than the default 12-ft width. The pedestrian flows were generally far below the default value of 50 pedestrians per hour using each crosswalk.

TABLE 9 Accuracy of Level of Service Prediction Using Actual Adjustment Factors and Default Timing

Actual Adjustment Factors/ Actual Timing	Prediction for St. Louis A.M.						
	A	B	C	D	E	F	*
A							
B		3					
C				1			
D		1					
E							
F							
*		1	2	2	1	1	4

Actual Adjustment Factors/ Actual Timing	Prediction for St. Louis P.M.						
	A	B	C	D	E	F	*
A							
B		1					
C		1	1	2			
D							
E							
F							
*				2	1	1	5

Actual Adjustment Factors/ Actual Timing	Prediction for Mid Missouri A.M.						
	A	B	C	D	E	F	*
A							
B		4	1				
C		1					
D		2					
E							
F							
*		3	1	1	1		1

Actual Adjustment Factors/ Actual Timing	Prediction for Mid-Missouri P.M.						
	A	B	C	D	E	F	*
A							
B			2				
C	1	2	2		1		
D							
E			1				
F							
*		3			2		1

Signal Timing Algorithm

Part B of Table 13 reflects use of the signal timing algorithm to predict the v/c derived from the actual signal timing. The actual adjustment factors were used in all cases. Whereas all correlations were significant at the 0.01 level, the correlation coefficients were not as high as those of Part A of Table 13.

In general the signal timing algorithm predicted the actual v/c well. However, a noticeable number of predictions differed significantly from the actual v/c, both through underestimation and overestimation of the v/c.

An algorithm to minimize the critical v/c for the intersection might have been used in place of the algorithm used in this study. One would assume that in that case the predicted v/c would never be greater than the actual v/c.

For planning purposes there is an obvious advantage to using an algorithm that would accurately predict the signal timing used in the field. One can envision that, if this had been the case, the r^2 for Part B of Table 13 would be closer to 1.

Default Adjustment Factors and Signal Timing Algorithm

Part C of Table 13 shows how well the use of both the default adjustment factors and the default signal timing predicted the actual v/c. Three of the correlations were significant at the 0.01 level and the fourth, mid-Missouri p.m. peak data, was significant at the 0.05 level.

For the St. Louis data the predicted v/c ratios were generally too high. For the mid-Missouri a.m. data the regression equation had a slope close to one and passed near the origin. However, the spread of the data from the curve was fairly large. For the mid-Missouri p.m. data, the regression curve differed markedly from a slope of one. Whereas the judicious removal of one or two data points could make the slope close to one, the remaining data points would still be far from a perfect fit to the ideal relationship.

RECOMMENDATIONS

Default Values

The worth of the default values used in a planning application of the operational procedure can be measured by how well the default values represent the actual values. When the actual signal timing plan was used, the v/c ratios derived from the default adjustment factors had a high correlation with the v/c ratios derived for the actual conditions. However, the predicted v/c ratios derived from default adjustment factors were generally too high. Similarly, the predicted level of service was often poorer than that for actual conditions.

TABLE 10 Accuracy of Level of Service Prediction Using Default Adjustment Factors and Default Timing

Actual Adjustment Factors/ Actual Timing	Prediction for St. Louis A.M.						
	A	B	C	D	E	F	*
A							
B		2					1
C							1
D		1					
E							
F							
*			1		2		8

Actual Adjustment Factors/ Actual Timing	Prediction for St. Louis P.M.						
	A	B	C	D	E	F	*
A							
B		1					
C			1				3
D							
E							
F							
*						1	8

Actual Adjustment Factors/ Actual Timing	Prediction for Mid Missouri A.M.						
	A	B	C	D	E	F	*
A							
B		4	1				
C		1					
D		2					
E							
F							
*		4					3

Actual Adjustment Factors/ Actual Timing	Prediction for Mid-Missouri P.M.						
	A	B	C	D	E	F	*
A							
B			1				1
C		3					3
D							
E				1			
F							
*		2	1	1	1		1

TABLE 11 Accuracy of Level of Service Prediction Using Default Adjustment Factors for Actuated Signals

Actual Adjustment Factors/ Actual Timing	Prediction for St. Louis AM & PM						
	A	B	C	D	E	F	*
A	1						
B							1
C							
D					1		
E							
F					1		1
*							1

Actual Adjustment Factors/ Actual Timing	Prediction for Mid Missouri AM & PM						
	A	B	C	D	E	F	*
A	1						
B		7	1				
C				1			
D				1			
E							
F							
*							1

TABLE 12 Categories of Intersection Critical v/c Ratios Resulting from Use of Actual Timing and Adjustment Factors

	Number of Intersections with Critical v/c		
	v/c < 1	1 < v/c < 1.2	v/c > 1.2
St. Louis	23	8	5
Mid-Missouri	36	3	3

TABLE 13 Regression Equations Derived from Using Predicted v/c Ratios To Estimate Actual v/c Ratios

How "x" Was Derived	Location[a]	Regression Equation[b]	r^2
A. Using Default Ajustment Factors with Actual Signal Timing	SL a.m.	y= 0.040 + 0.864x	0.929
	SL p.m.	y= 0.000 + 0.893x	0.964
	MM a.m.	y=-0.057 + 1.108x	0.848
	MM p.m.	y=-0.120 + 1.096x	0.802
B. Using Actual Adjustment Factors with Default Signal Timing	SL a.m.	y=-0.135 + 1.129x	0.741
	SL p.m.	y=-0.126 + 1.060x	0.677
	MM a.m	y=-0.040 + 1.023x	0.744
	MM p.m.	y=-0.007 + 0.916x	0.635
C. Using Default Adjustment Factors with Default Signal Timing	SL a.m.	y=-0.057 + 0.946x	0.585
	SL p.m.	y=-0.150 + 0.971x	0.594
	MM a.m.	y= 0.024 + 0.963x	0.463
	MM p.m.	y= 0.273 + 0.561x	0.186
D. Using Default Adjustment Factors for Actuated Signals	SL	y= 0.061 + 0.868x	0.982
	MM	y=-0.010 + 0.986x	0.888

[a] SL = St. Louis
MM = Mid-Missouri

[b] x = predicted v/c for intersection
y = v/c for intersection derived from adjustment factors and actual signal timing.

If an area average peak-hour factor had been used, the predicted v/c ratios would have averaged about 7 to 8 percent too high for both locations. Almost all of this average difference was due to lane width and pedestrian volume differences. A traffic organization will generally have data for or a reasonable estimate of existing lane width, percent trucks, and pedestrian flows. A planning application should encourage agencies to use their own default values for these variables. Similarly, an agency should be encouraged to develop appropriate estimates for a.m. and p.m. peak hour factors.

It is likely that an agency will often not have accurate information on approach grades, number of local buses stopping, and number of parking maneuvers. The present HCM default values should be used in those cases.

Signal Timing Plan

The signal timing algorithm performed adequately for the purposes of this study by generating reasonable timing plans. The algorithm was fairly accurate in predicting the v/c and level of service derived from the actual signal timing. The algorithm would be viewed by many as too simplistic to serve as a default signal timing algorithm for a planning application of the operational procedure.

Two alternative performance measures for a signal timing algorithm are apparent. One measure would be how well the algorithm predicts the signal timing plan that would be used at the signal. The other would be how closely the algorithm comes to optimizing some objective.

MHTD allows a wide latitude to the individual responsible for developing the signal timing plan. The guidelines are such that two individuals could easily develop significantly different, yet appropriate, plans for the same intersection. Further, it is the author's understanding that many signals are retimed in the field through observation of traffic during peak periods. It is doubtful that any algorithm would consistently predict both the phase plan and the green times used at MHTD intersections. Since various traffic agencies might use a variety of means to generate signal plans, a single algorithm would probably often be unsuccessful in predicting signal timing plans for a wide variety of agencies.

Many computerized algorithms exist for timing traffic signals. It is likely that an algorithm to minimize v/c or delay

could be incorporated into a planning application of the operational procedure. Since level of service is based on delay, a planning application of the operational procedure should include a signal timing procedure that either minimizes delay or comes close to that optimal solution. Such a procedure might not be the best approach for matching the actual level of service at signalized intersections. However, it would be consistent with the philosophy of the HCM that the most important measure of signalized intersection operation is delay. It also would encourage traffic agencies to view the purpose of signal timing to be the minimization of delay to the motorist.

Design

A planning application of the operational procedure is feasible. The same software could easily be used as an aid in the design of signalized intersections. At the design stage many of the input data would be available. With a reasonable signal timing algorithm incorporated into the software, an iterative design approach would be easily accomplished. A designer could examine a wide variety of alternative lane arrangements for v/c and level of service in much less than 1 hr.

Other Considerations

To gain wide acceptance, a planning application of the operational procedure must yield realistic results. The operational procedure indicated that the output volumes of many of the intersections were above the theoretical capacity. The reasons identified for these inconsistent results were higher-than-expected saturation flow rates and the lack of a method for dealing with right-turns-on-red. If these problems are not rectified, it is likely that a planning application will not receive as wide a use as possible.

A planning application should encourage agencies to calibrate saturation flow rates for their own intersections. It is likely that significant increases in accuracy would result.

An optional procedure to predict right-turns-on-red could also be beneficial. The procedure might be made available to the user when a lane group containing right turns has been identified as a critical lane group.

One difficulty in the availability of an HCM-type software package with a signal timing algorithm is that the signal timing algorithm might be viewed by some as the accepted way to time a traffic signal. The purpose of the HCM is to provide a means to measure the performance of facilities rather than to describe how traffic should be controlled. Care will be required if the planning application is not to be viewed as a guide for traffic signal timing.

CONCLUSIONS

For intersections similar to those of this study, a planning application of the HCM signalized intersection operational procedure could be practical and reasonably accurate. A planning application requires reasonable default values and a means to estimate an appropriate signal timing. Such an application can provide very accurate estimates of an intersection's critical v/c ratio and a likely estimate of achievable level of service. The only intersection-specific data required would be peak-hour volumes and lane usage.

A planning application should have the following characteristics:

1. The application should encourage an agency to develop its own appropriate estimates for peak-hour factor, percent trucks, and pedestrian volumes.
2. The application should encourage an agency to calibrate accurate estimates of ideal saturation flow rates for the intersections within its jurisdiction.
3. The application should use a signal timing algorithm that at least approximates the best level of service to be expected at the intersection.
4. Since some agencies will have more data readily available than will others, the application should allow the analyst to input site-specific values in place of default values.
5. If the application is also to serve as a design procedure, the application should provide a simple means to examine the results of changes in input data.
6. The level of accuracy for v/c and level of service predictions should be made clear to the user.
7. The signal timing algorithm should be presented as a representation of a reasonable signal timing rather than as a suggested signal timing.
8. A means to estimate the likely number of right-turns-on-red turning movements should be developed and considered for inclusion as an optional calculation for both the operational procedure and a planning application of the operational procedure.

A planning application with the above characteristics could be a useful tool for traffic engineers and planners. The application's predictions for v/c and level of service would add useful information to the planning process.

ACKNOWLEDGMENTS

The authors wish to thank Tom Dollus of MHTD for providing the data used in this study.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. *Highway Capacity Software, Release 1.50.* McTrans Center for Microcomputers in Transportation, University of Florida, Gainesville, 1986.
3. W. R. McShane and R. P. Roess. *Traffic Engineering.* Prentice Hall, Englewood Cliffs, N.J., 1990.

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

Implementing Travel Forecasting with Traffic Operational Strategies

ALAN J. HOROWITZ

An "adaptive" travel forecasting model that has the ability to account for the ways in which the traffic system operates is described. Within the model, trip distribution, mode split, and traffic assignment are all sensitive to delays at intersections as well as delays along uncontrolled road segments. To the extent possible, delay relationships were adapted from the 1985 *Highway Capacity Manual* (HCM). Separate relationships were included for all-way stop controlled, priority, and signalized intersections. The model was implemented as a specially modified version of the Quick Response System II (QRS II) software. The HCM signalized intersection procedures were difficult to incorporate because they result in delay/volume relationships that are nonmonotonic and discontinuous. Even with better-behaved delay relationships, it is unlikely that a unique solution could be obtained. Operational strategies can be either automatically calculated or specified by users, depending on their nature. The inclusion of operation strategies in the forecast does not greatly increase computation time, data requirements, or the needed level of user expertise. A form of elastic-demand incremental assignment could find at least one user-optimal equilibrium solution. An adaptive model can reduce dependence on base-year calibration for incorporating operational effects.

Long-term transportation plans are traditionally prepared by evaluating several fixed alternatives against expected future conditions. Sometimes a travel forecast reveals a major oversight in assumptions about the way an alternative is operating, in which case the alternative might be reworked and reevaluated. The determination of when an oversight in operation has occurred is entirely judgmental; many inconsistencies between a travel forecast and an alternative are routinely tolerated.

It is presently feasible to build travel forecasting models that can automatically account for the way facilities are operated. Conceivably, such models could set signal timing, remove on-street parking, determine signal coordination in a corridor, or choose traffic control devices for intersections. The model could make a large number of short-term, microscale design decisions that could not have been made without knowledge of future traffic volumes.

For example, planners normally expect travel forecasts to reflect delays at intersections. Intersection delays depend on signal timing, and, operationally, signal timing depends on traffic volumes. Consequently, a forecasting model should be capable of calculating important aspects of signal timing (number of phases, length of phases, and cycle length) for every intersection in the network as traffic is being assigned (*1–4*). In essence, the travel forecast can be adaptive in the same sense that an actuated signal control system is adaptive. Unless the model can make consistently good assumptions about signal timing, the implied capacities of arterials will be wrong and the resulting forecast will be wrong. This type of forecasting will be referred to here as "adaptive travel forecasting."

Center for Urban Transportation Studies and Civil Engineering and Mechanics, University of Wisconsin-Milwaukee, Room E387, EMS Building, P.O. Box 784, Milwaukee, Wis. 53201.

If one were to try to create a traffic-optimizing travel forecasting model, it would look something like a hybrid between a traditional urban transportation planning (UTP) model and an operations-level traffic model (such as SOAP, PASSER-II, or TRANSYT). Such a model would be very complex, requiring considerably more data, computer resources, and user expertise than either of its two constituents. A notable example of this type of model is Continuous Traffic Assignment Model (CONTRAM) from the Transport and Road Research Laboratory (*5*). The essential differences between a CONTRAM-like model and an adaptive travel forecasting model are discussed later in this paper.

Adaptive travel forecasting is a special case of the network design problem (*6,7*), which attempts to find societally optimized networks, perhaps involving new facilities as well as operational strategies. Unlike the network design problem, adaptive travel forecasting seeks to be entirely predictive of the impacts of specific alternatives. That is, the model attempts to forecast what will be, not what should be. Clearly, the model must deal with short-term operational strategies to arrive at reasonable link volumes, but these strategies have little value by themselves. The network design problem is inherently more demanding than adaptive travel forecasting.

This paper describes initial experiences with an adaptive travel forecasting model, which has the ability to modify traffic controls at numerous isolated intersections and gives users the ability to make other operational adjustments. The focus is on applications rather than theoretical issues. How much adaptation can reasonably be included? Can the model be made sufficiently consistent with existing methods of travel forecasting, traffic theory, the *Highway Capacity Manual* (HCM) (*8*), and traffic engineering practice? Can an equilibrium solution be found? Are data and computer requirements reasonable? Conversely, does adaptation impose limitations on the size of networks? Does adaptation interfere with elastic-demand assignments? Does adaptation make forecasts more difficult to interpret? Each of these questions will be addressed in the following sections.

ADAPTIVE TRAVEL FORECASTING MODEL

Adaptation was added to an existing travel forecasting package, Quick Response System II (QRS II). QRS II is typical

of the UTP models currently used by planning agencies and consultants in the United States, so it provides a fair test of the difficulties involved in implementing adaptation. Extensive revisions to the source code were required to establish signal timings, compute delays, find incrementally averaged trip tables, account for all turning movements, and identify the conflicting and opposing traffic for every approach.

Because of uncertainties of the effects of adaptation on forecasts, only phasing and signal timing at signalized intersections were made automatic. Cycle length, quality of progression, and the placement of stop signs were kept as manual (but explicit) adjustments to the network.

An adaptive model might have to simultaneously handle thousands of traffic-controlled intersections, so an early decision was made to hold the amount of data required for any given intersection to a bare minimum. To calculate intersection delay the model was designed to only require information about an approach's lane geometry, the saturation flow rate of its through lanes, and the form of traffic control. Additional data, such as link speed and street continuity, are already required by QRS II for the purposes of travel forecasting.

Delay

Separate delay procedures were implemented for signalized intersections, all-way stop controlled (AWSC) intersections, and priority intersections (one-way and two-way stops). To the extent possible, consistency with the 1985 HCM was maintained. The choice of the HCM procedures was made on the basis of their wide adoption by planners and traffic engineers and can be considered arbitrary from the standpoint of traffic flow theory. It is possible that delay relations from other countries could perform better in a travel forecasting application.

At this writing the HCM provides insufficient treatment of AWSC intersections. So instead, delay at AWSC intersections was calculated by an enhanced form of Richardson's M/G/1 queuing model (*9*). Additional terms were provided to handle delays from turning and from coordination between drivers on subject and opposing approaches. The M/G/1 model was successfully calibrated to data provided by Kyte (*10*), and it produces results that are consistent with the recently released interim procedures for AWSC capacity (*11*).

The HCM's procedure for priority intersections omits calculation of delay, does not account for equilibrium in lane utilization for multilane approaches, and does not provide capacities when there are large conflicting volumes. These problems were remedied. Delay was calculated as if stops were a random, single-server queue as suggested by the Swedish Highway Capacity Manual (*12*). More complete capacity relations were obtained from Baass (*13*).

Total approach delay for signalized intersections is calculated consistently with the HCM, except as follows:

1. As an expedient, delay is not separately calculated for exclusive right lanes. Rather, sufficient capacity to handle just right-turning traffic is added to the capacity of the TR or LTR group before application of the delay formula.
2. The possible presence of pedestrians is ignored in order to reduce data requirements.
3. Acceleration delay is calculated from link speed and an estimate of the number of stopping vehicles.
4. To avoid a serious discontinuity in the delay function, shared left lanes were allowed to act as exclusive left lanes only when the model determines that a protected left phase is required.
5. The HCM's stopped delay formula was slightly modified for volume-to-capacity ratios (X) greater than 1.0 to eliminate the possibility of infinite or negative delay values in the uniform delay term when a green time constitutes nearly the full cycle.

Total approach delay is found by taking a volume-weighted average of delay across all lane groups and all phases. A separate delay value was not calculated for left turns, because the HCM procedure cannot provide this value in all cases. The desirability of using separate left turn delays when provided by the procedure was not investigated in this study. Users retain the ability to add a left-turn penalty at individual approaches, if desired.

Phasing and Timing

Green times, saturation flow rates, and delays are intrinsically linked. As noted previously, the calculation of delay requires knowledge of lane-group capacity as given by saturation flow rates and green times. Green times depend on saturation flow rates as parts of flow ratios. Of course, the saturation flow rate for a lane group depends on left-turn volumes and the amount of opposing traffic.

Consequently, it becomes necessary to simultaneously solve for green times and saturation flow rates. Once these have been established, delay can be ascertained. This calculation must be performed for each of the many signalized intersections, for each hour in the analysis period, and at each traffic assignment iteration. Recognizing that the amount of calculation could be prohibitively large, it is necessary to place limits on the range of signalization strategies available for any given intersection.

These rules have been adopted:

1. Green time for a TR or LTR phase is allocated in proportion to the critical flow ratio for the phase.
2. To avoid very small green times, a minimum flow ratio can be established by the user for the purpose of signal timing.
3. A protected left phase is provided only if there is insufficient left-turn capacity during an LTR phase, considering both sneakers and gaps in opposing traffic.
4. If protection is required, the phasing is always equivalent to "dual leading lefts with overlap."

It is recognized that these rules do not produce the best signal timing, but they at least find an acceptable timing in accordance with traffic engineering practice. The issue of whether an adaptive model should find optimal, rather than conventional, signalization remains unresolved.

The concept of link capacity within UTP models is seriously weakened in an adaptive travel forecasting model. It is only known that traffic volumes should not exceed the saturation flow rate, less any approach capacity lost during phase changes. Otherwise, links compete for slices of time at intersections.

CREATING AND INTERPRETING AN ADAPTIVE TRAVEL FORECAST

Adaptive Model Operation

Figure 1 shows one possible way of operating an adaptive model. There are two loops. The inner loop automatically adjusts signal timing and incorporates congestion effects. The outer loop provides the user with opportunities to modify traffic control devices in accordance with traffic engineering practice.

This method of operating the model will produce an elastic-demand assignment. That is, both the distribution of trips throughout the urban area and the level of transit ridership will reflect vehicular flow conditions on the highways.

The allocation of traffic engineering principles to the two loops is somewhat arbitrary. As more is learned about adaptive travel forecasting, principles can be transferred from the outer loop to the inner loop. For example, textbooks provide simple rules for determining optimal cycle length at isolated intersections; these rules could be moved to the inner loop provided we also have some way to determine which intersections are truly isolated.

This iterative procedure raises the serious methodological issue of whether it is possible to obtain an equilibrium solution in accordance with Wardrop's first principle—a user-optimal assignment—in a reasonable amount of time on a very large network. (Networks in QRS II could be as big as 585 zones and 4,500 links on a microcomputer; other packages permit nonadaptive networks many times this size). It will be shown later in this paper that it is at least sometimes possible to obtain an equilibrium solution with large-network methods and that it is always possible to determine whether any given solution is an equilibrium one.

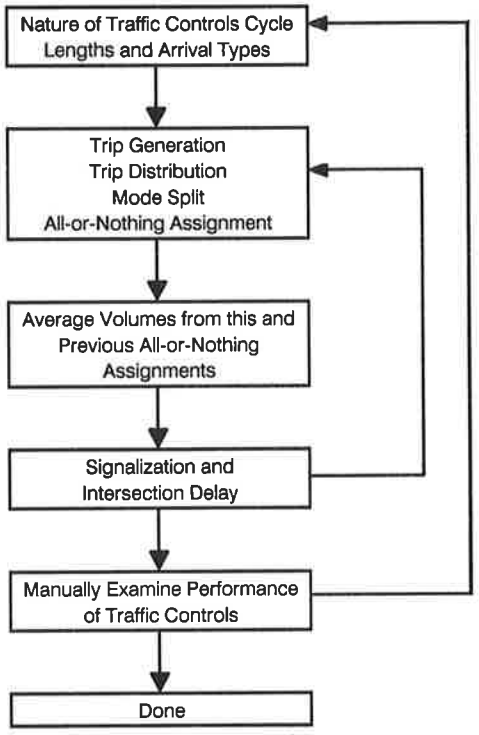


FIGURE 1 Operating an adaptive travel forecasting model.

Obtaining an Equilibrium Solution

A promising heuristic for obtaining an equilibrium solution involves an averaging of traffic volumes from many all-or-nothing assignments. Such an averaging step is fundamental to the most widely cited equilibrium assignment techniques: Frank-Wolfe decomposition for fixed-demand assignments (*14*), Evans's algorithm for elastic-demand assignments (*15*), and convergent incremental assignment (*16*).

Nonlinear optimization methods for large networks, such as Frank-Wolfe decomposition or Evans's algorithm, cannot be applied for two reasons. First, the adaptive model lacks a closed-form and well-behaved delay/volume function. Second, delay at any given approach is a function of all possible movements at the intersection. The second problem by itself could possibly be overcome by assignment methods designed to handle "asymmetric" networks (*17*).

Of the three aforementioned techniques, only incremental assignment can be implemented with adaptive travel forecasting, as it is now understood. In effect, incremental assignment creates a weighted average of many all-or-nothing assignments. The weights are predetermined; they do not depend on knowledge of the delay/volume function. Incremental assignment converges to a user-optimal equilibrium solution for fixed-demand assignments (*16*), runs only slightly slower than Frank-Wolfe decomposition (*18*), and works well on a broad range of elastic-demand problems (*19,20*). Incremental assignment has already been tested in the United Kingdom on networks with traffic controls by the authors of JAM as reported by Lewis and McNeil (*21*). Since there does not yet exist theory to suggest that incremental assignment works properly on adaptive networks, its usefulness must be established empirically.

A solution to an adaptive network may not be unique, even if it is an equilibrium solution. Because the model reallocates limited resources (such as intersection capacity or favorable coordination) across facilities, it is entirely possible to have many good solutions (*1,22*).

Consider the network of Figure 2. It consists of a trip origin, a trip destination, a single signalized intersection, and four

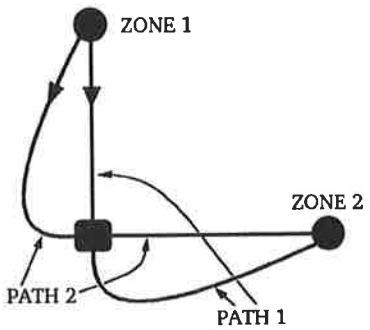


FIGURE 2 Two-zone, single-intersection network.

two-way links with capacities only at the intersection approaches. Turns at this intersection are fully restricted, so there are only two paths from the origin to the destination. Traffic is uncongested. Both paths have identical characteristics, but they compete for green time at the intersection. There are exactly three assignment solutions that could reasonably satisfy Wardrop's first principle:

A. All vehicles use Path 1,
B. All vehicles use Path 2, and
C. Both paths are used equally.

Further assume that the intersection is operating below capacity and that no minimum is established for the length of green phases. Solutions A and B would cause nearly the full cycle to be allocated to one path or another, thereby minimizing travel time for all vehicles. Solution C allocates green time equally to both streets, and it is the solution that would have been obtained in a nonadaptive network with a monotonically increasing delay/volume function. A quick inspection of the uniform delay term of the HCM delay formula reveals that Solution C is by far the least desirable from the standpoint of user cost.

It is possible to conclude from this example that equilibrium solutions to an adaptive network are not unique, and they would be likely to differ from those of a nonadaptive network. Furthermore, this example raises some tough methodological questions, which cannot be definitively answered here. How can we deal with multiple solutions to the same problem? Do we care to find more than one solution? Is there a tendency in adaptive travel forecasts toward all-or-nothing assignments?

Comparison with CONTRAM-Like Models

Since many traffic engineers are familiar with CONTRAM, it provides a good basis for comparing the current work. CONTRAM and QRS II are approaching the same methodological position from opposite directions but remain some distance apart. CONTRAM is fundamentally a traffic-optimization model that permits path choice; QRS II is fundamentally a travel demand/assignment model that contains explicit traffic flow relations. The two models have similar philosophical underpinnings but differ considerably in the types of problems they can address. CONTRAM is geared to small networks with fixed demands; QRS II is geared to large networks with elastic demands. CONTRAM's primary outputs are optimized traffic controls and indicators of the performance of the traffic system; QRS II's primary output is assigned traffic volumes. CONTRAM will yield results that are of little interest to those doing medium- or long-term travel forecasts, such as queue lengths and optimized green times. Interestingly, CONTRAM is still well ahead of conventional travel forecasting models by implementing a form of dynamic traffic assignment, something that planners are just now recognizing as important. However, CONTRAM's assignment algorithm cannot ensure that assigned volumes represent equilibrium conditions.

DELAY/VOLUME RELATIONSHIPS AT ADAPTIVE SIGNALIZED INTERSECTIONS

As indicated previously, delay must be found by simultaneously solving for saturation flow rates and green times. For this research the solution was found by the method of successive approximations.

Step 0. Determine the need for a protected left phase at each approach, so that the number of phases is known for the remaining steps. A by-product of this step is an initial estimate of saturation flow rates.
Step 1. Estimate green times.
Step 2. Estimate saturation flow rates according to the HCM procedure.
Step 3. Check for convergence. If not converged, go to Step 1.
Step 4. Calculate stopped and acceleration delay for each phase and lane group. Find the volume-weighted average for each approach.

There are other possible algorithms. Because the number and type of phases are determined before finding the length of phases, this algorithm will result in only one of many possible signal timings. No attempt is made to select an optimal timing. The ability of this algorithm to converge was individually checked on approximately 800 intersections of varying characteristics. Approximately six iterations are required to obtain four significant digits in the saturation flow rates.

Delay/volume relationships for approaches to signalized intersections in an adaptive model are considerably different from those typically seen for fixed-capacity facilities. Figure 3 shows a typical intersection with delays on all approaches varying as volume changes on a single, subject approach. Each approach at this four-way intersection has two shared lanes, there are no turning vehicles, and volumes at conflicting and opposing approaches are held at 800 vph. The volume on the

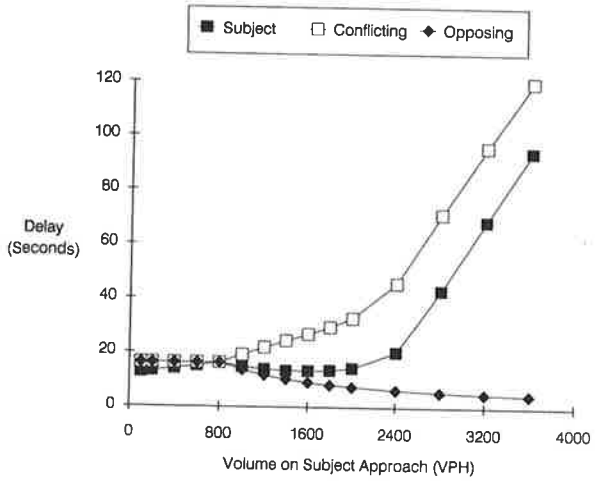


FIGURE 3 Delay on all approaches of a signalized intersection as a function of volume on a single approach (0 percent right turns, 0 percent left turns, 800 vph at opposing and conflicting approaches, no exclusive lanes, 3,600 vph ideal saturation flow rate, 20 mph speed).

subject approach is varied from 50 to 3,600 vph (the ideal saturation flow rate).

The delay at all approaches is strongly affected by variations of volume on the subject approach. The fact that the subject and opposing delay curves have roughly similar shapes is coincidental. Delay on the opposing approach declines with increasing subject approach volume beyond 800 vph. The decline is due to the increasingly ample green time to handle a given volume. Figure 3 also shows that delay on the subject approach is not even monotonic. Subject approach delay rises to a local maximum at 800 vph (the fixed volume on conflicting and opposing approaches), then declines to a local minimum at 1,600 vph, before increasing again.

Other tested intersections exhibit different curve shapes, different positions of local maxima and minima, and different values of delay. However, all tested intersections show a direct relationship between delay on the conflicting approaches and volume on the subject approach, declining values of delay on the opposing approach, and a clear lack of monotonicity.

Although they are not readily seen in Figure 3, multiple discontinuities can exist in the delay/volume relationships. Major culprits are the steps in the HCM's procedure dealing with left turns from shared lanes and the need to decide on an integer number of phases. Discontinuities can be eliminated only by deviating substantially from the HCM and accepted traffic flow theory.

Nonmonotone and discontinuous delay/volume relationships further complicate the assignment algorithm. At best, such messy relationships can introduce additional equilibrium solutions. At worst, they can prevent convergence to any equilibrium solution.

TEST OF ADAPTIVE TRAVEL FORECASTING

The UTOWN network, originally created for testing UTPS, did not contain traffic controls (see Figure 4). This network is known to be hostile to assignment algorithms. It was modified by incorporating signalized intersections and two-way stops (primarily at freeway off-ramps). Otherwise, an attempt was made to keep the two networks as similar as possible.

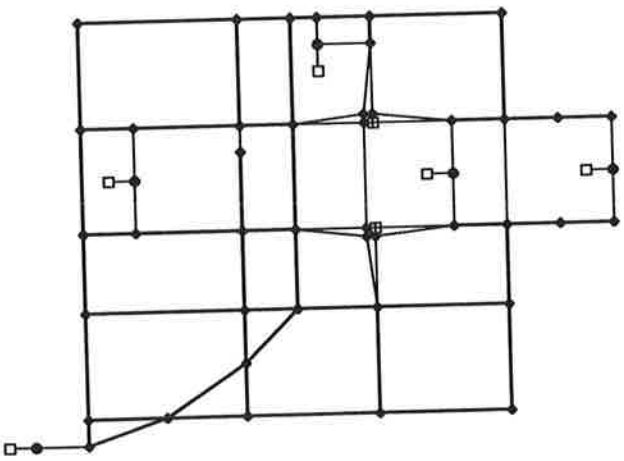


FIGURE 4 Nonadaptive UTOWN network without traffic control.

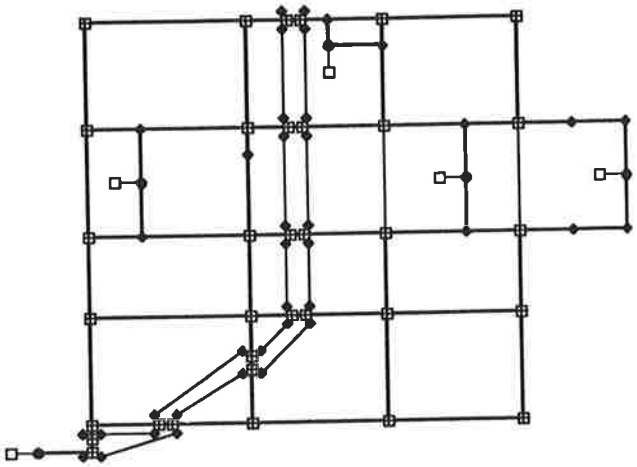


FIGURE 5 Adaptive UTOWN network with traffic control.

The modified, adaptive UTOWN network is shown in Figure 5. Both networks simulated a single peak hour with three trip purposes, but without mode split.

Convergence to an equilibrium solution needs to be checked, but the standard methods derived from Frank-Wolfe decomposition will not work in this case. Since we are interested in a user-optimal assignment, each trip should be assigned to a shortest path between its origin and destination. Therefore, it is possible to determine when equilibrium has been achieved by checking whether the used paths are indeed the shortest paths. A simple test can be devised that compares total travel time between two assignments.

Step 1. Run the model through the desired number of iterations. Obtain estimates of volumes. Recalculate the link travel times. Compute total travel time with the estimates of link volumes and the new travel times.

Step 2. Using the averaged trip table and new travel times from Step 1, run an all-or-nothing assignment. Do not recalculate link travel times. Compute total travel time.

Step 3. Compare the total travel times from Steps 1 and 2. The total travel time from Step 2 will always be the smaller. If they are nearly the same, convergence to an equilibrium solution has been achieved. If they differ significantly, there could be two causes: (*a*) more iterations are required or (*b*) the algorithm failed.

A similar test is provided in UTPS for fixed-demand assignments.

The test was performed on both UTOWN networks for varying numbers of iterations of elastic-demand, incremental assignment. As seen in Table 1, incremental assignment can produce an equilibrium solution on a network with explicit traffic controls. After 200 iterations the difference between Steps 1 and 2 was inconsequential. Equilibrium was effectively achieved after about 20 iterations. A comparison of Table 1 with Table 2 indicates that the adaptive travel forecast converges faster than the nonadaptive one.

A comparison of assigned volumes between the two networks indicated little agreement. The introduction of explicit traffic controls had a large effect on the forecast. In the original nonadaptive network the capacities of many intersections

TABLE 1 Convergence of Incremental Assignment on the Adaptive UTOWN Network

Iterations	Total Travel Time Step 1	Step 2	% Difference
1	1141630	1017512	12.220
2	1288223	1044020	23.339
5	1072852	1018350	5.352
10	1028061	1014823	1.013
20	1012969	1009968	0.297
50	1019219	1015559	0.360
100	1016405	1015892	0.050
200	1015288	1014971	0.031

TABLE 2 Convergence of Incremental Assignment on the Nonadaptive UTOWN Network

Iterations	Total Travel Time Step 1	Step 2	% Difference
1	2188242	1080782	102.468
2	1405340	1086944	29.293
5	1181057	1120440	5.410
10	1162928	1137757	2.212
20	1148820	1117023	2.847
50	1137249	1131264	0.529
100	1133918	1131264	0.235
200	1132658	1130776	0.166

were greatly exceeded even though individual link capacities were maintained.

For the adaptive network, approximately 11 percent of the computation time was devoted to intersection simulations. A greater amount of computation time would be required for a multihour assignment. Memory requirements increased slightly because of the need to keep track of the many turning movements for each all-or-nothing assignment.

One means of judging whether an adaptive assignment tends toward an all-or-nothing assignment is to count the number of used links. Both UTOWN networks only have 20 assignable origin-destination pairs (discounting intrazonal trips), so an all-or-nothing assignment requires less than half of the network. Table 3 gives the percentage of feasible link directions that were assigned a meaningful amount (defined as 1 percent of the saturation flow rate after 200 iterations) of traffic. Freeway ramps in the adaptive network were ignored for consistency. The results suggest that adaptive networks require a smaller percentage of link directions, but not as few as all-or-nothing assignment.

A version of QRS II that contains adaptation has recently been distributed to many users, so a considerable amount of experience is starting to be accumulated. In general, the observations obtained from tests of small networks have been confirmed on full-sized networks.

TABLE 3 Number of Possible and Used Directions

Network	Possible	Assigned	%Assigned
UTOWN Adaptive AON	126	58	46
UTOWN Adaptive Incremental	126	78	62
UTOWN NonAdaptive AON	112	51	46
UTOWN Adaptive Incremental	112	78	70

CHOOSING TRAFFIC CONTROLS

An adaptive travel forecasting model encourages planners to alter the types of traffic controls at intersections, if warranted by the assigned volumes. For instance, a choice could be made between signals and signs. If signs are chosen, a further choice must be made as to which approaches will get them. It is easy to imagine an adaptive network that automates this choice process in accordance with the *Manual on Uniform Traffic Control Devices* (MUTCD).

A simpler rule, often seen in traffic operations models, is to minimize delay. For example, the intersection of Figure 2 would be best served by a two-way stop, not a signal. For the preferred solutions, A and B, signs could be located on the one path with no assigned volume. None of the vehicles would ever be required to stop, so delay is further minimized.

Figure 6 shows total delay at an intersection with three alternative forms of traffic control. Subject and opposing volumes are varied together. Conflicting volumes are held constant at 200 vph, each approach has only one lane, and half of the volume makes a turn. It is seen that the three types of traffic controls perform almost equally well at a volume of 400 vph on the subject and opposing approaches. Below 400 vph the two-way stop is superior; above 400 vph the signal is superior. As expected from the MUTCD, other similar tests indicate that the point at which all controls are equally effective varies with the amount of conflicting volume.

It is outwardly practical to let the model decide on the nature of traffic control, provided that computer resources are sufficient to evaluate each arrangement of signs and signals. Such a model could be made consistent with the MUTCD. Unfortunately, allowing the model to choose signs or signals introduces additional discontinuities in the delay function. The discontinuities would further hamper the search for a good, stable solution.

Assuming that we are able to reasonably and automatically replicate the MUTCD at a single intersection, it is possible for both the alternatives of signs and signals to represent a

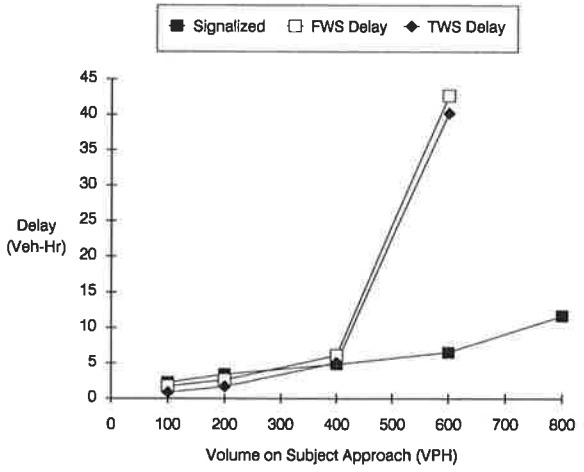


FIGURE 6 Total delay on all approaches for a four-way stop, a two-way stop, and a signal (opposing volume same as subject volume, conflicting volumes at 200 vph, 25 percent right turns, 25 percent left turns, one lane at all approaches, 20 mph speed).

user-optimal equilibrium solution. A signal is not just an upgraded sign; it can influence travelers' paths. Consider an intersection with two-way signing, where the volumes are large on the major street and potentially large on the minor street. Traffic on the minor street is subject to long delays from conflicting traffic, so the assignment algorithm holds the minor street volumes to a level below any of the MUTCD warrants. If the signs are replaced by a signal, the minor street volumes would increase dramatically. The signal is now warranted, simply because it is there. Users of QRS II report that this dilemma arises often enough in real networks to be worrisome.

Future-year networks are influenced by tradition and habit, especially in the placement of traffic control devices. This inertia can be built into the model by using the base-year delays as a starting point for future-year forecasts. For example, the network of Figure 2 is highly sensitive to the starting delays. The slightest delay disadvantage for a given street at the beginning of the algorithm would be sufficient for it to lose all of its traffic. Even if the placement of signs were made to be automatic (moved to the inner loop in Figure 1), the model would be reluctant to rearrange them or to upgrade them to a signal. It may be best to apply adaptive travel forecasting to make incremental adjustments to the current configuration rather than to attempt to consider all possible arrangements of traffic controls.

RELATIONSHIP OF ADAPTATION TO NETWORK CALIBRATION

Planners routinely "calibrate" their nonadaptive networks. That is, they adjust turn penalties, link speeds, and link capacities to obtain better agreement with base-year traffic counts. An extensive calibration exercise eliminates forecast/operational inconsistencies in the base-year network.

The need to calibrate is disturbing in itself; a truly good model should be able to provide accurate forecasts without much fiddling. Even more disturbing is the common practice of using the calibrated penalties, free speeds, and capacities for future-year forecasts. It cannot be assumed that these network attributes will be stable over time.

There are many possible alternatives to calibration for improving the match to existing traffic counts. Better delay/volume relationships would certainly help. The ability to forecast operational characteristics (such as cycle length, phase lengths, and coordination strategies) that affect delay also would help. Using base-year delays as a starting point can introduce some beneficial inertia. Since real highway systems are adaptive, it is certainly a mistake to use a calibration process that would rigidly fix (either explicitly or implicitly) operational characteristics that are known to be variable.

LEVELS OF ADAPTATION

As adaptive travel forecasting gains acceptance, planners will need to seriously consider the appropriate amount of adaptation for their networks. The HCM, of course, does not discuss adaptive travel forecasting, but it does indicate how adaptation can occur. The following levels of adaptation could be invoked, to various degrees, for any given network.

Level 0—no adaptation. Capacity is rigidly fixed on all streets and intersection approaches.

Level 1—low-cost traffic engineering improvements for isolated intersections without changing the type of traffic control. Capacity varies with the amount and nature of conflicting and opposing traffic. (Examples are signal timing and conversion of a through lane to an exclusive lane.)

Level 2—major traffic engineering improvements for isolated intersections. Capacity varies with the amount of and nature of conflicting, opposing, and subject approach traffic. (Examples are installation of signals and relocation of bus stops.)

Level 3—traffic engineering improvements involving a system of intersections. Capacity and delay vary with the nature of surrounding intersections. (An example is signal coordination.)

Level 4—minor geometric changes at isolated intersections. Capacity varies principally with volume on the subject approach. (Examples are adding exclusive lanes, removing on-street parking, and increasing curb radii.)

If all levels of adaptation were fully included in the network, the assignment would be constrained only by cost or by operational limitations, making it similar to the network design problem. With the procedures described here, planners can try to handle Levels 2, 3, and 4 subjectively.

There is no scientific way to determine the levels of adaptation for any given forecast. However, it is reasonable to expect all long-term forecasts to be adaptive to the extent that obvious design flaws or operational deficiencies in the highway system are eliminated. A good working assumption is that continuing efforts will be made to eliminate bottlenecks due to poor geometry, especially those with low-cost solutions. An important implication of adaptation is that planners may be able to ignore many small and isolated reductions in capacity when building their future-year networks.

CONCLUSIONS

Adaptation can provide additional realism to travel forecasts. An adaptive travel forecast requires (*a*) delay relationships that are properly sensitive to traffic controls and (*b*) the ability to modify the way in which the traffic controls are operated. We currently possess sufficient knowledge of traffic flow to allow a network to adapt its operational characteristics to forecast volumes. But we do not yet possess a complete understanding of the effect of adaptation on equilibrium traffic assignments.

An adaptive travel forecast estimates delays at intersection approaches through complex intersection simulations. The HCM contains some well-accepted relationships for this purpose. The implied delay/volume functions are discontinuous and nonmonotonic. Nonetheless, it is still possible (at least some of the time) to obtain acceptable equilibrium solutions with a form of incremental traffic assignment. This technique also permits the network to have elastic demands. A simple test is available to determine whether an equilibrium solution has been reached.

Because an adaptive network can reallocate resources among competing facilities, there need not be a unique solution.

Furthermore, there appears to be a tendency for an adaptive network to generate fewer used paths than a nonadaptive one.

Comparisons of small networks suggest that adaptive networks do not require much more computational effort, appreciably more computer memory, or much more user expertise. There are additional data requirements, but these are tolerable.

Adaptation, including better intersection delay relationships, promises to reduce the independence on network calibration. Adaptation can help avoid the locking of base-year operational characteristics into future-year forecasts.

Further research is needed to determine the prevalence of multiple equilibrium solutions in full-sized networks and the extent to which unwanted solutions can be avoided by setting starting delays consistent with existing traffic. Additional future research should compare the computed intersection delays with those experienced on actual networks.

ACKNOWLEDGMENTS

The author is indebted to Roger Tobin of GTE Laboratories for sharing the proof of the test of user-optimal equilibrium and his knowledge of asymmetric network problems. Portions of this research were supported by the Federal Highway Administration. The author expresses gratitude to the many QRS II users who provided their networks and insights.

REFERENCES

1. M. J. Smith. Traffic Signals in Assignment. *Transportation Research B*, Vol. 19B, No. 2, 1985, pp. 155–160.
2. R. W. Bentley and T. A. Lambe. Assignment of Traffic to a Network of Signalized City Streets. *Transportation Research A*, Vol. 14A, 1980, pp. 57–65.
3. M. J. Smith and M. Ghali. The Dynamics of Traffic Assignment and Traffic Control: A Theoretical Study. *Transportation Research B*, Vol. 24B, No. 6, 1990, pp. 409–422.
4. P. Marcotte. Network Optimization with Continuous Control Parameters. *Transportation Science*, Vol. 17, No. 2, May 1983, pp. 181–197.
5. D. R. Leonard and P. Gower. *User Guide to CONTRAM Version 4*. Transportation and Road Research Laboratory, Supplementary Report 735, 1982.
6. T. L. Magnanti and R. T. Wong. Network Design and Transportation Planning: Models and Algorithms. *Transportation Science*, Vol. 18, No. 1, Feb. 1984, pp. 1–55.
7. C. S. Fisk. A Conceptual Framework for Optimal Transportation Systems Planning with Integrated Supply and Demand Models. *Transportation Science*, Vol. 20, No. 1, 1986, pp. 37–47.
8. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
9. A. J. Richardson. A Delay Model for Multiway Stop-Sign Intersections. In *Transportation Research Record 1112*, TRB, National Research Council, Washington, D.C., 1987, pp. 107–112.
10. M. Kyte. Estimating Capacity of an All-Way Stop-Controlled Intersection. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.
11. *Transportation Research Circular 373: Interim Materials on Unsignalized Intersection Capacity*. TRB, National Research Council, Washington, D.C., July 1991.
12. A. Hansson. Swedish Highway Capacity Manual: Part 2, Capacity of Unsignalized Intersections. In *Transportation Research Record 667*, TRB, National Research Council, Washington, D.C., 1978, pp. 4–11.
13. K. G. Baass. The Potential Capacity of Unsignalized Intersections. *ITE Journal*, Oct. 1987, pp. 43–46.
14. L. LeBlanc, E. Morlock, and W. Pierskella. An Efficient Approach To Solving the Road Network Equilibrium Traffic Assignment Problem. *Transportation Research*, Vol. 9, 1975, pp. 309–318.
15. S. P. Evans. Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. *Transportation Research*, Vol. 10, 1976, pp. 37–57.
16. W. B. Powell and Y. Sheffi. The Convergence of Equilibrium Algorithms and Predetermined Step Sizes. *Transportation Science*, Vol. 16, 1982, pp. 45–55.
17. S. Dafermos. Traffic Equilibrium and Variational Inequalities. *Transportation Science*, Vol. 14, No. 1, 1980, pp. 42–54.
18. A. J. Horowitz. Convergence Properties of Some Iterative Traffic Assignment Algorithms. In *Transportation Research Record 1220*, TRB, National Research Council, Washington, D.C., 1990, pp. 21–27.
19. A. J. Horowitz. Tests of an Ad Hoc Algorithm for Elastic-Demand Equilibrium Traffic Assignment. *Transportation Research B*, Vol. 23, No. 4, 1989, pp. 309–313.
20. A. J. Horowitz. Convergence of Certain Traffic and Land-Use Equilibrium Assignment Models. *Environment and Planning A*, 1991, in press.
21. S. Lewis and S. McNeil. Developments in Microcomputer Network Analysis Tools for Transportation Planning. *ITE Journal*, Oct. 1986, pp. 31–35.
22. B. B. Heydecker. Some Consequences of Detailed Junction Modeling in Road Traffic Assignment. *Transportation Science*, Vol. 17, No. 3, Aug. 1983, pp. 263–281.

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

Left-Turn Adjustment Factors for Saturation Flow Rates of Shared Permissive Left-Turn Lanes

FENG-BOR LIN

For capacity analysis of signalized intersections, a left-turn adjustment factor is used in the 1985 *Highway Capacity Manual* (HCM) to account for the effect of left turns on saturation flow rates. When shared permissive left-turn lanes are the subject of analysis, the HCM uses a theoretical model to determine the related adjustment factors. Questions concerning the reliability of this model have been raised, and a recent study sponsored by the Federal Highway Administration has suggested that the HCM model be replaced by a set of new models. The new models, however, have serious flaws. An improved model that provides logical explanations of the causal relationships between the left-turn adjustment factor and its contributing factors is described. The contributing factors include opposing flow rate, number of opposing lanes, flow rate in the lane adjacent to the shared lane, proportion of left turns in the shared lane, proportion of left turns in opposing flow, proportion of opposing vehicles arriving in red interval, cycle length, green interval, and change interval. A numerical example is given to illustrate the applications of the improved model.

In the capacity analysis of signalized intersection, the 1985 *Highway Capacity Manual* (HCM) (*1*) requires that the saturation flow rate of a lane or a lane group be determined from the following formula:

$$S = S_o N F f_{LT} \quad (1)$$

where

- S = saturation flow rate (vphg) (vehicles per hour of effective green interval);
- S_o = ideal saturation flow rate, taken to be 1,800 vphg per lane;
- N = number of lanes in a lane group;
- F = the product of seven adjustment factors related respectively to lane width, heavy vehicles, approach grade, parking, blocking effects of local buses, area type, and right turns; and
- f_{LT} = adjustment factor for left turns.

When an analysis involves shared permissive left-turn lanes, the determination of the left-turn adjustment factor becomes a rather difficult problem.

A shared permissive left-turn lane refers to a lane from which opposed left-turn vehicles and vehicles of other directional movements can move into the intersection in a permissive left-turn signal phase. The presence of opposed left turns in such a lane disrupts vehicular movements and complicates the determination of f_{LT}. The 1985 HCM relies on a theoretical model to deal with this problem. The HCM concept in estimating f_{LT} has been used by Levinson (*2*) to develop a model for estimating the capacity of shared left-turn lanes. Questions concerning the reliability of the HCM model have been raised, however, and a set of new models was recommended in a recent study sponsored by FHWA (*3*). The new models were developed from regression analysis of field data. They are easy to use but do not properly account for the causal relationships between f_{LT} and its contributing factors. For example, f_{LT} can be expected to vary with opposing flow rate, yet the model recommended for single lane approach assumes implicitly that f_{LT} is independent of opposing flow rate. This flow may be partially responsible for the fact that the estimates obtained on the basis of that model have little correlation with observed values (*3*).

To provide an alternative, this paper describes an improved model that explains logically the causal relationships between f_{LT} and its contributing factors. This model deals with the left-turn adjustment factors for shared lane only (i.e., $N = 1$); it is developed on the basis of theoretical reasonings, field data, and computer simulation. A numerical example is provided to illustrate the applications of the model.

RESEARCH APPROACH

Equation 1 indicates that, if the saturation flow rate for a given F can be determined under a wide range of conditions, then a data base can be established for modeling f_{LT}. Because the vehicular movements in a shared left-turn lane can be interrupted by blocked left-turn vehicles, their saturation headway cannot be meaningfully defined and measured in the field to estimate the corresponding saturation flow rate. This problem can be overcome by taking advantage of the following relationship for $N = 1$ and for conditions represented by $F = 1.0$:

$$f_{LT} = \frac{S}{S_o} = \frac{CQ_{\max}}{G_e S_o} \quad (2)$$

where

- $Q_{\max}$ = capacity of a shared lane for $F = 1.0$ (level, 12-ft-wide approach lane and ideal conditions for other factors related to F),
- C = cycle length (sec), and
- G_e = effective green (sec).

Department of Civil and Environmental Engineering, Clarkson University, Potsdam, N.Y. 13699-5710.

For a given combination of C and G_e, Equation 2 shows that finding f_{LT} is the same as finding Q_{max}. Q_{max} can be determined by estimating the number of vehicles per cycle that can move out of an intersection. This expected number of departures includes the following components: early left turns, M_1; unblocked straight-through departures, M_2; departures in leftover green, M_3; and departures after green interval, M_4. Details of these components will be discussed later.

The modeling of Q_{max} is divided into two parts. The first part concerns a basic flow pattern as shown in Figure 1 (*top*). The notations used in this figure are defined as follows: Q_{01} = flow rate in the inside opposing lane (vph), Q_{02} = flow rate in the outside opposing lane (vph), and Q_a = flow rate in the lane adjacent to the shared lane (vph). The opposing lanes of the basic pattern do not contain left-turn vehicles. Such a situation may exist at a T-intersection or at a four-leg intersection where left turns from the inside opposing lane are prohibited. The maximum number of opposing lanes considered in this study is two.

The second part of the modeling effort involves the development of a mechanism to transform a flow pattern that contains left turns in Q_{01} into an equivalent basic flow pattern. This transformation involves the conversion of Q_{01} into an equivalent straight-through opposing flow $(Q_{01})_e$ for the estimation of the capacity of a shared lane. The transformation process is shown in Figure 1 (*bottom*). In this figure, P_o represents the proportion of left turns in the inside opposing lane.

Q_{max} is considered to be a function of such variables as Q_{01}, Q_{02}, Q_a, P_o, cycle length C, green interval G, signal change interval Y, proportion of left turns in shared lane P_s, and so forth. This study employs a combination of theoretical considerations, field data, and computer simulation to identify the causal relationships between Q_{max} and its contributing variables. The simulation model used in this study is a microscopic model developed at Clarkson University. This model can realistically simulate the stochastic movements of vehicles at intersections controlled by a variety of traffic signals. Details of this model are described elsewhere (*4*). An example comparison of the simulation outputs and field observations is given in Table 1.

One concern in modeling Q_{max} and its related f_{LT} is whether the effects of signal coordination should also be considered.

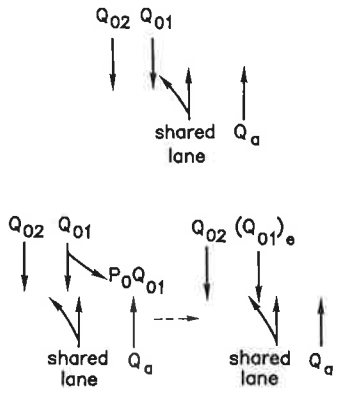


FIGURE 1 Basic flow pattern (*top*) and pattern transformation (*bottom*).

A simulation analysis reveals that the capacity of a shared left-turn lane may be affected by the presence of prominent cyclic platoons in the opposing lanes as a result of signal coordination. Nevertheless, as shown in Figure 2, the capacities of shared left-turn lanes when the arrivals are random tend to lie mostly within 50 vph of the values for coordinated signal operations. Such discrepancies are not alarmingly large in the context of capacity analysis. To avoid unnecessary complications, the arrivals will be assumed to be random for the purpose of modeling Q_{max} and f_{LT}. To account partially for the fact that arrivals are not necessarily random, the proportion of arrivals in red interval is included as a variable in the modeling process.

MODEL FOR BASIC FLOW PATTERN

Given the four components of departures per cycle, M_1, M_2, M_3, and M_4, the capacity of a shared lane for $F = 1.0$ can be determined as

$$Q_{max} = \left(\sum_{n=1}^{4} M_n\right) \frac{3{,}600}{C} \tag{3}$$

The modeling of each of the departure components is discussed.

Early Left Turns, M_1

Early left turns refer to those leading left-turn vehicles in various cycles that turn in front of the leading opposing vehicle shortly after green onset. Given the proportion of left turns in a shared left-turn lane (P_s) and the probability of early left turn α for a leading left-turn vehicle, the expected number of early left turns per cycle can be estimated as

$$M_1 = \alpha P_s \tag{4}$$

The values of α are often less than 0.5. Therefore, when P_s is much smaller than 1.0, M_1 becomes negligibly small.

Unblocked Straight-Through Departures, M_2

After the green onset, those straight-through vehicles ahead of the first left-turn vehicle can move out without facing the possibility of being blocked. To facilitate the estimation of such departures, the directional movements of the vehicles in a shared left-turn lane are classified into a series of events as shown in Figure 3. In this figure, K_2 represents the expected maximum number of straight-through vehicles that can move into the intersection before the green interval G expires. The value of K_2 can be determined as

$$K_2 = \frac{G - L_s}{H_s} \tag{5}$$

where L_s is lost time due to starting delays (sec), and H_s is saturation headway when only straight-through vehicles are present (sec).

TABLE 1 Comparison of Observed and Simulated Signal Operations

		Average Green, sec				Average Stopped Delay, sec/veh	
		Observed		Simulated			
Case	Phase	Mean	S.D.[1]	Mean	S.D.	Observed	Simulated
A	1	6.5	2.6	5.1	2.1	7.7[2]	6.6
	2	32.4	25.2	30.3	24.6		
B	1	33.8	18.2	31.9	17.6	11.1[3]	10.7
	2	5.4	2.1	5.1	2.3		
	3	24.0	0.0	24.0	0.0		
C	1	27.8	12.8	27.6	12.1	14.7[3]	14.9
	2	12.4	6.3	13.0	6.3		
D	1	18.8	9.2	17.5	8.1	Not available	Not available
	2	10.7	5.8	9.5	5.2		
E	1	29.3	7.6	30.2	9.1	42.3[4]	43.7
	2	20.5	0.7	20.2	0.9		
F	1	12.9	3.9	12.6	3.5	14.0[3]	13.5
	2	9.2	4.1	9.4	4.0		
	3	33.6	7.3	32.1	6.7	3.1[5]	2.7
G	1	10.8	6.0	12.1	3.7	25.1[6]	25.6
	2	30.0	0.8	30.0	0.0		
	3	19.1	0.5	19.2	0.0		
H	1	13.4	3.4	13.4	3.0	41.9[7]	41.6
	2	30.0	0.2	30.0	0.1		
	3	19.1	0.7	19.9	2.0		

[1]S.D. = Standard deviation
[2]single-lane flow with right turns and left turns
[3]exclusive left-turn flow
[4]shared-permissive left-turn flow (85% left turns)
[5]exclusive right-turn flow with right-turn-on-red
[6]shared-permissive left-turn flow (93% left turns)
[7]shared-permissive left-turn flow (95% left turns)

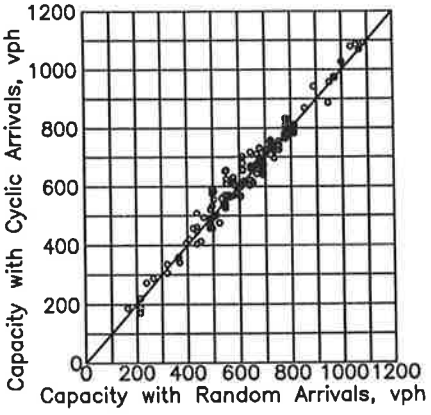


FIGURE 2 Capacities with cyclic platoon arrivals in opposing lanes versus capacities with random arrivals.

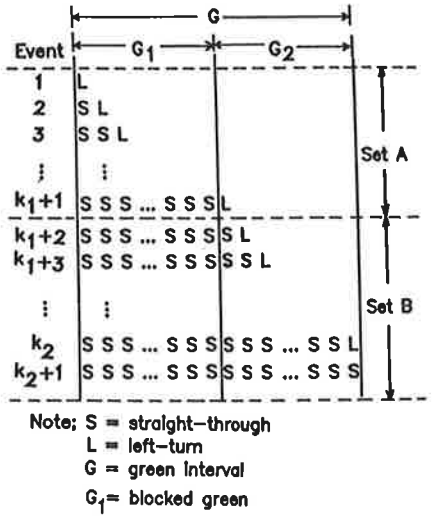


FIGURE 3 Classification of arrival sequences.

The expected number of unblocked straight-through departures per cycle can be estimated as

$$M_2 = \begin{cases} K_2 & \text{if } P_s = 0 \quad (6a) \\ \sum_{n=0}^{K_2-1} n(1-P_s)^n P_s + K_2(1-P_s)^{K_2} = & \\ [1 - P_s - (1-P_s)^{K_2}]/P_s & \text{if } P_s > 0 \quad (6b) \end{cases}$$

Departures in Leftover Green, M_3

In Figure 3, the green interval G is divided into two components: G_1 and G_2. G_1 is the average portion of the green interval consumed by the queueing vehicles in the opposing lanes before these vehicles cross the conflicting point. The conflicting point can be considered to be a representative location at which a leading left-turn vehicle that is blocked would come to a stop to wait for a suitable gap in the opposing flow. On the basis of G_1, the events shown in the figure are grouped into Set A and Set B. Set A events allow a maximum of K_1 straight-through vehicles to move into the intersection before a left-turn vehicle becomes the leading vehicle in the shared lane. The number of straight-through vehicles ahead of the first left-turn vehicle in Set B ranges from $K_1 + 1$ to K_2. The value of K_2 is determined from Equation 5, and K_1 can be determined in a similar manner as

$$K_1 = \frac{G_1 - L_s}{H_s} \geq 0 \quad (7)$$

After the departures of unblocked straight-through vehicles, a mix of left-turn and straight-through vehicles can move out by using leftover green intervals. For each of the Set A events, the leftover green interval is G_2. For the Set B events, the leftover green intervals depend on the portion of the green interval already consumed by the unblocked straight-through vehicles.

To facilitate the estimation of G_1 and the leftover green intervals, let us define the following additional variables:

i = inside opposing lane ($i = 1$) or outside opposing lane ($i = 2$),
S_{0i} = saturation flow rate of the ith opposing lane (vph),
x_i = number of queueing vehicles in the ith opposing lane at green onset,
m_{0i} = average number of queueing vehicles in the ith opposing lane at green onset,
q_{0i} = arrival rate in the ith opposing lane during green and signal change intervals (vph),
$q_{12} = q_{01} + q_{02}$ = sum of arrival rates q_{01} and q_{02} (vph),
R_o = proportion of arrivals in red in opposing lanes, and
β = time required for queueing vehicles to go from the stop line until they clear the conflicting point (sec).

On the basis of these definitions, m_{0i} and q_{0i} can be determined as

$$m_{0i} = \frac{Q_{0i} R_o C}{3{,}600} \quad (8)$$

and

$$q_{0i} = \frac{Q_{0i}(1 - R_o)\, C}{G + Y} \quad (9)$$

If there are x_i queueing vehicles in the ith opposing lane at the green onset, the portion of the green interval t_i consumed by these and subsequent queuing vehicles before they all cross the conflicting point may be estimated as

$$t_i = \begin{cases} 0 & \text{if } x_i = 0 \quad (10a) \\ \frac{3{,}600 x_i + L_s q_{0i}}{S_{0i} - q_{0i}} + L_s + \beta \leq G & \text{if } x_i > 0 \quad (10b) \end{cases}$$

Therefore, if $t_2 \leq t_1$, the queuing vehicles in the inside opposing lane would govern the time required to discharge all queueing vehicles in a given cycle. Otherwise, the queueing vehicles in the outside lane would govern. In other words, the queuing vehicles in the inside lane would govern if the following inequality holds:

$$\frac{3{,}600 x_1 + L_s q_{01}}{S_{01} - q_{01}} \geq \frac{3{,}600 x_2 + L_s q_{02}}{S_{02} - q_{02}} \quad (11)$$

This inequality can be rewritten as

$$x_2 \leq \frac{1}{3{,}600}\left[\frac{3{,}600 x_1 + L_s q_{01}}{S_{01} - q_{01}}(S_{02} - q_{02}) - L_s q_{02}\right] \quad (12)$$

Let X be the largest integer of x_2 that satisfies Equation 12 and $P(x_i)$ be the probability of having x_i queueing vehicles in Lane i at the green onset. For a given x_1, the portion of the green interval consumed by opposing queueing vehicles would be t_1 if x_2 has a value equal to or less than X. On the other hand, the portion of the green interval consumed by opposing queueing vehicles would be t_2 if x_2 has a value larger than X. Therefore, the expected value of G_1 can be estimated as

$$G_1 = \sum_{x_1=0}^{\infty} P(x_1)\left[t_1 \sum_{x_2=0}^{X} P(x_2) + \sum_{x_2=X+1}^{\infty} t_2 P(x_2)\right] \quad (13a)$$

If the arrivals in red are random, the $P(x_i)$ in Equation 13a can be determined from the following Poisson distribution:

$$P(x_i) = \frac{m_{0i}^{x_i} e^{-m_{0i}}}{x_i!} \quad (13b)$$

When only one opposing lane is present and arrivals are random, Equation 13a can be reduced to

$$G_1 = \frac{3{,}600 m_{01}}{S_{01} - q_{01}} + \left(\frac{L_s q_{01}}{S_{01} - q_{01}} + L_s + \beta\right)(1 - e^{-m_{01}}) \leq G \quad (14)$$

If two opposing lanes are present, Equation 13a can be approximated by a simplified equation. Let Q_H be the larger one of Q_{01} and Q_{02}. And, if $q_{01}/S_{01} \geq q_{02}/S_{02}$, let $m_H = m_{01}$,

$q_H = q_{01}$, $S_H = S_{01}$, $S_L = S_{02}$, and $q_L = q_{02}$. Otherwise, let $m_H = m_{02}$, $q_H = q_{02}$, $S_H = S_{02}$, $S_L = S_{01}$, and $q_L = q_{01}$. Then the simplified equation can be written as

$$G_1 = \frac{3{,}600 m_H}{S_H - q_H} + \left(\frac{L_s q_H}{S_H - q_H} + L_S + \beta\right)(1 - e^{-m_H}) + 2\gamma_1 e^{\gamma_2 - \gamma_3(1-\gamma_1)} \leq G \quad (15a)$$

where

$$\gamma_1 = \frac{q_L S_H}{q_H S_L} \quad (15b)$$

$$\gamma_2 = \frac{(0.042 + 0.01R_o)Q_H C}{3{,}600} \quad (15c)$$

and

$$\gamma_3 = e^{\frac{.08Q_H C}{3{,}600}} - 1 \quad (15d)$$

Equation 15a is the same as Equation 14 when only one opposing lane is present.

Twelve samples of field data were collected from three intersections to test Equations 13a and 15a. The observed values of G_1 and the estimates obtained from these equations are given in Table 2. The discrepancies between the observed and the estimated values were small.

Two other models have been recommended for estimating G_1 in the HCM and in the FHWA study (*2*). For comparison, the estimated values of G_1 obtained from these models are also given in Table 2. The model recommended in the FHWA study measures G_1 in terms of the time required for the front end of the last queueing vehicle to reach the stop line after green onset. Therefore, the estimates obtained from this model are increased by an amount equal to β. The reference line used in the model given in the HCM is unknown, and therefore no adjustments are made.

Given G_1, the leftover green interval for Set A events is $G_2 = G - G_1$. To facilitate further analysis, this leftover green interval is modified into an effective leftover green T_a, where

$$T_a = \begin{cases} G - G_1 & \text{if } G_1 \geq L_s \quad (16a) \\ G - G_1 - L_s\left(1 - \dfrac{G_1}{L_s}\right) & \text{if } G_1 < L_s \quad (16b) \end{cases}$$

This equation implies that if G_1 is greater than or equal to the lost time L_s, there is no need to adjust for starting delays because the lost time is completely accounted for by G_1. When G_1 is smaller than L_s, however, G_1 may not fully account for the starting delays associated with the vehicles in the shared lane. Equation 16b provides an adjustment for such a situation.

For the determination of the average leftover green interval for Set B events, one may proceed with the determination of the probability of an event being in Set B. This probability is $(1 - P_s)^{K_1+1}$, where P_s is the proportion of left turns in the shared lane. Therefore, for the Set B events and $0 < P_s < 1$, the average number of straight-through vehicles $\overline{K}_b$ that can move out before a left-turn vehicle becomes the leading vehicle in the shared lane can be determined as

$$\overline{K}_b = \frac{1}{(1 - P_s)^{K_1+1}}\left\{\left[\sum_{n=K_1+1}^{K_2-1} n(1 - P_s)^n P_s\right] + K_2(1 - P_s)^{K_2}\right\} = \frac{1}{P_s}[1 + K_1 P_s - (1 - P_s)^{K_2 - K_1}] \quad (17)$$

TABLE 2 Observed and Estimated Values of G_1

						G_1				
Case	Q_{01}	Q_{02}	C	G	R_o	Actual	Eq. 13a	Eq. 15a	HCM	FHWA
1	323	266	81.8	23.0	.79	20.0	20.0	20.2	13.0	23.0
2	278	246	81.8	23.0	.88	19.3	19.2	20.4	11.3	23.0
3	350	360	81.8	23.0	.75	20.4	21.6	23.0	16.4	23.0
4	388	0	70.6	30.0	.67	19.1	-	18.9	11.2	19.2
5	415	0	72.2	30.0	.68	19.4	-	20.6	12.6	20.3
6	502	0	71.2	30.0	.62	22.4	-	23.5	15.9	21.0
7	494	0	72.3	30.0	.71	24.7	-	24.4	16.0	24.6
8	440	0	68.9	30.0	.52	19.9	-	18.4	12.6	16.2
9	356	0	80.0	22.4	.80	21.5	-	21.0	14.2	22.2
10	274	0	103.7	19.2	.92	19.2	-	19.2	15.2	19.2
11	274	0	82.8	17.2	.72	17.2	-	17.2	11.8	17.2
12	218	0	84.0	18.2	.78	13.4	-	14.8	9.1	16.4

Note: L_s = 2.0 sec (assumed value)
β = 2.2 to 3.8 sec
Saturation flow rate: 1500 to 1750 vphg
Lost time per phase = 4 sec (assumed for HCM model)

The corresponding values of $\overline{K}_b$ for $P_s = 0$ and $P_s = 1$ are

$$\overline{K}_b = \begin{cases} K_2 & \text{if } P_s = 0 \quad (18a) \\ 0 & \text{if } P_s = 1 \quad (18b) \end{cases}$$

The average portion of the green interval consumed by these $\overline{K}_b$ vehicles is approximately $\overline{K}_b H_s + L_s$. The corresponding effective leftover green interval T_b for the Set B events becomes

$$T_b = G - \overline{K}_b H_s - L_s \quad (19)$$

With the exception of the last event shown in Figure 3, a left-turn vehicle becomes the leading vehicle in the shared lane after unblocked straight-through vehicles have moved out. This leading vehicle has to use the gaps in the opposing flow to move out. Assuming that a waiting left-turn driver will only accept those gaps longer than τ sec, the average number of gaps J that will be rejected before a gap is accepted can be estimated as

$$J = \sum_{n=0}^{\infty} nZ(h \le \tau)^n[1 - Z(h \le \tau)] \quad (20)$$

where $Z(h \le \tau)$ is the probability that a gap h in the opposing flow is less than or equal to τ.

It can be shown that, for random arrivals, the average number of rejected gaps can be approximated as

$$J = e^{\frac{q_{12}\tau}{3,600}} - 1 \quad (21)$$

The value of τ can be considered to be equal to the median of the lengths of accepted gaps. Typical values of τ are between 4.5 and 5.5 sec. For such values of τ, the average length of each rejected gap can be approximated as $\tau/2$ without incurring significant errors in estimating the capacity of a shared left-turn lane. On the basis of this approximation, a waiting left-turn driver will wait an average of $J\tau/2$ sec before accepting a gap. After a decision is made to accept a gap, it will take an additional δ sec for the left-turn vehicle to cross the conflicting point and for the next vehicle to move up. Typical values of δ are between 2.0 and 2.5 sec.

Let H_x represent the expected portion of the green interval consumed by the first left-turn vehicle. Then, H_x can be determined as

$$H_x = \frac{\tau}{2}\left(e^{\frac{q_{12}\tau}{3,600}} - 1\right) + \delta \quad (22)$$

After the first left-turn vehicle has moved out, the vehicle following can be either a straight-through or a left-turn vehicle. The expected time H_y needed by either of such vehicles to move out is not amenable to simple analytical modeling. Nevertheless, when the opposing flow does not exist, the saturation headway H_o of the vehicles in the shared lane can realistically be estimated as

$$H_o = (1 - P_s)H_s + P_s H_e \quad (23)$$

where H_e is saturation headway only when unopposed left-turn vehicles are present and H_s is saturation headway when only straight-through vehicles are present.

When opposed left turns exist, the average departure headway can be expected to exceed H_o and to increase with the proportion of left turns in the shared lane P_s and the opposing flow rate q_{12}. A logical model characterizing this relationship between H_y, P_s, and q_{12} is

$$H_y = H_o e^{A\left(\frac{q_{12}}{100}\right)^B} \quad (24a)$$

For random arrivals, simulation reveals that the coefficients A and B in this equation can be estimated as

$$A = 0.18P_s^{0.68} \quad (24b)$$

and

$$B = 1.02P_s^{-0.15} \quad (24c)$$

The first left-turn vehicle in a Set A event consumes H_x sec of the effective leftover green T_a, and the subsequent vehicles consume an average of H_y sec each. Thus, the expected number of departures W_a related to Set A events is

$$W_a = \begin{cases} \dfrac{T_a}{H_x} & \text{if } T_a \le H_x \quad (25a) \\ 1.0 + \dfrac{T_a - H_x}{H_y} & \text{if } T_a > H_x \quad (25b) \end{cases}$$

Similarly, the expected number of departures W_b related to Set B events is

$$W_b = \begin{cases} \dfrac{T_b}{H_x} & \text{if } T_b \le H_x \quad (26a) \\ 1.0 + \dfrac{T_b - H_x}{H_y} & \text{if } T_b > H_x \quad (26b) \end{cases}$$

Set A events and Set B events account for a total probability of $1 - (1 - P_s)^{K_1+1}$ and $(1 - P_s)^{K_1+1}$, respectively. Therefore, the total expected departures during the effective leftover green in a cycle is

$$M_3 = W_a[1 - (1 - P_s)^{K_1+1}] + W_b(1 - P_s)^{K_1+1} \quad (27)$$

Departures After Green Interval, M_4

For capacity analysis of signalized intersections, the 1985 HCM assumes implicitly that two blocked vehicles can move out of the intersection after the green interval expires. On the basis of this assumed condition, simulation data generated in this study show that the following equation can provide reasonable estimates of M_4:

$$M_4 = 1.3 + 0.0033\, e^{-0.007G} P_s^{0.2} q_{12} \le 2.0 \quad (28)$$

This equation implies that M_4 varies from 1.3 to 2.0 vehicles. The value of M_4 is 1.3 vehicles when opposed left turns

do not exist ($P_s = 0.0$ or $q_{12} = 0.0$), and it reaches 2.0 vehicles when blocked vehicles are present in virtually every cycle after the green interval expires.

STRAIGHT-THROUGH EQUIVALENT OF Q_{01}

The present of left-turn vehicles in the inside opposing lane can increase the number and the size of the gaps usable to the drivers in a shared left-turn lane. Therefore, when the proportion of the left-turn vehicles in the opposing lane increases, the capacity of a shared left-turn lane also increases. The increase in capacity can be affected by several other factors. This phenomenon is shown in Figure 4 on the basis of simulation data. By comparing such data as shown in this figure with data for an inside opposing flow Q_{01} that contains no left turns, one can determine the straight-through equivalent $(Q_{01})_e$ of Q_{01}. An example of $(Q_{01})_e$ expressed as a ratio to Q_{01} is shown in Figure 5.

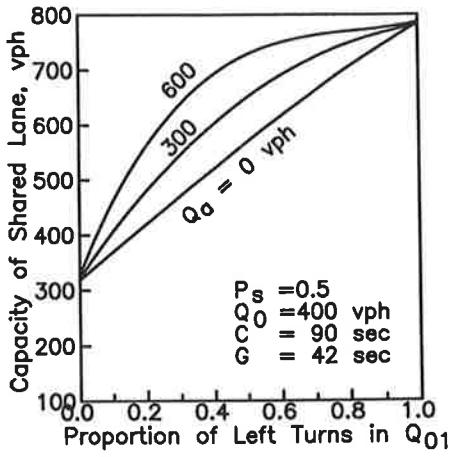


FIGURE 4 Effects of left turns in Q_{01} on capacity of shared left-turn lanes.

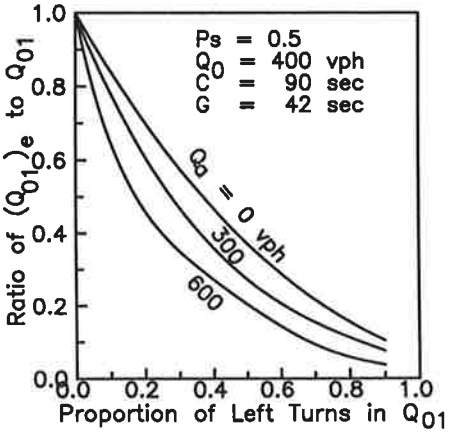


FIGURE 5 Characteristic relationships between Q_{01} and its straight-through equivalent $(Q_{01})_e$.

In general, the relationship between $(Q_{01})_e$ and Q_{01} can be represented by

$$(Q_{01})_e = Q_{01}(1 - \beta_1 P_o)e^{-\beta_2 P_o} \tag{29a}$$

where β_1 can be treated as a constant coefficient with a value of 0.97 and β_2 is a function of several variables.

To identify the relationship between β_2 and the influencing variables, a very large number of simulation runs were performed to develop a data base. On the basis of these simulation data, an analysis was carried out to isolate the effects of each variable on β_2. This effort produced a set of equations for determining β_2.

To facilitate the determination of β_2, let us define two functions, δ_1 and δ_2, as follows:

$$\delta_1 = -\left(e^{1.39\frac{G+Y}{C}} - 1\right)P_s \tag{29b}$$

and

$$\delta_2 = \left(0.0006 + 0.00233\frac{G+Y}{C} + 0.0021P_s\right)Q_a \tag{29c}$$

Then, for $Q_{01} \leq 400$ vph, the value of β_2 is

$$\beta_2 = 1.5\, e^{-2.7P_s} + \frac{0.9Q_{01}}{400} e^{\delta_1+\delta_2} \tag{29d}$$

And, for $Q_{01} > 400$ vph, the value of β_2 is

$$\beta_2 = 1.5\, e^{-2.7P_s} + 0.9\, e^{\delta_1+\delta_2} + \frac{Q_{01} - 400}{400}\left(4.5 - 3.6\frac{G+Y}{C} - 0.5\,P_s\right) \tag{29e}$$

The straight-through equivalent of each vehicle in Q_{01} (i.e., $(Q_{01})_e/Q_{01}$ as determined from Equation 29a through 29e) has several characteristics that are worth noting. First, larger Q_a and P_o increase the chance of an opposing left-turn vehicle being blocked and, thus, allow more vehicles in a shared left-turn lane to move out. Under such conditions each opposing vehicle becomes less of a factor affecting the capacity of a shared left-turn lane. This is the reason why, as shown in Figure 5, $(Q_{01})_e/Q_{01}$ decreases with Q_a and P_o. An increase in the opposing flow Q_{01} has similar effects. In contrast, $(Q_{01})_e/Q_{01}$ increases with P_s and $(G + Y)/C$.

APPLICATIONS

In general, the applications of the analytical model described involves the transformation of Q_{01} into a straight-through equivalent, the determination of Q_{max} for the basic flow pattern resulting from the transformation, and the use of Equation 2 to determine f_{LT}. A numerical example is given in Table 3 to illustrate the applications of the model. The example involves a flow pattern that has $Q_a = 400$ vph, $Q_{01} = 200$ vph with $P_o = 0.2$, $Q_{02} = 350$ vph, $R_o = 0.32$, and $P_s = 0.8$. The related signal control has a cycle length C of 50 sec, a green interval G of 30 sec for the permissive left-turn phase,

TABLE 3 Estimation of Q_{max} and f_{LT}—An Example

A. Transformation of Q_{01} into $(Q_{01})_e$

$\beta_1 = 0.97$ (in Eq. 29a); $\delta_1 = -1.26$ (Eq. 29b); $\delta_2 = 1.55$ (Eq. 29c);
$\beta_2 = 0.77$ (Eq. 29d); $(Q_{01})_e = 138$ vph (Eq. 29a)

B. Estimation of Q_{max} [set Q_{01} to $(Q_{01})_e = 138$ vph]

Determination of M_1
$M_1 = 0.16$ veh/cycle (Eq. 4)

Determination of M_2
$K_2 = 14$ (Eq. 5); $M_2 = 0.25$ veh/cycle (Eq. 6b)

Determination of M_3
$m_{01} = 0.61$ (Eq. 8); $m_{02} = 1.56$ (Eq. 8); $q_{01} = 138$ (Eq. 9);
$q_{02} = 350$ (Eq. 9); $q_{12} = 488$; In Eqs. 15a through 15d, $S_{01} = S_{02} = 1{,}800$;
$m_H = 1.56$; $S_H = S_L = 1{,}800$; $q_H = q_{02} = 350$; $q_L = q_{01} = 138$; $Q_H = Q_{02} = 350$;
$\gamma_1 = 0.39$; $\gamma_2 = 0.22$; $\gamma_3 = 0.475$; $G_1 = 8.5$ (Eq. 15a); $T_a = 21.5$ (Eq. 16a);
$K_1 = 3.3$ (Eq. 7); $K_b = 4.6$ (Eq. 17); $T_b = 18.8$ (Eq. 19); $H_x = 5.5$ (Eq. 22);
$H_o = 2.08$ (Eq. 23); $A = 0.155$ (Eq. 24b); $B = 1.055$ (Eq. 24c);
$H_y = 4.7$ (Eq. 24a); $W_a = 4.40$ (Eq. 25b); $W_b = 3.83$ (Eq. 26b);
$M_3 = 4.40$ veh/cycle (Eq. 27)

Determination of M_4
$M_4 = 2.6 > 2.0$ (Eq. 28); set $M_4 = 2.0$ veh/cycle

Determination of Q_{max}
$Q_{max} = 490$ vph (Eq. 3)

C. Estimation of f_{LT}

$f_{LT} = 0.45$ (Eq. 2 with $G_e = G = 30$ sec)

and a change interval of 4 sec. The saturation flows for straight-through movements and unopposed left turns are, respectively, 1,800 vphg ($H_s = 2.0$ sec) and 1,700 vphg ($H_e = 2.1$ sec). In addition, the following parameters are used: $\tau = 5.5$ sec, $\delta = 2.5$ sec, $\beta = 2.5$ sec, $\alpha = 0.2$, and $L_s = 2.0$ sec. The model gives an estimated Q_{max} of 490 vph and a f_{LT} of 0.45. In comparison, direct simulation gives a Q_{max} of 442 vph.

Over a wide range of conditions, the values of Q_{max} estimated from the analytical model and those determined from direct simulation are mostly within 50 vph of each other. This characteristic is shown in Figure 6. Figure 7 further shows the ability of the model to estimate Q_{max} and the related f_{LT} when a lane is changed from a straight-through-only lane ($P_s = 0.0$) to a shared left-turn lane ($0.0 < P_s < 1.0$) and, finally, to an exclusive left-turn lane ($P_s = 1.0$).

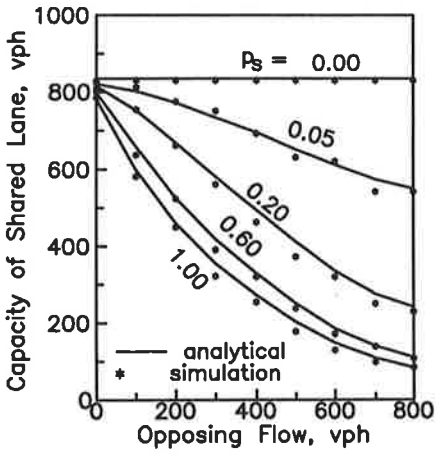


FIGURE 6 Capacities estimated from analytical model versus capacities determined from direct simulation.

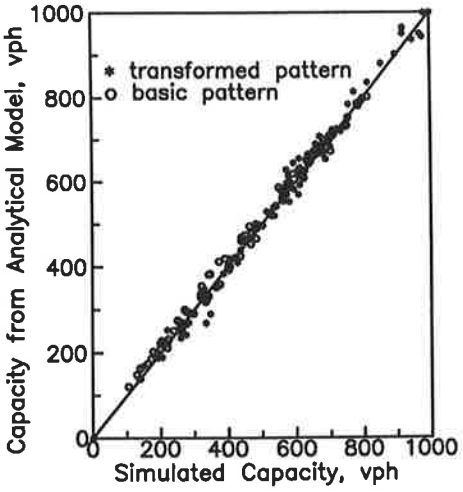


FIGURE 7 Simulated capacities and estimates obtained from analytical model for basic flow patterns with 90-sec cycle and 42-sec green interval.

CONCLUSIONS

The left-turn adjustment factor for a shared left-turn lane is a complex function of a number of variables. Reflecting this complexity, the analytical model developed in this study is much more complicated than the HCM and FHWA models. The added sophistication enables the resulting model to provide better explanations of the causal relationships between the adjustment factor and its contributing factors.

In comparison with elaborate microscopic simulation, the analytical model developed in this study can yield equally realistic estimates. Field data may be collected in future studies to test and modify the constant coefficients associated with the model.

ACKNOWLEDGMENTS

This paper is based on work supported in part by the New York State Science and Technology Foundation. The author wishes to thank John F. Edwards, a graduate student 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. H. S. Levinson. Capacity of Shared Left-Turn Lanes—A Simplified Approach. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 45–52.
3. R. P. Roess, J. M. Ulerio, and V. N. Papayannoulis. Modeling the Left-Turn Adjustment Factor for Permitted Left Turns Made from Shared Lane Groups. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 138–150.
4. F.-B. Lin. Knowledge Base on Semi-Actuated Signal Control. *Journal of Transportation Engineering*, ASCE, Vol. 117, No. 4, 1991, pp. 398–417.

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

Oversaturation Delay Estimates with Consideration of Peaking

NAGUI M. ROUPHAIL AND RAHMI AKÇELIK

A deterministic oversaturation queueing model that uses a generalization of the peak hour factor concept of the U.S. *Highway Capacity Manual* (HCM) as a simple variable demand model is described. The model is used to explore several issues related to oversaturation models. In particular, the relationship between the delay measurement methods (queue sampling and path trace) and the delay definitions used in the corresponding analytical delay models is investigated with a view to level of service assessment and performance prediction. The differences in delay definitions and delay measurement methods are negligible for undersaturated conditions (low to medium v/c ratios). However, as flows approach capacity (high v/c ratios below capacity) and exceed capacity (v/c ratio greater than 1), the selection of the duration of the flow period, delay definition, and delay measurement method affects delay estimates significantly. Substantial differences in delay and queue estimates are found between the cases of peak flow and maximum delay periods regardless of the delay measurement method. The use of the average delay experienced by individual vehicles in a maximum delay period creates problems in system performance analysis. A delay definition based on a maximum delay period reveals an inconsistency in relation to delays measured in the field. Whereas the HCM recommends that field delays be measured in the peak flow period, the maximum delay period does not coincide with the peak flow period. It is therefore important that the delay definition implied by the present HCM delay formula for signalized intersections be clarified.

The U.S. *Highway Capacity Manual* (HCM) (*1*) qualifies the signalized intersection delay equation given in Chapter 9 as:

> The delay equation may be used with caution for up to (a degree of saturation of) 1.2, but delay estimates for higher values are not recommended. Oversaturation, i.e. $x > 1.0$, is an undesirable condition that should be ameliorated if possible.

However, from a congestion management viewpoint, it is desirable to be able to predict oversaturation delays without any limitation.

A paper by Akçelik (*2*), which discussed the x^2 factor in the second (random plus oversaturation) term of the HCM delay model, is related to this issue. For background information on delay models in general, and the HCM delay equation in particular, the reader is referred to McShane and Roess (*3*).

Messer (*4*) analyzed oversaturation delays in relation to the justification of the x^2 factor. Through subsequent private communication with Messer, it is understood that the x^2 factor is intended to convert the delay in the peak flow period (15 min in the HCM) to a peak delay value that is the maximum delay observed sometime during or after the peak flow period.

This paper explores the issues related to oversaturation models by highlighting differences between various delay definitions (delay during the peak flow period versus a maximum delay period, and delay measured in the specified period versus delay experienced by all vehicles arriving during the specified period) and the corresponding delay measurement methods (queue sampling and path trace).

A deterministic (nonrandom) oversaturation queueing model is presented that uses a generalization of the peak hour factor concept of the HCM as a simple variable demand model. Consideration is given to the choice of the duration of the peak flow period and to the average flow rates and degrees of saturation in the peak and nonpeak flow periods.

A numerical example is given to demonstrate the effects of different delay definitions and the choice of the peak flow period on estimates of delay and queue statistics.

NOTATION

Symbol	*Definition*
T	Duration of the total flow period
T_p	Duration of the peak flow period in the total flow period ($0 < T_p \leq T$)
T_o	Duration of the oversaturation period (time from the start of the peak flow period until the queues clear), $T_o = (1 - \alpha)x_p T_p/(1 - \alpha x_p)$
q_p	Average flow rate in the peak flow period (during T_p)
q_n	Average flow rate in the nonpeak flow period (during $T - T_p$)
q_a	Average flow rate during the total flow period (during T)
α	The ratio of nonpeak and peak flow rates, $\alpha = q_n/q_p$
c_p	Peak period capacity throughout the oversaturation period (T_o)
c_n	Nonpeak period capacity outside the oversaturation period ($T - T_o$)
x_p	Peak period degree of saturation (v/c ratio), $x_p = q_p/c_p$
αx_p	Degree of saturation during the remainder of the oversaturation period following the peak flow period ($T_o - T_p$), $\alpha x_p = q_n/c_p$
x_n	Nonpeak period degree of saturation outside the oversaturation period ($T - T_o$), $x_n = q_n/c_n$
PFF	Peak flow factor: the ratio of average flow rates in the total and peak flow periods, PFF = q_a/q_p
PHF	Peak hour factor: special case of PFF where the total flow period T is 1 hr, PHF = q_a/q_p
PTF	Peak time factor: the ratio of durations of the peak and total flow periods, PTF = T_p/T
y	A variable that defines the delay period. This is the time from the start of the peak flow period to the start of a floating delay period of duration T_p in T (for $y = 0$, the

N. M. Rouphail, Urban Transportation Center, The University of Illinois at Chicago, Suite 700 South, 1033 West Van Buren Street, Chicago, Ill. 60607-9940. R. Akçelik, Australian Road Research Board, P.O. Box 156, Nunawading, Victoria 3131, Australia.

Symbol	Definition
	delay period is the peak flow period, and $y = y_m$ corresponds to the maximum delay period)
y_m	The value of y that gives the maximum value of delay (total or average) for any floating delay period of duration T_p in T
D_y	Total oversaturation delay for the delay period as defined by variable y
d_y	Average oversaturation delay for the delay period as defined by variable y
N_{sy}	Queue size at the start of the delay period as defined by variable y
N_{ey}	Queue size at the end of the delay period as defined by variable y
N_{ay}	Average overflow queue size for the delay period as defined by variable y

Flow and capacity (q, c) are in vehicles per hour (vehicles per second), total delay (D) is in vehicle-hours (vehicle-seconds), average delay (d) is in hours (seconds) per vehicle, and queue size (N) is in vehicles.

ISSUES AND DEFINITIONS

When demand flow of traffic in a lane (or lane group) at an intersection exceeds the capacity by a large margin as represented by a high degree of saturation (volume/capacity ratio $x >> 1.0$), overflow queues develop and persist over a considerable period of time. In such oversaturated conditions, stochastic variations in demand flows have minimal influence on the system operation, and a simple deterministic input-output queueing model is adequate for representing the resulting queueing phenomenon.

Deterministic oversaturation models are key predictors of delays and queues under highly congested conditions. Such models are also important in defining continuum models that allow for stochastic variations in demand flows, have time-dependent characteristics, and apply to undersaturated as well as oversaturated conditions.

A continuum model (i.e., an entire delay or queue length curve) with time-dependent characteristics can be obtained by means of the coordinate transformation method (*5*,*6*). This process essentially shifts the steady-state (stochastic) queueing model from its vertical asymptote to a time-dependent deterministic asymptote as shown in Figure 1. The resulting time-dependent model incorporates both random and oversaturation delays for high degrees of saturation. Therefore, the positioning of the time-dependent asymptote has profound implications for delay and queue estimation around capacity, which are the most relevant operating conditions in an intersection design context (*7*).

The deterministic oversaturation delay function provides a lower bound of delay for oversaturated conditions, which should apply independent of traffic control (e.g., signalized or unsignalized) or arrival characteristics (e.g., random or platooned).

Several definitional issues arise in deriving equations that express delay and queue statistics in an oversaturation queueing model: (*a*) Is the delay measurement method used in the field consistent with the delay models used in operational analysis? (*b*) Which combination of time period and delay measurement method should be used for level of service (LOS) assessment?

With regard to the first point, two basic methods can be identified. The HCM recommends that field delays be measured using a periodic queue sampling process (at 10- to 20-sec intervals). Total delay is then estimated as the area under the queue profile. Average delay is computed by dividing the total delay by the number of vehicle arrivals during the study interval. On the other hand, the path-trace method measures individual vehicle delays from arrival to departure time, even if the latter occurred beyond the observation period. Delay models used in Australia (*2*,*6*,*8*) are consistent with the path-trace method. The queue sampling and path-trace methods of delay measurement are compared in Figures 2a and 2b.

The second and key issue is concerned with the selection of a combination of time period and delay measurement method for the purpose of LOS assessment. Messer (*4*) points out that the maximum vehicle delays in an oversaturated period

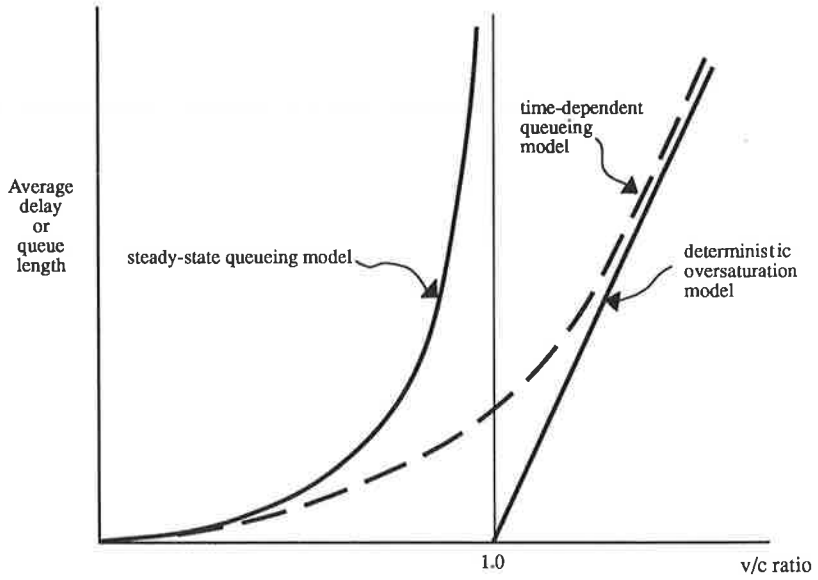


FIGURE 1 Steady-state and time-dependent queueing models.

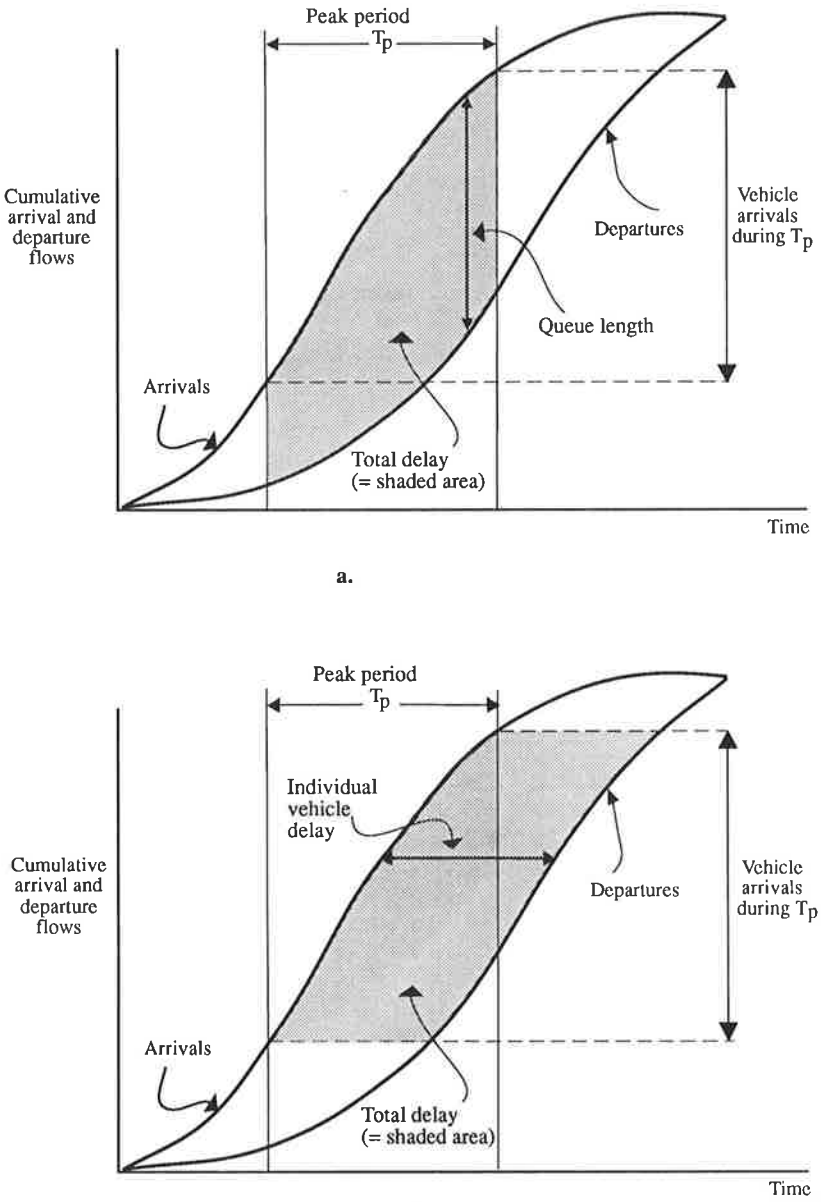


FIGURE 2 Queue sampling (*a*) and path-trace (*b*) methods of delay measurement.

typically occur at or beyond the termination of the peak flow period; in other words, the peak flow and peak delay periods do not necessarily coincide. He goes on to suggest that the maximum delay period should be used for LOS assessment. In that context, he relates the use of the x^2 factor in the incremental delay term in the HCM delay formula to the estimation of the peak delay in any floating 15-min period within the peak hour.

The formula given by Messer (*4*) to calculate the peak oversaturation delay corresponds to the queue sampling method. However, it is suggested that a delay formula corresponding to the path-trace method is more relevant to the LOS concept because this represents the delay experienced by individual vehicles. In this paper, a generalization of Messer's maximum delay formula and a formula based on the path-trace method are developed.

USE AND EXTENSION OF THE PEAK HOUR FACTOR CONCEPT

For the identification and evaluation of the period when the maximum delay occurs, a detailed analysis of the demand flow profile around the peak flow conditions is needed. This process, however, may be too complex and cumbersome to be applied in a simple operational analysis context as in the HCM since signal timing and capacity analysis for each time interval in the peak period is required. However, a simple

representation of the peaking characteristics can be gained through the use of the peak hour factor (PHF) parameter. For this purpose, a simplified demand profile is described with average flow rates q_p and q_n in the peak flow period (T_p) and the nonpeak flow period ($T - T_p$), where T is the total flow period ($0 < T_p \leq T$).

The PHF parameter (see Figure 3) characterizes the peaking of demand flows by relating the average flow rate q_a in the peak hour ($T = 1$ hr) and the average flow rate q_p in the peak flow period ($T_p \leq 1$ hr) through

$$\text{PHF} = q_a/q_p \tag{1}$$

By the principle of conservation of vehicles,

$$q_a T = q_p T_p + q_n(T - T_p) \tag{2}$$

From Equations 1 and 2 with $T = 1$ hr, the nonpeak flow rate q_n is expressed as

$$q_n = q_p \frac{(\text{PHF} - T_p)}{(1 - T_p)} \tag{3}$$

Note that PHF ≤ 1.0 and $q_n \geq 0$ since, by definition, $0 < T_p \leq 1.0$. When $T_p = T = 1$ hr, $q_a = q_p$ and $q_n = 0$, therefore PHF $= 1.0$. This corresponds to a constant demand rate during the total flow period.

The PHF parameter may be generalized by specifying a general value for the total flow period T (instead of $T = 1$ hr) during which the peak flow period T_p occurs ($0 < T_p \leq T$). For the generalized parameter, let us use the term peak flow factor, PFF, instead of PHF, and let us define a new parameter called the peak time factor, PTF:

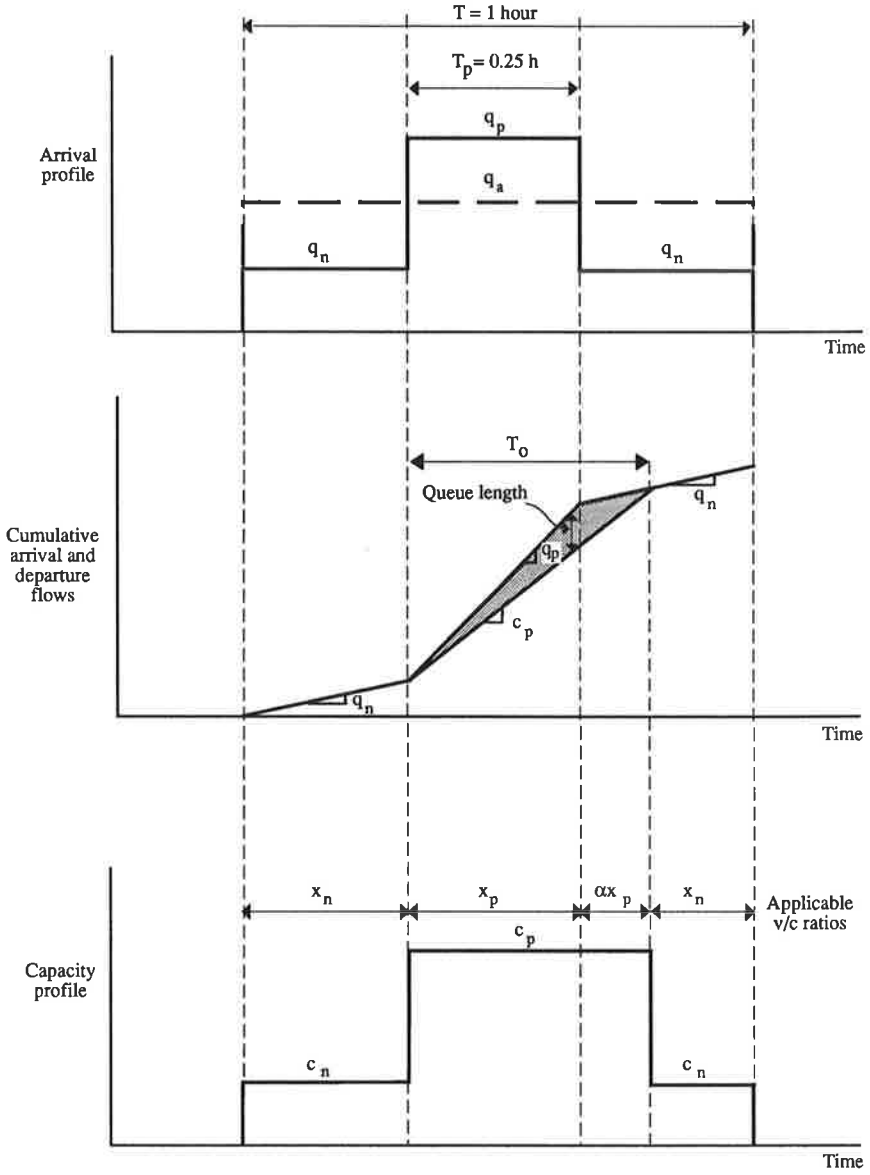


FIGURE 3 Demand, capacity, and queue profiles using the peak hour factor concept.

$$\text{PFF} = q_a/q_p \tag{4a}$$

$$\text{PTF} = T_p/T \tag{4b}$$

It can be shown that PTF ≤ PFF ≤ 1.0. In the HCM, T_p = 0.25 hr and T = 1 hr are used, yielding 0.25 ≤ PFF ≤ 1.0.

Rewriting Equation 3 for the general case gives

$$q_n = q_p \frac{(\text{PFF} - \text{PTF})}{(1 - \text{PTF})} \tag{5}$$

Defining parameter $\alpha = q_n/q_p$ ($0 \le \alpha \le 1$), Equation 5 can be expressed as

$$q_n = \alpha\, q_p \tag{6a}$$

where

$$\alpha = \frac{(\text{PFF} - \text{PTF})}{(1 - \text{PTF})} \tag{6b}$$

The peak period capacity, c_p, is considered to apply as long as oversaturation persists, since this situation represents heavy demand conditions (e.g., leading to maximum green times at traffic signals). The oversaturation period (i.e., the time from the start of the peak flow period until the time the oversaturation queue clears) is given by

$$T_o = \frac{(1 - \alpha)x_p T_p}{1 - \alpha x_p} \tag{7}$$

where $x_p = q_p/c_p$.

The validity of the oversaturation queueing model given in this paper is predicated on the assumption that the peak period queues must not grow after the termination of the peak flow period so that the oversaturation period is not indefinite (Equation 7). This constraint is expressed as

$$\alpha x_p < 1.0 \tag{8a}$$

This is equivalent to

$$x_p < 1.0/\alpha \quad \text{or} \quad q_n < c_p \tag{8b}$$

For example, applying the HCM values T = 1 hr, T_p = 0.25 hr, PFF = PHF, PTF = 0.25, Equation 6b gives

$$\alpha = \frac{\text{PHF} - 0.25}{0.75} \tag{9}$$

and from Equation 8b, the condition for the oversaturation queues to clear is

$$x_p \le \frac{0.75}{\text{PHF} - 0.25} \tag{10}$$

For example, when PHF = 0.9, x_p must not exceed 1.15 for the oversaturation queues to clear after the peak flow period.

DEVELOPMENT OF AN OVERSATURATION QUEUEING MODEL

An oversaturation queueing model is given here that extends Messer's original formulation (*4*) as follows:

- The model is used to derive the delay and queue statistics for either the peak flow period or a maximum delay period of duration T_p.
- Separate equations are given for estimating delays in accordance with the queue sampling and path-trace methods of measuring delays.

In Figure 2, cumulative arrival and departure patterns and resulting queues during an oversaturation period are shown. The specific queueing model used in this paper is depicted in Figures 3, 4a, and 4b. The dual flow rate approach used in this model is consistent with the PHF concept in the HCM. The model differs from the so-called low-definition approach used in the United Kingdom (*5*) in dividing the total flow period into peak and nonpeak periods with constant flow rates rather than using a constant average flow rate throughout the total flow period. It also differs from the low definition approach by assuming that the peak period capacity applies throughout the oversaturation period (i.e., the nonpeak capacity applies only after the oversaturation queues have cleared).

Let us define a variable (floating) delay period that starts at time y after the start of the peak flow period, terminates before the end of the oversaturation period ($0 \le y \le T_o - T_p$), and is of the same duration as the peak flow period (T_p). For $y = 0$, the delay period is identical to the peak flow period, and $y = y_m$ defines a maximum delay period. The delay and queue statistics for various combinations of delay period definition and delay measurement method are given as follows.

Delay and Queue Statistics Using the Queue Sampling Method of Measuring Delay

The total oversaturation delay measured by the queue sampling method as incurred in a floating delay period of duration T_p starting at time y after the onset of the peak flow period (see Figure 4a) is given by

$$D_y = 0.5c_p[(x_p - 1)(T_p^2 + 2T_p y - y^2) - y^2(1 - \alpha x_p)] \tag{11}$$

The number of vehicles experiencing the total delay given by Equation 11 is

$$N_y = q_p(T_p - y) + q_n y = q_p[T_p - y(1 - \alpha)] \tag{12}$$

Therefore, the average delay corresponding to Equation 11 is

$$d_y = D_y/N_y \tag{13}$$

The start and end overflow queue lengths for the delay period are

$$N_{sy} = c_p y(x_p - 1) \tag{14}$$

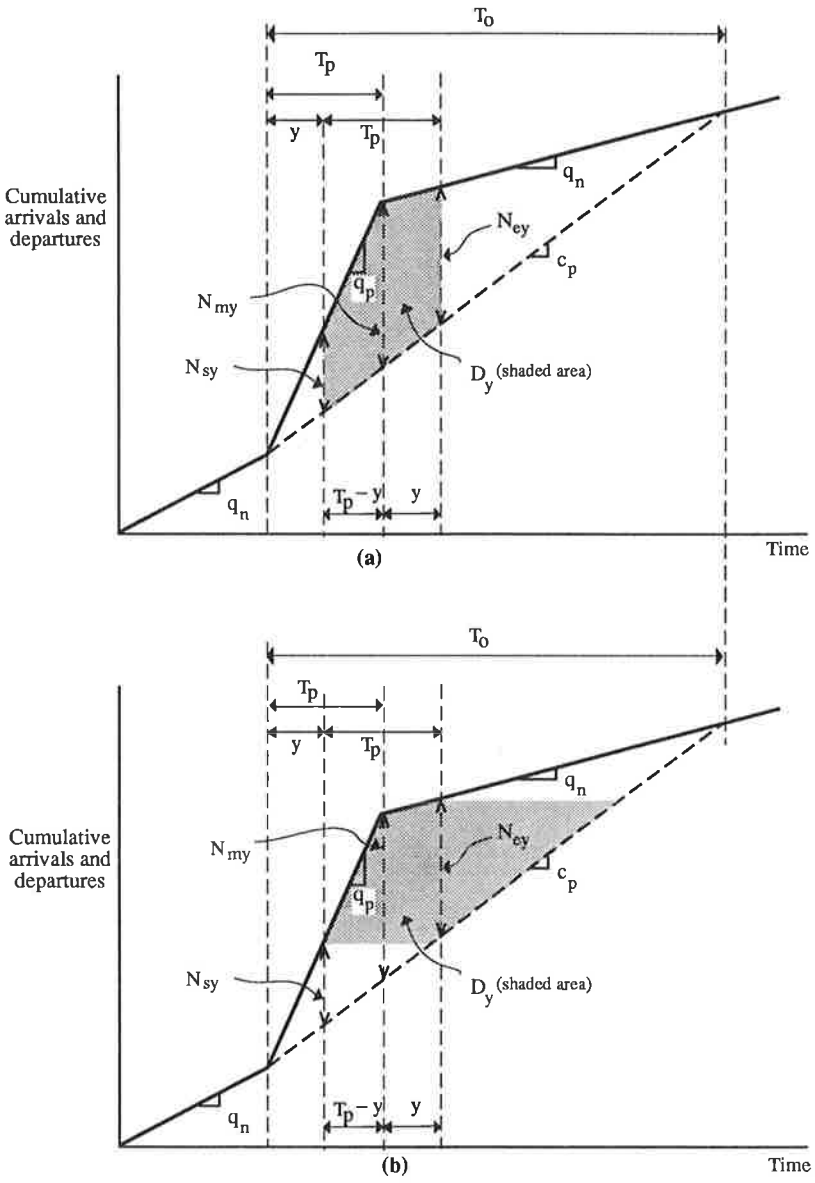


FIGURE 4 Oversaturation models with (*a*) queue sampling and (*b*) path-trace methods of delay measurement.

$$N_{ey} = c_p[T_p(x_p - 1) - y(1 - \alpha x_p)] \quad (15)$$

The average overflow queue length for the delay period of duration T_p is

$$N_{ay} = D_y/T_p \quad (16)$$

Application of the general equations for the queue sampling method of delay measurement to the maximum total delay and peak flow periods is given in the following subsections.

Delay and Queue Statistics for the Maximum Delay Period

The value of y that gives the maximum value of the total delay from Equation 11, y_m, can be determined by setting the derivative of the total delay with respect to y to zero:

$$y_m = \frac{T_p(x_p - 1)}{x_p(1 - \alpha)} \quad (17)$$

Note that Equation 17 always satisfies $y_m \leq T_o - T_p$ (i.e., the maximum delay period ends before the end of the oversaturation period). From Equations 11 and 17, the maximum total delay is

$$D_m = 0.5T_p^2c_p(x_p - 1)\left[1 + \frac{(x_p - 1)}{x_p(1 - \alpha)}\right] \quad (18)$$

The number of vehicles experiencing the maximum total delay given by Equation 18 is

$$N_m = c_pT_p \quad (19)$$

The corresponding value of the average delay is

$$d_m = 0.5T_p(x_p - 1)\left[1 + \frac{(x_p - 1)}{x_p(1 - \alpha)}\right] \quad (20)$$

The equation given by Messer (*4*) corresponds to Equation 20. Thus, Messer's formula gives the maximum delay for the queue sampling method of measurement. It should be noted that there is an inconsistency in the definition of the delay factor (k) used by Messer (bracketed term in Equation 20) since he calculated it as the ratio of the maximum delay with the queue sampling method (Equation 20) to the delay to individual vehicles arriving during the peak flow period, which implies the path-trace method of delay measurement (Equation 34). Furthermore, Equation 20 does not necessarily give the maximum value of the average delay experienced by individual vehicles since it is based on maximum total delay.

From Equations 14, 15, and 17, the start and end overflow queue lengths in the maximum delay period can be shown to be equal and have the value

$$N_{sm} = N_{em} = \frac{T_p c_p (x_p - 1)^2}{x_p(1 - \alpha)} \quad (21)$$

From Equations 16 and 17, the average overflow queue length in the maximum delay period is

$$N_{am} = 0.5 T_p c_p (x_p - 1) \left[1 + \frac{(x_p - 1)}{x_p(1 - \alpha)} \right] \quad (22)$$

Delay and Queue Statistics for the Peak Flow Period

Delay and queue statistics for the peak flow period with the queue sampling method of measurement can be derived by setting $y = 0$ in Equations 11 to 16. The nonpeak flow rate is not relevant to estimating queues and delays in this case (see Figure 4a). Therefore, the total delay in the peak flow period is

$$D_o = 0.5 T_p^2 c_p (x_p - 1) \quad (23)$$

The number of vehicles experiencing the total delay given by Equation 17 is

$$N_o = q_p T_p \quad (24)$$

The average delay measured in the peak flow period is

$$d = D_o / N_o = \frac{0.5 T_p (x_p - 1)}{x_p} \quad (25)$$

The start and end overflow queue lengths in the peak flow period are

$$N_{so} = 0 \quad (26)$$

$$N_{eo} = T_p c_p (x_p - 1) \quad (27)$$

and the average overflow queue length in the peak flow period is

$$N_{ao} = 0.5 T_p c_p (x_p - 1) \quad (28)$$

Delay and Queue Statistics Using the Path-Trace Method of Measuring Delay

The path-trace method measures delays experienced by individual vehicles, which is more relevant to the LOS concept than the queue sampling method, which relates to a system concept. Therefore, this method considers delays to vehicles arriving in that period regardless of departure times (see Figures 2b and 4b).

The total oversaturation delay measured by the path-trace method that is incurred in a floating delay period of duration T_p starting at time y after the onset of the peak flow period is given by

$$D_y = 0.5 c_p x_p \{(x_p - 1)(T_p^2 - y^2) + \alpha y [2 T_p (x_p - 1) - y(1 - \alpha x_p)]\} \quad (29)$$

The average delay corresponding to Equation 29 can be calculated from the general relationship described by Equation 13. This includes individual vehicle delays experienced beyond the delay period (i.e., after time $y + T_p$). However, all queue statistics are equivalent to those derived for the queue sampling method (Equations 14 to 16).

Application of the general equations for the path-trace method of delay measurement to the maximum average delay and peak flow periods is given in the following subsections.

Delay and Queue Statistics for the Maximum Delay Period

For LOS assessment purposes, the period maximizing the average delay rather than the total delay should be used. The value of y that gives the maximum value of the average delay from Equations 13 and 29, y_m, can be determined by setting the derivative of the average delay with respect to y to zero:

$$y_m = \frac{T_p}{1 - \alpha} \left[1 - \sqrt{1 - \frac{(x_p - 1)(1 - \alpha^2)}{\alpha (1 - \alpha x_p) + (x_p - 1)}} \right] \quad (30)$$

subject to $y_m \leq T_o - T_p$. The upper bound on y_m ensures that the maximum delay period ends before the end of the oversaturation period.

The maximum value of the average delay is obtained from $d_m = D_m / N_m$ using the total delay, D_y, from Equation 29 and the number of vehicles experiencing that total delay, N_y, from Equation 12 with $y = y_m$:

$$d_m = 0.5 \frac{(x_p - 1)(T_p^2 - y_m^2) + \alpha y_m [2 T_p (x_p - 1) - y_m (1 - \alpha x_p)]}{T_p - y_m (1 - \alpha)} \quad (31)$$

The start, end, and average overflow queue lengths for the delay period maximizing the average delay can be obtained by substituting $y = y_m$ in Equations 14 to 16.

Delay and Queue Statistics for the Peak Flow Period

Delay and queue statistics for the peak flow period with the path-trace method of measurement can be derived by setting

$y = 0$ in Equation 29. The resulting equations are equivalent to the equations given by Akçelik (*6*). In this case, the nonpeak flow rate is not relevant to estimating queues and delays to vehicles arriving in the peak flow period (see Figure 2b). Therefore, the total delay in the peak flow period is

$$D_o = 0.5T_p^2 c_p x_p (x_p - 1) \quad (32)$$

The number of vehicles experiencing the total delay given by Equation 22 is

$$N_o = q_p T_p \quad (33)$$

Therefore, the average delay experienced by vehicles arriving in the peak period is

$$d_o = 0.5T_p(x_p - 1) \quad (34)$$

The start, end, and average overflow queue lengths in the peak flow period (N_{so}, N_{eo}, and N_{ao}) are identical to those obtained by the queue sampling method (see Equations 26 to 28).

NUMERICAL EXAMPLE

A numerical example illustrating the use of the equations given in the preceding section for estimating queue and delay statistics for the peak flow and maximum delay (total or average) periods is given in this section. In particular, the following points are explored:

- Effect of flow profile aggregation (i.e., the choice of the duration of peak flow period) on delay and queue estimates,
- Effect of the delay measurement method (queue sampling or path trace) on predicted oversaturation delay, and
- The relationship between the average oversaturation delay incurred in the peak flow period and that incurred in the maximum delay period given the method of delay measurement.

In this example, demand flow data are assumed to be collected in eight 15-min intervals within a 2-hr peak ($T = 2$ hr). The observed flow profile is shown in Figure 5a. The demand profiles synthesized according to selected durations of the peak flow period (T_p) are shown in Figure 5b.

The example relates to a traffic stream in a signalized intersection approach lane. As a rough way of emulating the

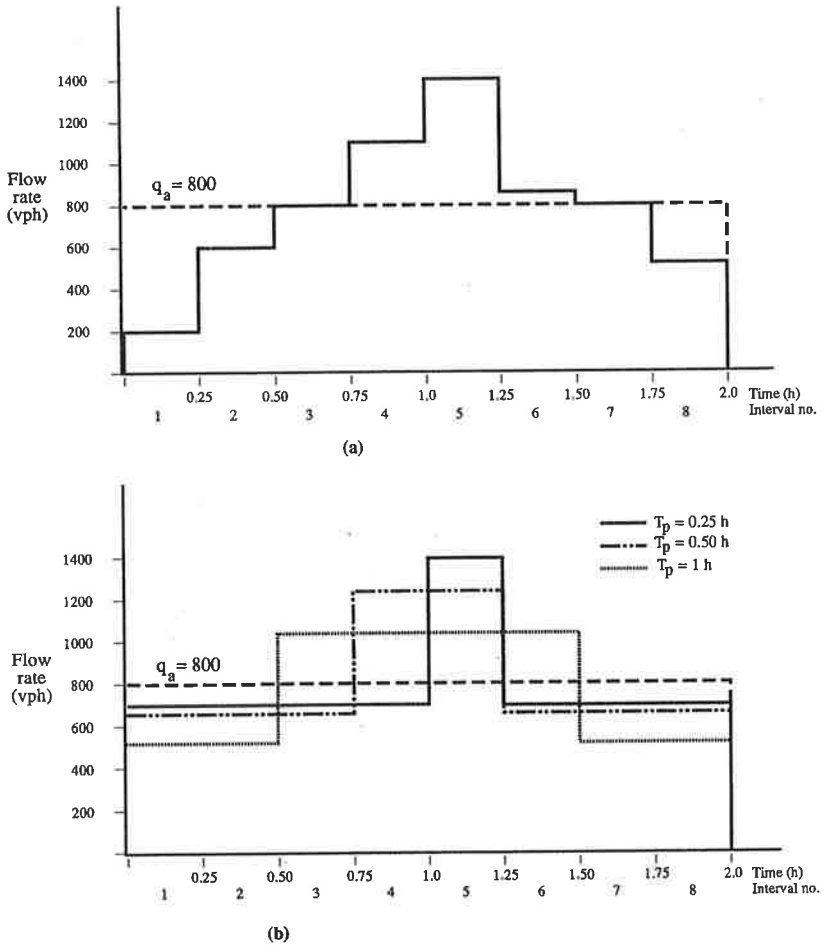


FIGURE 5 Numerical example ($T = 2$ hr): (*a*) actual demand profile; (*b*) synthesized demand profiles for indicated T_p.

operation of a vehicle-actuated signal controller, the capacity function is set at

$$c = \min \{c_m, q/x_d\} \quad (35)$$

where

c = capacity (veh/hr) for a given demand level q,
c_m = maximum movement capacity attained during the oversaturation period, that is, as long as overflow queues exist during and after the peak flow period (the maximum is a result of limitations on maximum green time, cycle length, etc.), and
x_d = design (or practical) degree of saturation at demand level q.

In this example, c_m = 1,000 veh/hr and x_d = 0.90 are used. The average flow rate over the 2-hr period is set at q_a = 800 veh/hr.

The capacity model given in Equation 35 yields higher capacities with increasing flow levels up to a maximum value. In some cases, the signalized intersection capacities may decrease as flow levels increase [e.g., where such factors as opposed turns, short lanes, and shared lane blockages are dominant (*8*)]. At roundabouts and other signalized intersections, capacities always decrease with increasing flow levels because of the underlying gap acceptance process. The results from the model based on Equation 35 should therefore not be generalized.

The results are summarized in Tables 1 to 5 as explained below. In all tables, flow rates are given in vehicles per hour, times in hours, queue lengths in vehicles, total delays in vehicle-hours, and average delays in seconds per vehicle.

TABLE 1 Flow and Capacity Parameters for Selected Peak Flow Period Lengths (c_p = 1,000 veh/hr)

Selected T_p	PTF (Eqn 4b)	Average q_p	PFF (Eqn 4a)	α (Eqn 6b)	q_n (Eqn 6a)	x_p	αx_p	T_o (Eqn 7)
0.25	0.1250	1400	0.571	0.510	714	1.400	0.714	0.600
0.50	0.250	1250	0.640	0.520	650	1.250	0.650	0.857
0.75	0.375	1133	0.705	0.528	598	1.133	0.598	0.998
1.00	0.500	1050	0.762	0.524	550	1.050	0.550	1.111

TABLE 2 Delay and Queue Statistics Using the Queue Sampling Method: Maximum Delay Period

T_p	x_p	αx_p	y_m	D_m	d_m	N_{sm}	N_{em}	N_{am}
0.25	1.400	0.714	0.146	19.79	284.9	58.3	58.3	79.2
0.50	1.250	0.650	0.208	44.27	318.8	52.1	52.1	88.6
0.75	1.133	0.598	0.187	46.85	224.9	24.9	24.9	62.5
1.00	1.050	0.550	0.100	27.50	99.0	5.0	5.0	27.5

TABLE 3 Delay and Queue Statistics Using the Queue Sampling Method: Peak Flow Period

T_p	x_p	αx_p	y	D_o	d_o	N_{so}	N_{eo}	N_{ao}
0.25	1.400	0.714	0.0	12.50	128.6	0.0	100.0	50.0
0.50	1.250	0.650	0.0	31.25	180.0	0.0	125.0	62.5
0.75	1.133	0.598	0.0	37.41	158.8	0.0	100.0	50.0
1.00	1.050	0.550	0.0	25.00	85.7	0.0	50.0	25.0

TABLE 4 Delay and Queue Statistics Using the Path-Trace Method: Maximum Delay Period

T_p	x_p	αx_p	y_m	D_m	d_m	N_{sm}	N_{em}	N_{am}
0.25	1.400	0.714	0.165	18.88	286.9	66.0	52.8	78.7
0.50	1.250	0.650	0.250	42.50	322.0	62.5	37.5	87.5
0.75	1.133	0.598	0.239	45.48	226.8	31.8	3.8	61.3
1.00	1.050	0.550	0.111	27.45	99.4	5.6	0.0	27.5

TABLE 5 Delay and Queue Statistics Using the Path-Trace Method: Peak Flow Period

T_p	x_p	αx_p	y	D_o	d_o	N_{so}	N_{eo}	N_{ao}
0.25	1.400	0.714	0.0	17.50	180.0	0.0	100.0	50.0
0.50	1.250	0.650	0.0	39.06	225.0	0.0	125.0	62.5
0.75	1.133	0.598	0.0	42.50	180.0	0.0	100.0	50.0
1.00	1.050	0.550	0.0	26.30	90.0	0.0	50.0	25.0

1. Table 1 gives the demand and capacity flow parameters in the peak and nonpeak periods. Capacity for the peak flow period is derived using the peak flow rate (q_p) in Equation 35. Since all peak flows are above 1,000 veh/hr, the peak period capacity is the same for all T_p cases ($c_p = c_m$ = 1,000 veh/hr). Table 1 also gives the oversaturation period (T_o) during which the peak period capacity (c_p) applies. The nonpeak period capacities (c_n) are not given in Table 1 since they are not used in the deterministic oversaturation model (Figure 4).
2. Tables 2 and 3 provide comprehensive delay and queue statistics for the maximum delay and peak flow periods obtained using the queue sampling method.
3. Finally, Tables 4 and 5 give comparable delay and queue statistics corresponding to the path-trace method.

A study of the results given in Tables 1 to 5 indicates a strong correlation between the duration of the selected peak interval and the corresponding delay and queue statistics. This is explained by the fact that the peak and nonpeak flow rates are averages within and outside the peak period. A peak flow period with longer duration (larger T_p) implies a longer peak, but a smaller peak flow rate. This trade-off is evident when comparing the average delay and queue length values within each table for different T_p values. In most cases, a 30-min interval yielded the highest delay. Thus, the blanket use of a fixed peak and analysis periods as in the HCM (T_p = 0.25 hr and T = 1 hr) is not supported by this example. In fact, such blanket definitions may not be consistent with the intended use of the delay models, which are meant to simulate the performance of a (possibly) saturated peak within an undersaturated total flow period.

The differences in queue and delay statistics obtained from the queue sampling and path-trace methods of delay measurement can be seen by comparing the results in Tables 2 and 4 for maximum delays or Tables 3 and 5 for average delays in the peak flow period. Maximum delays estimated by the two methods are similar. The path-trace method gives higher delay for the case of analysis for the peak flow period, which is due to the allowance for oversaturation delays experienced after the peak flow period.

As expected, substantial differences in delay and queue statistics were observed in the cases of peak flow and maximum delay periods regardless of the delay measurement method

(comparing the results in Tables 2 and 3 or those in Tables 4 and 5). The level of difference is also seen to be affected by the choice of the duration of the peak flow period.

CONCLUSION

This paper has presented a deterministic oversaturation queueing model, which generalizes the peak hour factor concept of the U.S. HCM (*1*). Using this simple variable demand model, several issues related to oversaturation models have been explored. In particular, consistency of delay definitions and delay measurement methods has been investigated. A numerical example has been used to illustrate the application of the model. For the discussion of a full time-dependent model allowing for both random and oversaturation delays, the user is referred to Akçelik and Rouphail (*7*).

The queue and delay estimates are highly sensitive to the selected peak flow period duration irrespective of the delay definition or the delay measurement method. In fact, variations caused by the choice of the peak flow period duration are as significant as those resulting from the use of a different delay definition or delay measurement method.

For the example analyzed, the ability to vary the duration of the peak flow period revealed that a 30-min peak period was more critical in terms of resulting delays and queues than the 15-min peak duration specified in the HCM.

As expected, substantial differences in delay and queue estimates are observed between the cases of peak flow and maximum delay periods regardless of the delay measurement method. In the numerical example, the maximum delays estimated by the queue sampling and path-trace methods are similar, but the path-trace method produces higher delays for the peak flow period.

The use of the average delay experienced by individual vehicles in a maximum delay period appears to have some merit in terms of LOS. However, this creates several problems in system performance analysis (including estimation of operating cost, fuel consumption, and pollutant emission).

First, the number of vehicles experiencing this delay is smaller than the number of vehicles arriving in the peak flow period. By applying this delay to the peak flow period, the total delay would be overestimated. Therefore, the use of this delay should be restricted to LOS assessment purposes only. For the purpose of system performance design and evaluation, total oversaturation delay should be used.

Second, a delay definition based on a maximum delay period reveals an inconsistency in relation to delays measured in the field. Simply stated, whereas the HCM recommends that field delays be measured in the peak flow period, the maximum delay period does not coincide with the peak flow period. Thus, the two delays are not comparable.

Furthermore, if the use of maximum delay is adopted, the path-trace rather than the queue sampling method should be used, since the former is more relevant to LOS assessment in reflecting the delays experienced by individual vehicles.

It is therefore important that the delay definition implied by the present HCM delay formula for signalized intersections be clarified in view of the comments presented in this paper. Specifically, if the x^2 factor in the incremental delay term of the HCM delay formula is intended to produce a maximum delay estimate for oversaturated conditions as put forward by Messer (*4*), the delay estimates from the HCM delay formula should not be expected to correspond to delays measured in the 15-min peak flow period by the queue sampling method specified in the HCM.

Thus, it would be advisable to consider two distinct delay models for system performance and LOS assessment purposes. In the former, delays incurred throughout the oversaturation period would be considered for estimating total delay, operating cost, fuel consumption, and pollutant emissions. The latter should strictly apply to the maximum delay period using the path-trace method for LOS assessment.

The differences in delay definitions and delay measurement methods that have been emphasized in this paper are relevant to oversaturated conditions only. For undersaturated conditions represented by low to medium v/c ratios, the effect of the time-dependence of demand flows (i.e., the duration of the peak flow period) on delays and queues is negligible, and therefore the delay definitions used in the delay formulas have little effect on delay estimates. Similarly, the queue sampling and path-trace methods of delay measurement should yield similar delays under low to medium v/c ratios.

However, as flows approach capacity (undersaturated but high v/c ratios near capacity) and exceed capacity (v/c ratio greater than 1), the selection of the duration of the flow period, delay definition, and delay measurement method affect delay estimates significantly. Because of the dual nature of the delay-flow functions, the use of a factor that applies to all flow conditions [such as the x^2 factor in the incremental (random plus oversaturation) term of the HCM delay equation or the progression factor that multiplies both terms of the equation] is not appropriate for modeling oversaturation effects (*2,9*). A time-dependent continuum model satisfying these requirements is described in a follow-up paper (*7*).

It is realized that the dual flow model presented in this paper is still a simplification of the variable demand model case. Nevertheless, the concept builds on flow data that are gathered routinely as part of intersection operational analysis studies. The end user must realize, however, that some prior investigation is needed to select sensible durations for the peak and total flow periods. As a guide, a 15-min peak flow period appears to be the smallest aggregation period for which volumes can be assumed uniform. The total flow period is more difficult to ascertain except that it is advisable that no overflow queues should be present at either its start or its termination. Determining the critical duration of the peak period interval (in multiples of 15-min) should take into account the level of peaking. Short peak flow periods are required in high peaking cases (low PHF or PFF values) to allow for long oversaturation periods resulting from a high v/c ratio in the peak flow period. In the numerical example, a 30-min peak flow period produced the largest delay and queue estimates. Further work is required on the effect of the choice of the location and duration of the peak flow period in terms of the total system performance in the total flow period.

The practical difficulty of measuring the true demand profile, which requires measuring arrival flows at the end of the queue, should also be recognized. Volume counts at the stop line cannot identify oversaturation since the stop line flows can never exceed the capacity (in the example shown in Figure

5, Intervals 4 and 5 have demands that exceed the capacity of 1,000 veh/hr, but the stop line counts would yield an apparent demand of 1,000 veh/hr). On the other hand, the stop line method would count the excess demand in subsequent intervals. This would indicate less peaking (a higher PHF or PFF value) than the real demand profile. Stop line volume counts supplemented by queue counts (*10*) could be used to estimate the true demand for oversaturated conditions.

ACKNOWLEDGMENTS

The authors thank Ian Johnston of the Australian Road Research Board for permission to publish this paper. The work reported in the paper was carried out during Nagui Rouphail's sabbatical leave at ARRB.

REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
2. R. Akçelik. The Highway Capacity Manual Delay Formula for Signalized Intersections. *ITE Journal*, Vol. 58, No. 3, 1988, pp. 23–27.
3. W. R. McShane and R. P. Roess. *Traffic Engineering*. Prentice Hall, Englewood Cliffs, N.J., 1990.
4. C. Messer. The 1985 Highway Capacity Manual Delay Equation. Compendium of Technical Papers, 60th Annual Meeting of the Institute of Transportation Engineers, Orlando, Fla., 1990, pp. 205–209.
5. R. M. Kimber and E. M. Hollis. *Traffic Queues and Delays at Road Junctions*. TRRL Laboratory Report 909, Berkshire, England, 1979.
6. R. Akçelik. *Time-Dependent Expressions for Delay, Stop Rate and Queue Length at Traffic Signals*. Internal Report AIR 367-1. Australian Road Research Board, 1980.
7. R. Akçelik and N. M. Rouphail. *Estimation of Delays at Traffic Signals for Variable Demand Conditions*. Working Paper WD TE91/005, Australian Road Research Board, 1991.
8. R. Akçelik. SIDRA for the Highway Capacity Manual. Compendium of Technical Papers, 60th Annual Meeting of the Institute of Transportation Engineers, Orlando, Fla., 1990, pp. 210–219.
9. N. M. Rouphail. Cycle-by-Cycle Analysis of Congested Flow at Signalized Intersections. *ITE Journal*, Vol. 61, No. 3, 1991, pp. 33–36.
10. D. S. Berry. Volume Counting for Computing Delays at Signalized Intersections. *ITE Journal*, Vol. 57, No. 3, 1987, pp. 21–23.

The views expressed in the article are those of the authors and not necessarily those of ARRB.

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

Car-Following Model Based on Fuzzy Inference System

SHINYA KIKUCHI AND PARTHA CHAKROBORTY

Car-following theory has been receiving renewed attention for its use in the analysis of traffic flow characteristics and vehicle separation control under the IVHS. A car-following model that uses the fuzzy inference system, which consists of many straightforward natural language–based driving rules, is proposed. It predicts the reaction of the driver of the following vehicle (acceleration-deceleration rates) given the action of the leading vehicle. A range of possible reaction is predicted and expressed by the fuzzy membership function. The model is applied to the analysis of traffic stability and speed-density relationship. For traffic stability, the results are compared with those derived from the deterministic approach. The speed-density relationship derived from the model is compared with a set of actual flow data. The predicted range is found to be reasonable. The proposed fuzzy approach helps explain the scatter of the actual data as possibility rather than random variation.

For the past several decades traffic flow has been generally analyzed under the premise that all drivers behave in a similar manner and that a general law exists governing the flow characteristics in the traffic stream. On the basis of this premise, characteristics of flow have been analyzed from both the microscopic and the macroscopic standpoints. Most studies have considered that a deterministic relationship exists between the action of a vehicle and the reaction of the vehicles that follow. Whereas the existence of this cause and effect relationship is not disputable, the reactions of a driver to the actions of other drivers are perhaps not based on a deterministic one-to-one relationship, but on a set of vague driving rules developed through experience. The way in which the rules are applied may differ with different drivers, and even for the same driver, it differs with different conditions. The rules are not rigid but are natural language based. For example, if the leading vehicle (LV) decelerates, then the following vehicle (FV) should decelerate; or, if the distance between the LV and FV becomes very short, the FV should decelerate and try to increase the distance. Such a linguistic reasoning pattern is suited for an analysis using fuzzy logic and approximate reasoning techniques. Fuzzy set theory and logic allows the mathematical treatment of subjective judgment and inference, and in recent years fuzzy logic has been applied to many practical problems involving controls and decisions under the environment of the imprecise human reasoning process.

This paper proposes a fuzzy rule–based car-following model that assumes that a decision made by a driver is the result of a fuzzy reasoning process and then predicts the possibilities of the reaction of the FV.

Civil Engineering Department, University of Delaware, Newark, Del. 19716.

CAR-FOLLOWING MODELS: TRADITIONAL APPROACH

This section is divided into two subsections. The first subsection describes the car-following models developed by the General Motors research group (GM Model) and their assumptions. The second subsection discusses traffic stability and speed-density relationships in the car-following context.

The car-following theory evolved in the 1950s. Among the researchers who pioneered in the field, Pipes (*1*, pp. 164–166) developed a microscopic model that assumed that the minimum safe distance between vehicles was a function of speed. His work was followed by that of Forbes (*1*, pp. 116–167). While Pipes modeled the traffic flow assuming that drivers maintain a constant distance headway, Forbes assumed that drivers maintain a constant time headway. However, by far the largest contribution was made by the GM's research team (*2–5*). Some of the GM models are discussed here.

Models and Assumptions

The GM models were based on the premise that the reaction of the FV at time t depends on the sensitivity of the FV and the strength of the stimulus given by the LV at time $t - \Delta t$, where the strength of the stimulus is measured in terms of the relative velocity between the LV and the FV, the reaction of the FV is measured by the acceleration or deceleration rate, the time difference, Δt, is equal to the perception/reaction time, and the sensitivity term maps the unit of a stimulus to a reaction. The GM team developed five models that have the same general structure but differ from one another in the sensitivity term. The fifth model is a generalized representation of the first four models:

$$\ddot{x}_{n+1}(t + \Delta t) = \frac{\alpha_{t,m}(\dot{x}_{n+1}(t + \Delta t))^m}{[x_n(t) - x_{n+1}(t)]^\ell} \cdot [\dot{x}_n(t) - \dot{x}_{n+1}(t)] \qquad (1)$$

where

$\ddot{x}_{n+1}(t + \Delta t)$ = acceleration or deceleration rate of $(n + 1)$th car at time $t + \Delta t$,
$\dot{x}_n(t)$ = speed of nth car at time t,
$x_n(t)$ = position of nth car at time t,
ℓ = parameter for sensitivity to distance $x_n(t) - x_{n+1}(t)$,

m = parameter for sensitivity to speed $\dot{x}_{n+1}(t + \Delta t)$, and
$\alpha_{\ell,m}$ = constant.

This model has the following characteristics:

1. The interaction between stimulus and reaction has a one-to-one correspondence. The notion that a driver's reaction pattern is imprecise is not fully represented. Ceder (*6*) expressed a similar concern. Representation of a human behavioral pattern may be better explained by an approximate reasoning process than a deterministic equational model.
2. The FV reacts even to minute changes in relative velocity between the LV and FV in a deterministic manner.
3. Sensitivities of the FV to the positive and negative relative velocities are the same. Equation 1 suggests that if the FV accelerates at y ft/sec^2 when the relative speed is β ft/sec, then it decelerates at $-y$ ft/sec^2 when the relative speed is $-\beta$ ft/sec. It has been observed that drivers react differently when the distance between cars is increasing or decreasing. Leutzbach (*7*) also states that "drivers pay closer attention to spacing decreases (decrements) than to spacing increases (increments) simply on the basis of their own safety."

Applications of GM Model

This section discusses two topics to which the car-following models have been applied: traffic stability and macroscopic speed-density relationships.

Traffic Stability

Traffic stability is a study of how stability is restored in the traffic flow after the leader of a platoon "destabilizes" the flow by accelerating or decelerating. The traffic stability analyses have focused on how the vehicle spacing changes with time. Two types of stability patterns have been studied in the past: local stability and asymptotic stability. Extensive analyses of local stability patterns were conducted for different car-following models (i.e., different combinations of m and ℓ in Equation 1) by Herman et al. (*5*), Chandler at al. (*2*), and Herman and Potts (*8*). Herman and Potts (*8*) present the results from three different cases: (*a*) $m = \ell = 0$, (*b*) $m = \ell = 0$ but with two values of α, and (*c*) $m = 0$ and $\ell = 1$. Only the first case ($m = \ell = 0$) has been analyzed mathematically.

While these analyses provide insight into what happens in reality, each has its own limitations. For example, the constant sensitivity case (m $= \ell = 0$) implies that the reaction to a given relative velocity is independent of the distance between LV and FV. Though an improvement over the previous one, even the reciprocal spacing ($m = 0, \ell = 1$) model has shortcomings; one of them is that no difference between the stimuli of positive and negative relative velocity is made in the sensitivity term. Another drawback of this model is that the FV's reaction is independent of the velocity of the FV. It can be argued that as velocity increases the reaction to positive relative velocity is subdued and negative relative velocity enhanced. The model represented in Equation 1, though still deterministic in nature, is the closest to reality. However, with nonzero coefficients of m and ℓ, the difference differential equation of Equation 1 becomes difficult to solve.

Asymptotic stability is concerned with how the instability introduced by the LV propagates down a line of traffic. This is an interesting topic in the sense that it may explain certain causes of accidents and congestion. Herman et al. (*5*) and Herman and Potts (*8*) have also presented results from their study on asymptotic stability, and these remain the most extensive study on this topic.

Speed-Density Relationship (u-k Relationship)

The bridge built by Gazis et al. (*4*) between the microscopic car-following model with $m = 0$ and $\ell = 1$ and Greenberg's macroscopic speed-density relationship was a significant step toward unifying the microscopic and macroscopic approaches. This effort has made it possible to show that other macroscopic speed-density models can also be derived from different assumed values of m and ℓ in Equation 1: $m = 0$ and $\ell = 2$ for Greenshields; $m = 1$ and $\ell = 2$ for Underwood; and $m = 1$ and $\ell = 3$ for Northwestern's (*1*, p. 304).

The relationship between the microscopic and macroscopic models allows the examination of the validity of the microscopic model by the observed *u-k* relationship. The facts that the observed data points in the *u-k* relationship are scattered and the observed and predicted characteristics have significant discrepancies suggest that a problem might lie in the deterministic approach.

This has been pointed out by some. For example, Ross (*9*) states, "The idea that there is deterministic relationship between speed and density, be it straight line or curve, is simply untenable. The most obvious problem is that speed-density observations always have much more scatter than can be explained by any reasonable amount of experimental error." This concern was echoed by Gilchrist and Hall (*10*): "The scatter in the traffic data is sufficient to cast doubt on the narrow linear representation of any relationship between traffic flow variables." Underwood (*11*) developed probability distributions for speed for different volumes.

These comments, combined with the belief that drivers do not behave in a rigid deterministic manner, lead us to consider a model based on a fuzzy inference system.

FUZZY RULE–BASED MODEL FOR THE CAR-FOLLOWING PROBLEM: RATIONALE

In the car-following situation, one follows a set of driving rules built over time through experience. Examples of the rules that the FV might apply are as follows: (*a*) accelerate if the LV accelerates, and (*b*) decelerate and keep longer distance if the LV decelerates and the distance between cars is short.

Each rule is built on natural language, and no exact boundary for the applicability of the rule is defined. Hence, many of the rules may be applied (or "fired") simultaneously in the mind of the driver, and the driver may not be completely certain of the appropriateness of his action. The probability approach, which has traditionally been used to analyze un-

certainty, however, cannot deal with linguistic variables such as "fast" and "slow"; further, it must follow a rigid set of rules defining the properties of the probability function.

If we postulate that a driver's reaction is one of several possible actions available, the variation of the reaction pattern and the scatter of the observed u-k relationship may be explained. A fuzzy set, which will be explained later, is actually the set of elements with the possibility of being in the set of discourse (*12*). In recent years, fuzzy sets have been used to represent the approximate reasoning and decision process. This approach may offer an alternative explanation of the car-following phenomenon.

ELEMENTS OF FUZZY SET THEORY

This section presents elements of fuzzy set theory that are relevant to the construction of the proposed model. More detailed explanation of fuzzy theory can be found elsewhere (*13–15*).

Fuzzy Sets

A fuzzy set is a set for which the criterion for belonging to the set is not dichotomous. The membership of the set is defined by a grade (or degree of compatibility or degree of truth) whose value is between 0 and 1. A membership function determines the grade and is defined as

$$h_A(x): X \rightarrow [0,1] \tag{2}$$

where A is a fuzzy set defined on the universal set X.

The notion of "high speed" or "low speed," for example, can be represented by fuzzy sets whose membership functions define the perception of high or low in terms of numerical value of speed. Similarly, an approximate integer constitutes a fuzzy set that is normal and convex. "Approximately 5" may have the following membership function:

$$\text{"Approximately 5"} = 3/0.4 + 4/0.8 + 5/1.0 + 6/0.6 + 7/0.4$$

Arithmetic operations on fuzzy numbers are defined using the extension principle. For a detailed description of fuzzy arithmetic, readers are referred to Dubois and Prade (*16*) and Kaufmann and Gupta (*17*).

Operations of Fuzzy Sets

Among the set operations relevant to the subsequent discussions are union, intersection, and complement, defined by Equations 3, 4, and 5, respectively.

$$h_{A\cup B}(x) = h_A(x) \vee h_B(x) \tag{3}$$

$$h_{A\cap B}(x) = h_A(x) \wedge h_B(x) \tag{4}$$

$$h_{\bar{A}}(x) = 1 - h_A(x) \tag{5}$$

In these equations, $\wedge$ indicates the minimum and $\vee$ the maximum of the operands [$h_A(x)$ and $h_B(x)$, in this case].

Fuzzy Inference

Under fuzzy logic, the inference process includes fuzzy input and a fuzzy relationship, as follows:

Input: x is somewhat A $\quad (x = A')$

Rule: if x is A then y is B $\quad (\text{R}{:}x = A \rightarrow y = B)$

Conclusion: y is somewhat B $\quad (y = B')$ (6)

where all or some of A, A', B, and B' are fuzzy sets, and the rule represents a fuzzy cause-and-effect relation between x and y. The first part of the rule, "x is A," is called the premise, and the second, "y is B," is called the consequence. The validity of the consequence depends on the compatibility between the input and the premise of the rule. In other words, the degree to which "y is B" is true is dictated by the degree of match between "x is somewhat A" and "x is A."

A fuzzy inference system can be composed of more than one rule with each rule consisting of more than one premise variable, as follows:

Input: $x_1 = A'$ and $x_2 = B'$

Rule 1: If $x_1 = A_1$ and $x_2 = B_1$, then $y = C_1$

Rule 2: If $x_1 = A_2$ and $x_2 = B_2$, then $y = C_2$

. . . . . .

Rule i: If $x_1 = A_i$ and $x_2 = B_i$, then $y = C_i$

Conclusion: $y = C'$ (7)

The compatibility between the input and the premise of a rule i, W_i, is examined as follows:

$$W_i = \{\vee_{x_1} [h_{A'}(x_1) \wedge h_{Ai}(x_1)]\} \wedge \{\vee_{x_2} [h_{B'}(x_2) \wedge h_{Bi}(x_2)]\} \tag{8}$$

When n different rules are applied (or "fired") for the given input, the degree of compatibility between the input and the premise is computed for each rule, and then the conclusion is the average of the individual consequences, C_i's, weighted by W_i's:

$$C' = \frac{\sum W_i \cdot C_i}{\sum W_i} \tag{9}$$

where C' is still a fuzzy number.

This operation is, in fact, an interpolation of C_i's. This method is an extension of the one proposed by Takagi and Sugeno (*18*) where their consequence C_i's are crisp numbers.

Expression 9 is computed as follows:

1. Normalize the values of W_i:

$$\eta_i = \frac{W_i}{\sum W_i} \tag{10}$$

2. Multiply fuzzy number C_i by η_i to obtain a new fuzzy number D_i:

$$h_{Di}(\lambda) = \max_{\lambda = y \times \eta_i} h_{Ci}(y) \tag{11}$$

3. Add Di's for all i's for which the rules apply:

$$h_{C'}(y) = \max_{y = \lambda_1 + \lambda_2 + \ldots + \lambda_N} \{h_{D1}(\lambda_1), h_{D2}(\lambda_2), \ldots, h_{DN}(y - \lambda_1 - \lambda_2 - \ldots - \lambda_{N-1})\} \tag{12}$$

where $h_{C'}(y)$ is the membership function of the conclusion.

A comprehensive discussion of fuzzy logic is given by Zimmermann (*14*).

FUZZY RULE BASED CAR-FOLLOWING MODEL

The model consists of two modules: a fuzzy inference system and a system that executes the inference system.

Fuzzy Inference System

The inference system infers the reaction of the FV in acceleration (or deceleration) rates in response to the action of the LV. Using the following notation, the structure of the system is presented.

d : distance between LV and FV (in specific value)
s : relative speed between LV and FV (in specific value)
a : rate of change of speed of LV (in specific value)
DS_i : perceived distance (in fuzzy number)
RS_i : perceived relative speed (in fuzzy number)
ALV_i : perceived rate of change of speed of LV (in fuzzy number)
AFV_i : reaction of FV in acceleration (or deceleration) rate (in fuzzy number)
AFV' : predicted reaction of FV in acceleration (or deceleration) rate given the input (in fuzzy number)
Input : $x_1 = d, x_2 = s, x_3 = a$
Rule i : If $x_1 = DS_i$, and $x_2 = RS_i$, and $x_3 = ALV_i$, then $y = AFV_i$
. . . . . .
Rule n : If $x_1 = DS_i$, and $x_2 = RS_i$, and $x_3 = ALV_i$, then $y = AFV_i$
Conclusion : $y = AFV'$

In the following, input, rules (the premise, consequence, and structure), and the conclusion of the inference system of the proposed model are discussed.

Input. Since the purpose of the model is to predict the behavioral pattern of the FV when a specific condition is given, the input is a set of parameter values that would affect the FV decision. They are

- Distance between FV and LV (ft),
- Speeds of FV and LV (ft/sec) (to obtain the relative speed), and
- Acceleration or deceleration rate of the LV (ft/sec^2).

Rule: premise. The premise variables of a rule are the distance between the LV and FV (DS), relative speed of the vehicles (RS), and the acceleration (or deceleration) rate of the LV (ALV). The quantity of each of the first two variables is grouped into 6 natural language–based categories, while the last variable is grouped into 12 such categories (6 for acceleration and 6 for deceleration). Each of these categories is a fuzzy set. They are presented in Table 1.

For all categories, triangular membership functions are assumed. For categories that represent DS, the membership function varies with the speed of the FV because it is believed that the notion of safe distance is relative to the speed at which the FV is traveling.

The reason for considering acceleration and deceleration separately is based on our belief that the intensity of FV's reaction is different when the LV is accelerating and deceleration, as discussed in the second section.

Rule: consequence. The consequence of a rule is the FV's reaction in terms of acceleration or deceleration rate expressed in fuzzy quantity (AFV). Each fuzzy quantity can be represented by a natural language term such as VERY STRONG DECELERATION. The reaction of FV should be similar in nature to that of LV since FV wishes to maintain the relative speed near zero. Thus, the membership function of AFV should be similar to that of ALV, but it is modified by the categories chosen for DS and RS in the premise.

If the category of DS in a rule is ADEQUATE, the AFV (in fuzzy number) is computed as follows:

$$\{(RS_i + ALV_i \bullet \Delta t)/\gamma\} = AFV_i \tag{13}$$

where Δt is the time interval at which the rules are applied (or the time intervals at which the inference is run; in our model $\Delta t = 1$ sec); γ is the time in which FV wishes to "catch up" with LV. We choose $\gamma = 2.5$ sec, which keeps the FV's acceleration and deceleration rates within a realistic range (less than approximately 10 ft/sec^2).

The numerator of Equation 13 represents the relative speed at time $t + \Delta t$. Dividing it by γ, we obtain the rate of speed change required for FV to restore zero relative speed. RS_i, ALV_i, and AFV_i are all fuzzy numbers.

If the category of DS in Rule i is different from ADEQUATE, the value of AFV_i is modified. The modification is done by sliding the membership function of AFV_i to the right or to the left (making it larger or smaller) according to DS_i's deviation from the category ADEQUATE. For each deviation to a shorter distance category, AFV_i is reduced by -1 ft/sec^2; for each deviation to a longer distance category AFV_i

TABLE 1 Categories (Fuzzy Sets) of Premise Variables

		Distance btwn. LV and FV (DS)	Relative speed (RS)	Actions of LV (ALV)	
				Acceleration	Deceleration
Categories	(1)	very small	FV slower	strong	strong
	(2)	small	FV slightly slower	somewhat strong	somewhat strong
	(3)	adequate	near zero	normal	normal
	(4)	more than adequate	FV slightly faster	mild	mild
	(5)	large	FV quite faster	very mild	very mild
	(6)	very large	FV faster	none	none

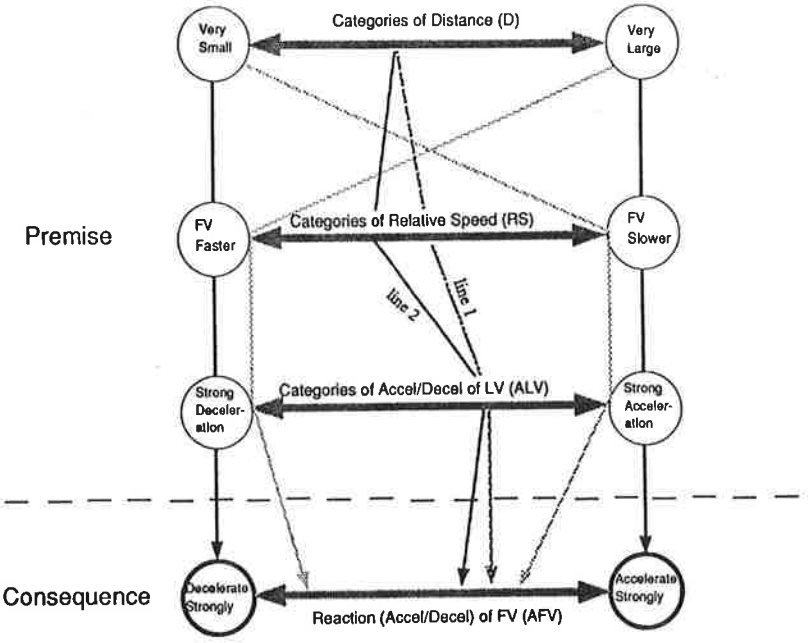


FIGURE 1 Formation of rules.

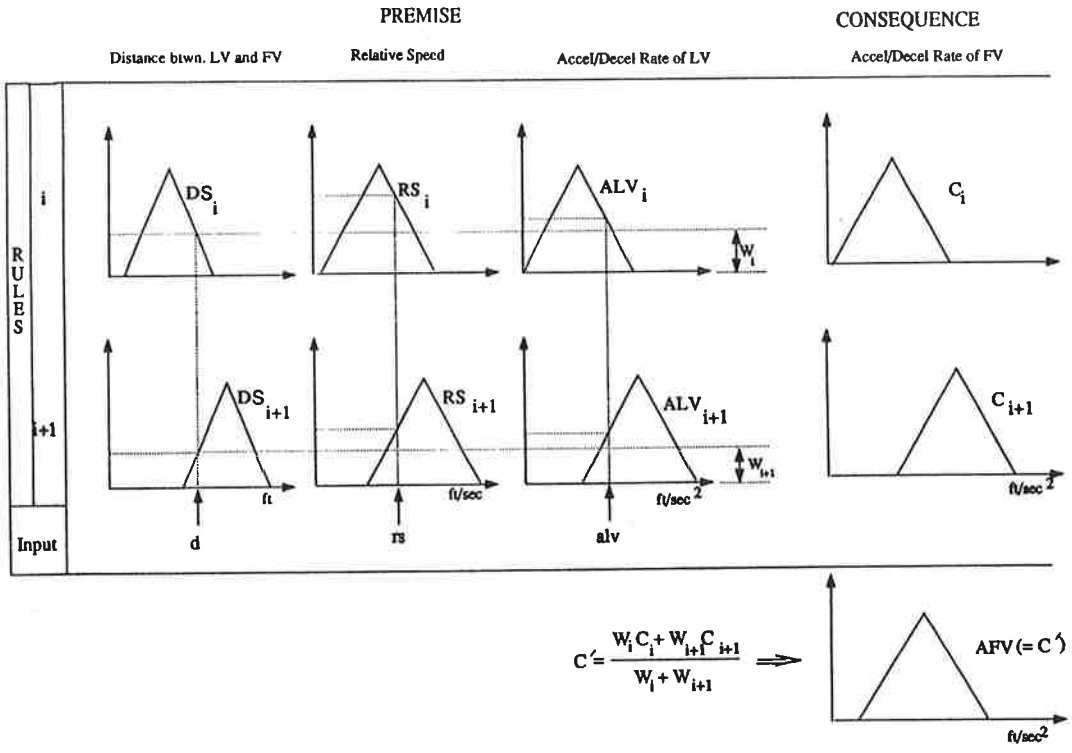


FIGURE 2 Execution of the fuzzy inference system.

is increased by +1 ft/sec^2. That is, AFV_i is determined by

$$\{(RS_i + ALV_i \bullet \Delta t)/\gamma\} + \beta_{DSi} \bullet \phi = AFV_i \quad (14)$$

where β_{DSi} is the number of categories for which DS_i deviates from ADEQUATE (it can be a positive or negative number depending on whether the deviation is to a longer distance or a shorter distance, respectively), and ϕ in this case is 1 ft/sec^2.

Rule: structure. Each rule is a conditional statement in the sense that, given a set of conditions represented by the premise variables, the consequence is predicted. The following is an example:

If Distance (*DS*): ADEQUATE,

Relative Speed (RS): NEAR ZERO, and

Acceleration of LV (ALV): MILD,

then FV should accelerate MILDLY.

The selection of categories of the premise variables and consequences are based on the method discussed previously. Figure 1 shows how a combination of the categories of the premise variables results in a particular consequence (which should lie between STRONG ACCELERATION and STRONG DECELERATION). For instance, the rule in the example could be represented by Line 1 in the figure. It is interesting to observe that a line connecting the upper circles of the premise leads to the very large acceleration of FV, and the line connecting the bottom circles leads to very large deceleration, thus setting the two extreme cases.

The conclusion. The level of compatibility between the input and premise of a rule i, W_i, is determined by the operation shown in Equation 8, except that in this case there are three premise variables. For all the rules for which the value of W_i is greater than zero, the conclusion is computed according to Equation 9 (or Equations 10, 11, and 12) where C's are the FV's acceleration (or deceleration) rate expressed in fuzzy number. The process of deriving the conclusion according to Equation 9 is shown in Figure 2 for the case in which two rules are applied to the same input. (In our example, on the average, three to four rules were fired for the same input).

Execution of the Model

The model executes the inference system at small time increments (1 sec in our example). At each time increment, the action of LV can be changed; for example, in one time increment, it accelerates at a given rate; at the next time interval, it accelerates at another rate. The speed and position of the FV relative to the LV are then updated after each time increment. The time delay between the actions of the LV and FV due to the perception and reaction process is assumed to be 1 sec in our example.

ANALYSIS: TRAFFIC STABILITY AND SPEED-DENSITY RELATIONSHIP

This section applies the model to the analyses of traffic stability and speed-density relationships, representing application to microscopic and macroscopic analyses of traffic flow, respectively. The output of the model is a fuzzy number. The lines that will be shown as model output in Figures 3, 4, 5, and 6 represent the values whose membership grade is 1. Lines A and B in Figure 7 correspond to the value at a membership grade of 0.2.

Traffic Stability

Traffic stability is examined from the local and asymptotic stability standpoint.

Local Stability

After an initial disturbance, the distance between LV and FV stabilizes into a pattern; this pattern is examined for different input conditions.

Figure 3 compares traffic stability obtained from the GM model (for $m = 0$, $\ell = 1$) with the one obtained from the proposed model. The two models are compared under the following conditions.

- Initial distance between LV and FV, 133 ft;
- Speed change of LV: Case 1, LV decelerates from 44.1 to 28.1 ft/sec in 2 sec and remains constant; Case 2, LV decelerates from 52.1 to 28.1 ft/sec in 3 sec and remains constant.

Case 1 corresponds to the example presented by Herman and Potts (*8*). Figures 3a and 3b differ in the assumed value of α in the GM model: $\alpha = 25.1$ ft/sec in Figure 3a, $\alpha = 17$ ft/sec in Figure 3b. Lines 1 and 2 represent the results of the GM model for Cases 1 and 2, and Lines 3 and 4 represent the results of the proposed model for Cases 1 and 2, respectively. Line 1 of Figure 3a is actually the same as the one presented by Herman and Potts (*8*, Figure 15).

Since the final speeds are the same in Cases 1 and 2, the results of Cases 1 and 2 should converge as time increases. This is the case in the proposed model (Lines 3 and 4 eventually merge). However, in the GM model, Lines 1 and 2 remain separate both in Figures 3a and 3b. Figure 3b shows that, for an arbitrarily chosen value of $\alpha = 17$ ft/sec in the GM model, the result of Case 2 is almost identical to the one derived from the proposed model. As seen in the forthcoming figures in this section, for the same final speed the proposed model yields the same stable distance between LV and FV regardless of the initial condition.

Figure 4 shows how the speed change of LV affects the distance D between LV and FV in the proposed model. LV changes its speed from 60 to 50 ft/sec in 2 sec, changes back to 60 ft/sec in 2 sec, and thereafter continues to travel at 60 ft/sec. Each of the four lines represents a different initial distance between LV and FV (100, 120, 150, and 180 ft). It is seen that D settles to approximately 125 ft regardless of the initial distance. However, the way D settles to 125 ft differs with the initial distance. When the initial distance is near 125 ft (final stable distance), D fluctuates more before settling to the stable distance. This suggests that the model can represent the susceptibility of the FV to the action of LV.

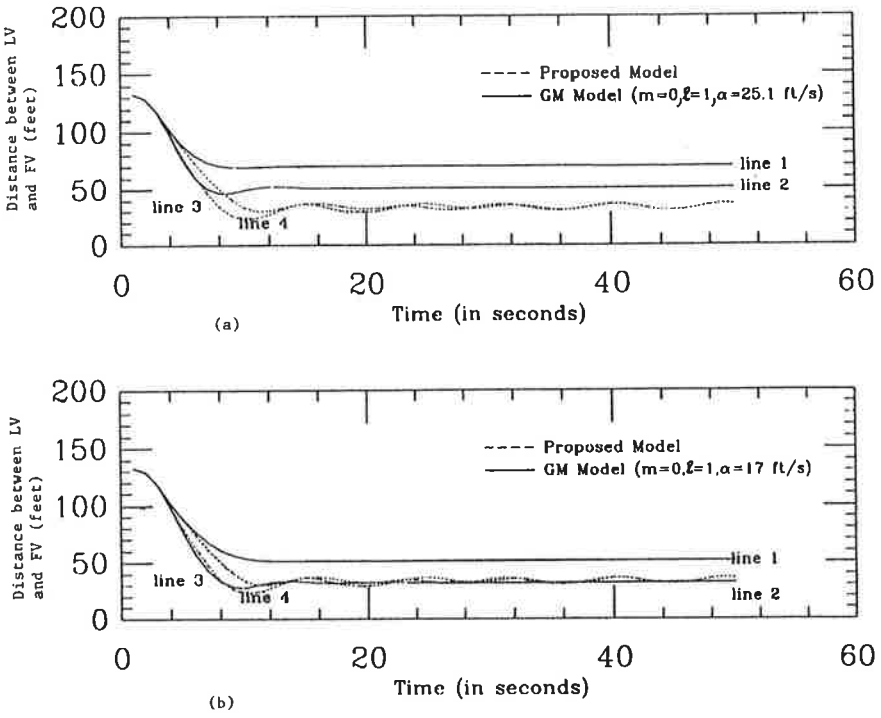


FIGURE 3 Comparison of the traditional car-following model with the proposed model.

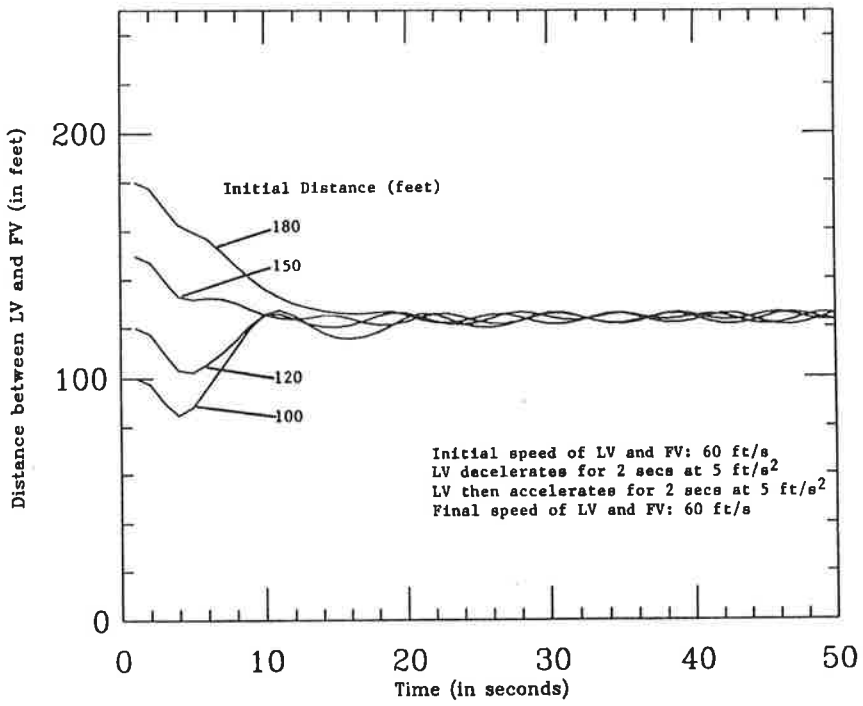


FIGURE 4 Local traffic stability: different initial distances between LV and FV.

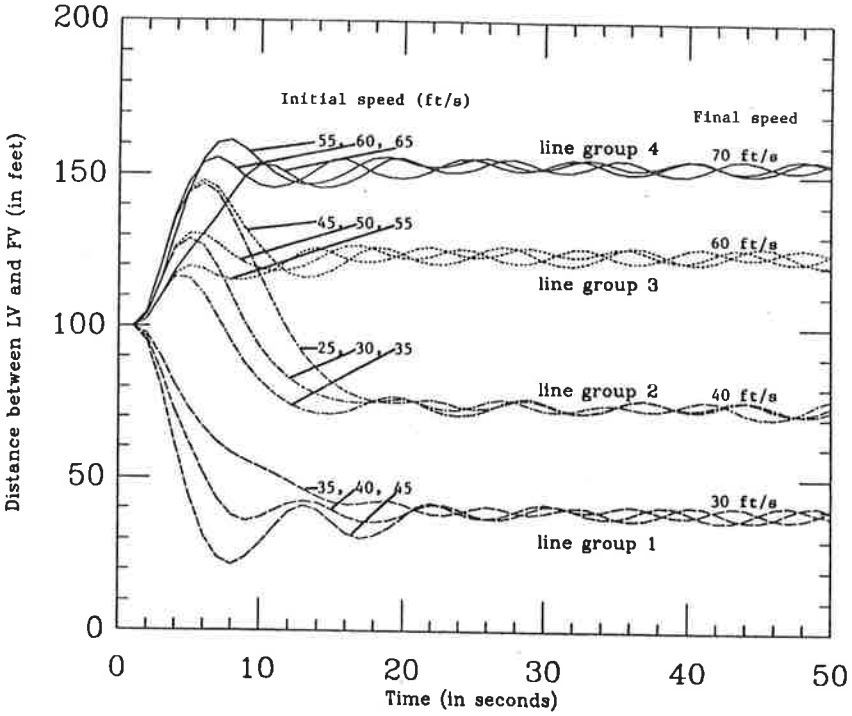


FIGURE 5 Local traffic stability: different initial and final speeds.

Figure 5 shows how D, the distance between LV and FV, depends on the LV's final speed using the proposed model. In all cases, the initial distance is 100 ft. Line Groups 1, 2, 3, and 4 present the final speeds of 30, 40, 60, and 70 ft/sec, respectively. The three lines in each group represent different initial speeds, as noted in the figure. For all cases, the LV is assumed to attain the final speed in 3 sec. Regardless of the initial speed, D approaches a higher constant value for a higher final speed.

Asymptotic Stability

Figure 6 shows how the distance between individual cars in a platoon can vary when the first car in the platoon decelerates and then accelerates under the proposed model. In this example, the platoon consists of five cars. The first car decelerates from 50 to 40 ft/sec in 2 sec, then accelerates back to 50 ft/sec in 2 sec, and thereafter travels at a constant speed of 50 ft/sec. Line 1 represents the variation in distance between the first and second cars, Line 2 between the second and third cars, and so forth. It is interesting to note that the pattern of variation in distance between the third and fourth car (Line 3) is different from the rest.

Speed-Density Relationship

After the distance between LV and FV settles to a stable value, D^*, the speed-density (u-k) relationship is analyzed. This was performed for different final stable speeds. Density is computed in the number of vehicles per mile. Since the value of D^* obtained from the model is a fuzzy number, the density obtained from D^* is also a fuzzy number, and thus the predicted u-k relationship is a fuzzy relationship. The computed fuzzy u-k relationship is compared with the plot of observed data in Figure 7.

In the figure the band formed by Lines A and B is the range of possible densities for the speed. The range corresponds to the density whose membership grade is 0.2 or greater. The value 0.2 is chosen only for the purpose of reference in this paper. Line C corresponds to the locus of the density whose membership grade is 1.

When density is high, the vehicles are expected to travel in the car-following pattern. Therefore, a reasonable match between the predicted and observed u-k relationships is expected. This notion is supported by the figure for the range where density is greater than approximately 40 vehicles per mile (vpm).

When density is low, the vehicles are expected to travel independent of one another. Therefore the car-following pattern (stimulus-response interaction) is not likely to be sustained, and hence, the proposed model would not be valid; thus the lines A, B, and C are shown only for density values greater than 40 vpm, the region where flow is reasonably dense.

The data points in the previous figure are derived from the speed-occupancy data obtained from Queen Elizabeth Way, Ontario, Canada (courtesy of Fred Hall). The conversion from the occupancy measure to the density measure is performed on the basis of an average vehicle length of 20 ft.

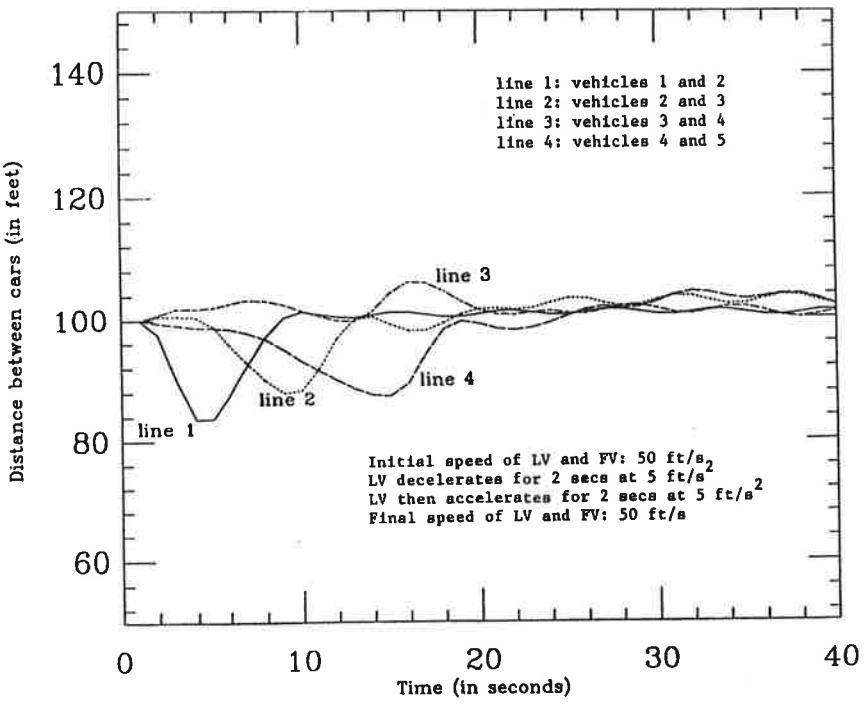


FIGURE 6 Asymptotic traffic stability for five-car platoon.

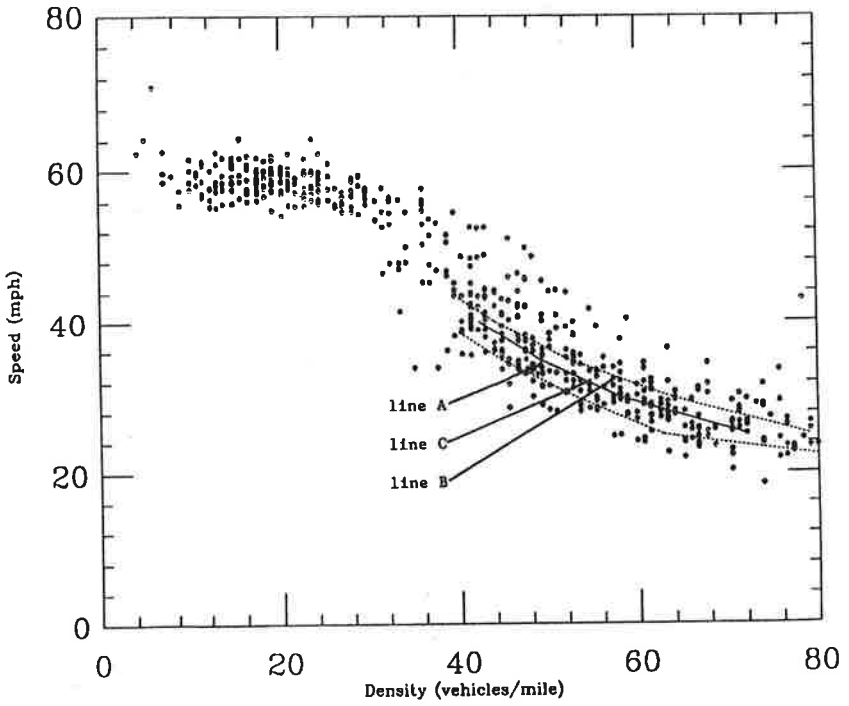


FIGURE 7 Speed-density relationship: predicted relationship and observations.

CONCLUSIONS

Decisions and actions of a driver are believed to follow a reasoning process based on vague logic. The model proposed in this paper applies the fuzzy inference system and simulates the car-following phenomenon. The output of the inference system is the FV's reaction (acceleration or deceleration rate) in fuzzy number in small time increments. By integrating this output over time, the movements of FV relative to LV are simulated. The purpose of the paper is to present the methodology of building the model. The shapes of specific membership functions used in the model must still be verified through field data collection.

The proposed model is a response to the concern that drivers do not exercise the dichotomous decision criteria assumed in the traditional deterministic car-following models. The model proposed has the following characteristics:

1. Driver's decision criteria are handled by fuzzy inference logic, which allows several decision rules to fire at the same time for a given set of input. As a result, the final output incorporates the ambiguity of the decision process.
2. The inference rules are a collection of natural language–based straightforward driving rules. The number of rules can be adjusted, and each rule can be independently modified to suit the decision criteria.
3. The output is a fuzzy number that represents a range of possible acceleration (or deceleration) rates of the FV. Thus, it captures the characteristics of traffic flow as the conglomeration of an individual driver's possible actions. Under the deterministic models, the variation of data points is viewed as random variation from a norm.
4. The result is realistic and consistent with the general expectation from a car-following model: for the same final speed, the distance between LV and FV eventually converges to the same value regardless of the initial condition. The "drift," oscillation of the distance between LV and FV, can also be captured.

The proposed approach to the car-following problem should have a number of applications, including control of vehicle separation under the IVHS. For the traffic flow analysis, the model can be extended to derive a possibility-based speed-volume relationship. Such a relationship would allow us to analyze the capacity as a fuzzy number and to recognize the level of service as the fuzzy measure of traffic conditions, instead of as the traditional rigidly bounded measure.

ACKNOWLEDGMENT

The authors are grateful to Fred L. Hall of McMaster University, Ontario, Canada, for providing the traffic flow data used in this study.

REFERENCES

1. A. D. May. *Traffic Flow Fundamentals*. Prentice-Hall, Englewood Cliffs, N.J., 1990.
2. R. E. Chandler, R. Herman, and E. W. Montroll. Traffic Dynamics: Studies in Car-Following. *Operations Research*, Vol. 6, 1958, pp. 165–184.
3. D. C. Gazis, R. Herman, and R. B. Potts. Car-Following Theory of Steady State Traffic Flow. *Operations Research*, Vol. 7, 1959, pp. 499–505.
4. D. C. Gazis, R. Herman, and R. W. Rothery. Non-Linear Follow the Leader Models of Traffic Flow. *Operations Research*, Vol. 9, 1960, pp. 545–567.
5. R. Herman, E. W. Montroll, R. B. Potts, and R. W. Rothery. Traffic Dynamics: Analysis of Stability in Car-Following. *Operations Research*, Vol. 7, 1959, pp. 86–106.
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. W. Leutzbach. *Introduction to the Theory of Traffic Flow*. Springer-Verlag, 1988.
8. R. Herman and R. B. Potts. Single-Lane Traffic Theory and Experiment. *Proc., Symposium on the Theory of Traffic Flow*, Dec. 1959, pp. 120–146.
9. P. Ross. Modelling Traffic Flow. *Public Roads*, Dec. 1987, pp. 90–96.
10. R. S. Gilchrist and F. L. Hall. Three-Dimensional Relationships Among Traffic Flow Theory Variables. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 99–108.
11. R. T. Underwood. *Speed, Volume and Density Relationships*. Thesis. Bureau of Highway Traffic, Yale University, New Haven, Conn., 1960.
12. L. A. Zadeh. Fuzzy Sets as a Basis for a Theory of Possibility. *Fuzzy Sets and Systems*, Vol. 1, 1978, pp. 3–28.
13. G. J. Klir and T. A. Folger. *Fuzzy Sets, Uncertainty, and Information*. Prentice-Hall, Englewood Cliffs, N.J., 1988.
14. H.-J. Zimmerman. *Fuzzy Set Theory and Its Applications* (2nd ed.). Kluwer Academic Publishers, 1990.
15. B. Kosko. *Neural Networks and Fuzzy Systems*. Prentice-Hall, Englewood Cliffs, N.J., 1991.
16. D. Dubois and H. Prade. *Fuzzy Sets and Systems, Theory and Applications*. Academic Press, New York, 1980.
17. A. Kaufmann and M. M. Gupta. *Introduction to Fuzzy Arithmetic, Theory and Applications*. Van Nostrand Reinhold Company, New York, 1985.
18. T. Takagi and M. Sugeno. Fuzzy Identification of Systems and Its Applications to Modeling and Control. *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 15, No. 1, 1985, pp. 116–132.

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

Statistical Properties of Vehicle Time Headways

R. T. LUTTINEN

The properties of vehicle time headways are fundamental in many traffic engineering applications. The shape of the empirical headway distributions is described by density estimates, coefficients of variation, skewness, and kurtosis. The hypothesis of exponential tail is tested by Monte Carlo methods. The independence of consecutive headways is tested using autocorrelation analysis, runs tests, and goodness of fit tests for geometric bunch size distribution. The power of these tests is enhanced by calculating combined significance probabilities. The variation of significance over flow rates is described by "moving probabilities." It is shown that speed limit and road category have a considerable effect on the statistical properties of vehicle headways. The results also suggest that the renewal hypothesis should not be accepted under all traffic conditions.

Vehicle time headways play an important role in many traffic engineering applications, such as vehicle-actuated traffic signals, gap availability, and pedestrian delay. Mathematical analysis and simulation of these systems are usually based on theoretical models. The models should be verified against the properties of real world headways. This paper presents some statistical properties of vehicle headways on Finnish two-way, two-lane roads.

The properties of headways have been extensively studied, especially in the 1960s. Some of the earlier work is reviewed and compared with recent data. More powerful statistical techniques are also presented.

DATA COLLECTION AND PRELIMINARY ANALYSIS

Data Collection

The data were collected in 1984 and 1988 by the Laboratory of Traffic and Transportation Engineering at Helsinki University of Technology. A traffic analyzer with two inductive loops on both lanes recorded for each passing vehicle its serial number, time headway (time from front bumper to front bumper in units of 1/100 sec), net time headway (time from back bumper to front bumper in units of 1/100 sec), speed (in units of 1 km/hr) and length (in units of 1/10 m).

The study sites had speed limits of 50, 60, 70, 80, and 100 km/hr. The roads with lower speed limits had a lower overall standard, but all the road sections were reasonably level and straight, and no steep hills, intersections, or traffic signals were near. The samples were classified into two road categories with speed limits 50 to 70 km/hr and 80 to 100 km/hr.

TL Consulting Engineers, Ltd., Vesijärvenkatu 26 A, 15140 Lahti, Finland.

The observations from high-speed (80 to 100 km/hr) roads were measured in 1984. These data have been previously analyzed by Pursula and Sainio (*1*) and Pursula and Enberg (*2*). Because the data were collected for capacity studies, the observations are concentrated in high volumes. Two-way volumes, also, are higher on high-speed roads. On low-speed roads the observations are more concentrated in low volumes.

More than 73,000 headways were recorded on 19 locations. Speed data were corrected according to radar measurements. Data sets with more than 1 percent overtakers were discarded. The samples were analyzed for trend using the method described later. Samples with more than 10 percent heavy vehicles (length > 6 m) were excluded from further analysis. Sixty-four trendless samples were chosen for further analysis.

The data that passed the preliminary phase consist of 64 samples and 16,570 observations (75 to 900 observations per sample). The flow rates vary from 140 to 1840 veh/hr.

Trend Analysis

The temporal variation of traffic is due to deterministic and random factors. To get generally applicable results about headway characteristics, it is necessary to consider stationary conditions. All nonrandom variation should be removed from the measurements as far as possible. Two approaches have been commonly used to overcome the problem of nonrandom variation: the samples are collected either as fixed time slices or using trend analysis.

In the first method the measurements are investigated in fixed time slices of length short enough to exclude any significant trend, typically 30 sec to 10 min (*1,3,4*). The number of headways in one sample is usually too small for statistical analysis, so it is necessary to group samples having nearly equal means. This may cause distortion in the empirical headway distribution because of inappropriate distribution of the sample means.

Because of these problems the second sampling method was chosen. Trend analysis was performed with a computer program (TRENDANA) showing graphically each headway, 15-point moving average, cumulative vehicle count, and the speed of each vehicle. The data were analyzed sequentially using trend tests. The sample size was incremented by 50 until the test reported trend at 5 percent level of significance. The sample was then decremented until the level of significance for trend was 30 to 70 percent with sample size more than 100 and sampling period between 5 and 40 min. Under low volumes the sample size or period length condition had to be relaxed sometimes. If a satisfactory sample was not found,

the first observations that apparently caused the trend were removed and the process was repeated.

The program supports three trend tests: weighted sign (WS) test (*5*), Kendall's rank correlation (RC) test (*6*), and exponential ordered scores (EOS) test (*7*).

Empirical power curves of these tests were evaluated using Monte Carlo methods. The EOS test was found most powerful with RC test close behind (*8*). Because of greater computational effort, the EOS test was used only in more detailed analysis. Preliminary analysis was performed using faster tests.

SHAPE OF HEADWAY DISTRIBUTION

Density Estimates

The density function of the headway distribution is estimated by the histogram method with origin at 0 and bin width 1. Figure 1 shows a surface plot of density estimates on low-speed roads at different flow levels. Headway 0 is assigned Frequency 0. Only samples having more than 100 observations are included.

The distribution is skewed to the right. The proportion of headways less than 1 sec is small. The mode is rather constant (1.5 sec) under all speed limits and flow rates. Because the distribution is unimodal and skewed to the right, the measures of location occur in the following order: mode, median, mean (*9–11*).

Peak heights of the empirical distributions are shown in Figure 2. The peak rises as the flow increases. On high-speed roads the peak value is higher than on low-speed roads. On low-speed roads the peak rises steeply under medium and low flow rates. Under high flow rates the speed limit looses its significance.

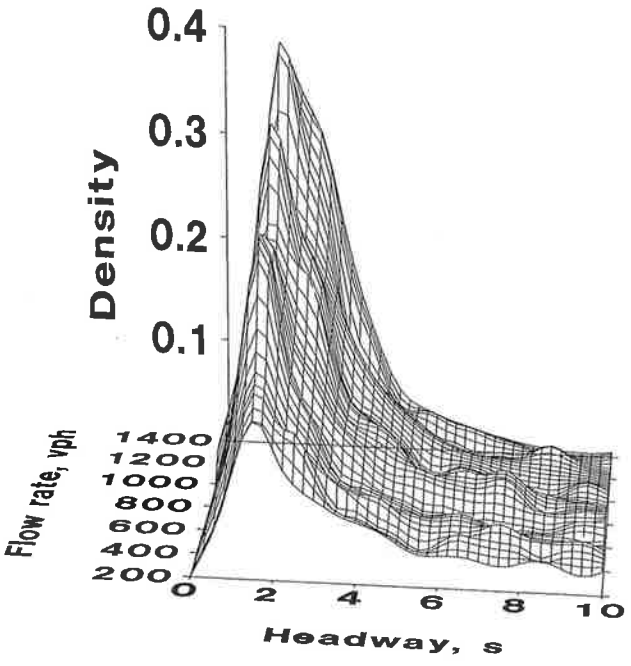


FIGURE 1 Surface plot of headway density estimates on low-speed (50 to 70 km/hr) roads.

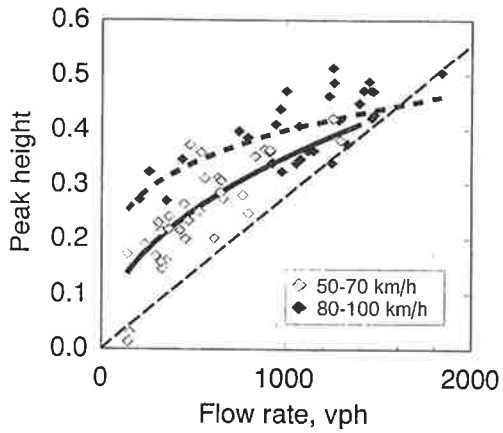


FIGURE 2 Peak height of empirical headway distributions for speed limits 50 to 70 km/hr (solid curve) and 80 to 100 km/hr (dashed curve). Thin dashed line is the peak height of the exponential distribution.

Coefficient of Variation

The sample coefficient of variation (CV) is the proportion of sample standard deviation to sample mean:

$$CV = s_T/\mu_T \tag{1}$$

In distribution functions CV is the proportion of standard deviation to expectation. The negative exponential distribution has CV equal to 1.

Polynomial curves have been fit (Figure 3) to the data for high-speed and low-speed roads. The curves are forced to 1 at flow rate $q = 0$. This is based on the assumption of Poisson tendency in low density traffic.

Some basic properties of CV can be observed in Figure 3:

1. Under heavy traffic the proportion of freely moving vehicles is small. The variance of headways is accordingly small. This phenomenon is reflected in the figure by $CV < 1$ at high flow levels.
2. The Poisson tendency of low density traffic has a theoretical (*12*) as well as an intuitive basis: under light traffic

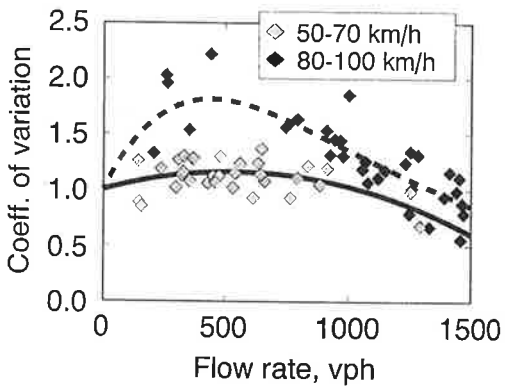


FIGURE 3 Coefficient of variation of the headway data for speed limits 50 to 70 km/hr (solid curve) and 80 to 100 km/hr (dashed curve).

vehicles can move freely, and randomness of the process increases. *CV* is therefore expected to approach 1 as flow approaches 0.

3. Under medium traffic there is a mixture of leading and trailing vehicles. This increases the variance above pure random process, and *CV* rises above 1. This is in contrast to the statement by May (*11*) that *CV* approaches 1 under low flow conditions but decreases continuously as the flow rate increases.

4. High-speed roads have greater *CV* than low-speed roads. In the present data the opposite flow rate is higher on high-speed roads, thus reducing overtaking opportunities. Other explanatory factors may be higher variation of speeds and greater willingness to overtake on high-speed roads. On low-speed roads the intersections are more densely spaced. So, there are more joining and departing vehicles, and trip lengths are usually shorter.

These observations gain at least partial support from other authors, as seen in Figure 4. The figure also shows that the studies are based on quite different data. The data sets of Breiman et al. (*13*), Buckley (*3*), and May (*10*) come from a freeway lane. The data of Dunne et al. (*14*) come from a two-lane rural road. *CV* is greater than 1 in all samples of Dunne et al., less than 1 in all samples of May, and near 1 in the samples of Buckley. The samples of Buckley have values similar to the present data from low-speed roads. Finnish studies (*1*) suggest that coefficient of variation on freeways, especially on the first lane, is lower than on two-lane highways.

Skewness and Kurtosis

The proportion of the first two moments was discussed earlier. The third and fourth moments about the mean, skewness and kurtosis, give more information about the shape of the distribution. Skewness is a measure of symmetry. Symmetric distributions have null skewness. If the data are more concentrated on the low values, as in headway distributions, skewness is positive. Kurtosis is a measure of how "heavy" the tails of a distribution are.

Figure 5 shows the sample kurtosis against the square of sample skewness. This relationship is sometimes used as a guide in selecting theoretical distributions (*15*). There is a strong linear relationship, which suggests the usefulness of this measure in model selection.

Points and lines of some theoretical distributions are shown for comparison. As skewness grows, kurtosis increases more slowly than in either the gamma or the lognormal distributions. The gamma distribution is closer to observed values than the lognormal distribution, even though the lognormal distribution is a better model for a headway distribution. The exponential distribution reduces to a point and totally loses the variety in the empirical headway distributions.

Exponential Tail Hypothesis

Several headway models are combinations of two distributions: one for leaders and the other for followers (*3,16–20*). The assumption is usually made that the leaders' headway distribution is exponential.

The tails (large headways) of empirical headway distributions were tested for exponentiality. Goodness of fit tests are based on the Anderson-Darling statistic. Such tests are more powerful than the better-known Kolmogorov-Smirnov and chi-squared tests (*21*). Because the parameters of the distribution are estimated from the sample, nonparametric tests give too conservative results (*22,23*). So, the significance of the tests was estimated using parametric tests and Monte Carlo methods (*24*). The number of replications was 10,000. The tests were performed using threshold values (t_0) from 0 to 14.5 in increments of 0.5.

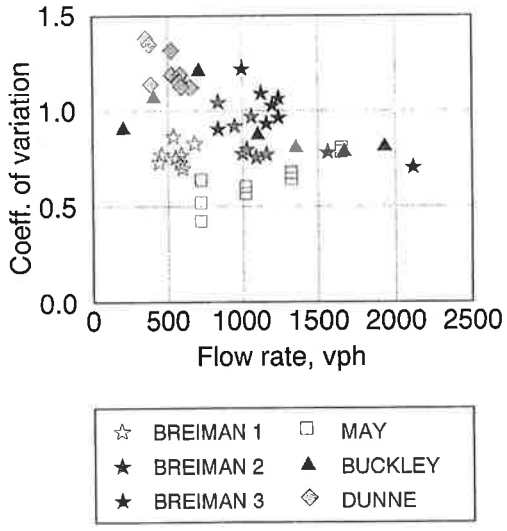


FIGURE 4 Coefficient of variation of headway distributions from different sources. (The number after "Breiman" stands for the freeway lane.)

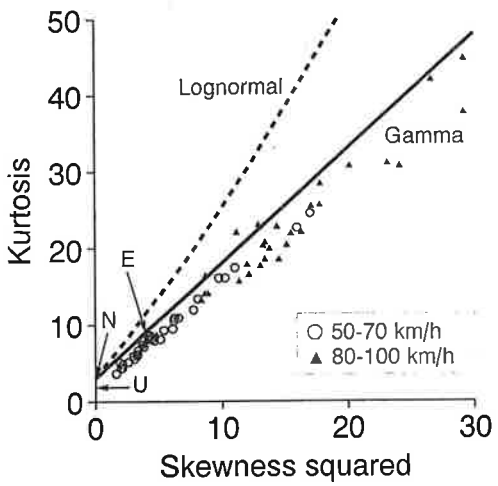


FIGURE 5 Kurtosis and squared skewness for the headway samples and some theoretical distributions. Exponential (E), normal (N), and uniform (U) distributions reduce to points shown by arrows.

To get a more powerful test, the significance probabilities from several samples were combined using the method proposed by Fisher (*25*). If the null hypothesis is true in all samples and the samples are independent, the probabilities are uniformly U(0,1) distributed. If probabilities p_i are U(0,1), the statistic

$$z = -2 \sum_{i=1}^{n} \ln p_i \quad (2)$$

has chi-squared distribution with $2n$ degrees of freedom. So, the combined significance (P) is the probability that a variable Z having chi-squared distribution with $2n$ degrees of freedom is greater than z:

$$P = Pr\{Z > z\} = 1 - F_{chi2}(z; 2n) \quad (3)$$

Figure 6 shows the combined significance levels against threshold values (t_0) at low- and high-speed roads and at different flow levels. Wasielewski (*4*) found no departures from the exponential distribution at threshold values greater than or equal to 4 sec on freeways. On the basis of Figure 6 this value appears too low on two-lane roads. The threshold value for not rejecting the hypothesis of exponential tail is about 8 sec. On low-speed roads lower threshold values may be possible. Miller (*16*) also found $t_0 = 8$ sec appropriate. Because of large t_0, a headway distribution that has positive skewness (such as gamma and lognormal distributions) should be considered for the followers.

Another indicator of the threshold is the influence of the speed of a vehicle on the speed of the trailing vehicle. A driver considering himself as a follower adjusts his speed to the speed of the vehicle ahead. This speed adjustment decreases the variation of relative speeds (speed differences) among successive vehicles. At some time distance the interaction of speeds disappears, and variation of relative speeds remains constant among vehicles having larger headways than the threshold. Figure 7 shows the standard deviation (s_v) of relative speeds against headway. Headways are combined in 1-sec intervals ($t-1$, t). At short headways s_v is small and increases as the headway increases. At large headways s_v is rather constant, but has greater variation because of smaller sample sizes. High-speed roads have larger s_v than low-speed roads.

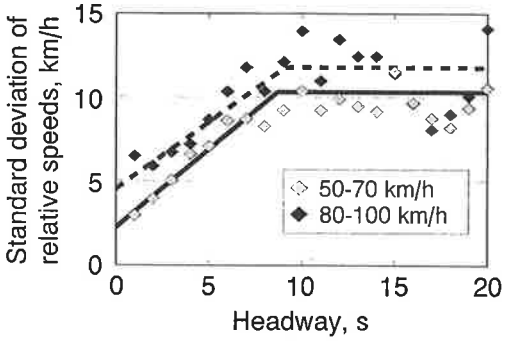


FIGURE 7 Standard deviation of relative speeds against time headways on low-speed (solid line) and high-speed (dashed line) roads.

A piecewise linear model was fit to the data—rising slope for headways less than the threshold and constant value for headways greater than the threshold. The s_v values were weighted by the number of observations in the interval. On both low- and high-speed roads the threshold value of about 9 sec was obtained. The original *Highway Capacity Manual* (*26*) applies similar methods with the same result. Similar analysis of motorway data by Branston (*27*) gives values of 4.5 sec and 3.75 sec for nearside and offside lanes, respectively.

Because all vehicles having headways ≤ 8 sec are not followers, the distribution of their relative speeds is a mixture of free speeds and constrained speeds. By fitting a mixed normal distribution to the relative speed data (Figure 8) the proportion of free-flowing vehicles among headways ≤ 8 sec is estimated to be 20 percent on high-speed roads and 15 percent on low-speed roads. The proportion of trailing vehicles is then estimated to be approximately the same as the proportion of headways ≤ 3.1 sec and ≤ 5.0 sec, respectively.

In the present data the flow rate at which the headway coefficient of variation (Figure 3) reaches its maximum has about 60 percent trailing vehicles. Yet, more extensive data sets and more accurate measuring equipment are needed for

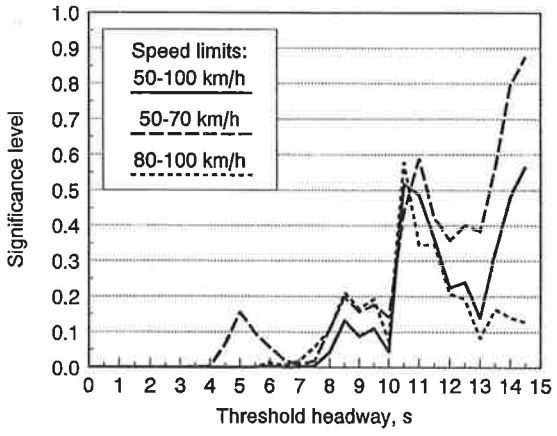


FIGURE 6 Goodness-of-fit tests for exponential tail of headway distributions.

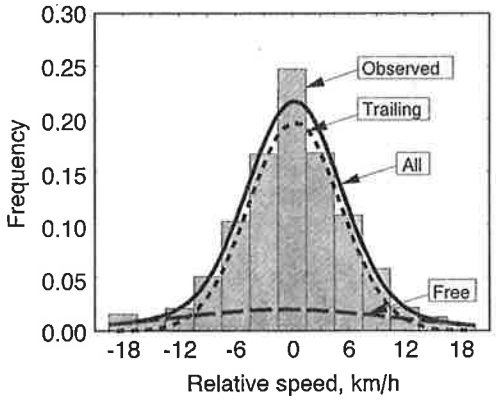


FIGURE 8 Relative speed distribution for headways ≤ 8 sec on high-speed roads as a mixed normal distribution.

further analysis, especially because relative speeds have larger measurement errors than absolute speeds.

RENEWAL HYPOTHESIS

Autocorrelation

The shape of the headway distribution describes the frequency of headways of different length. One step further in the statistical analysis is to examine the order in which the headways take place. A common assumption is the renewal hypothesis, which states that headways are independent and identically distributed. This hypothesis makes many theoretical analyses much easier. On the other hand, correlation between consecutive headways could give additional information for adaptive traffic control systems.

The autocorrelation coefficient is a measure of correlation between observations at given distances (lags) apart. In a sample of n observations the estimate of the autocorrelation coefficient at Lag k is

$$\hat{r}_k = \frac{\sum_{j=1}^{n-k} (T_j - \mu_T)(T_{j+k} - \mu_T)}{\sum_{j=1}^{n} (T_j - \mu_T)^2} \tag{4}$$

where

$$\mu_T = 1/n \sum_{j=1}^{n} T_j \tag{5}$$

The coefficient estimates are asymptotically $N(0, 1/n)$ distributed.

Autocorrelation coefficients indicate whether the observations are from a renewal process ($r_k = 0, k = 1, 2, \ldots$). The most important coefficient in this respect is r_1. The test is

$$H_0: r_1 \leq 0$$

$$H_1: r_1 > 0 \tag{6}$$

This is a one-sided test in contrast to the two-sided tests normally used (*14,20,28*). The tests for negative autocorrelation gave clearly nonsignificant values.

The variation of significance probabilities (p_i) over flow rates (q_i) is described by "moving probability" (Figure 9). The k-point moving probability at flow rate Q_j is

$$P_k(Q_j) = 1 - F_{\text{chi2}}(z_j; 2k) \tag{7}$$

where

$$z_j = -2 \sum_{i=j}^{j+k-1} \ln p_i \tag{8}$$

$$Q_j = 1/k \sum_{i=j}^{j+k-1} q_i \qquad j \in \{1, \ldots, n - k + 1\}$$

On high-speed roads there is significant positive autocorrelation among consecutive headways. There is a spike in the

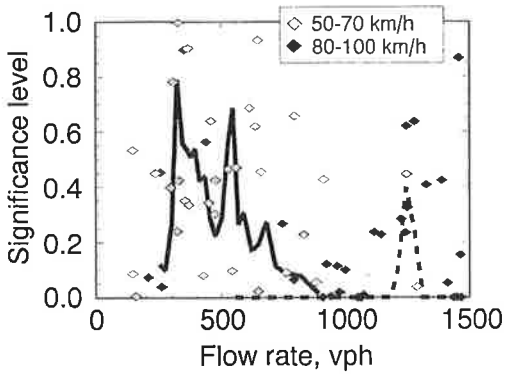


FIGURE 9 Significance of sample autocorrelation coefficients at Lag 1. Nine-point moving probabilities for low-speed (solid curve) and high-speed (dashed curve) roads.

curve, for which no explanation except random variation is found. (A similar although lower spike is in Figures 10 and 11.) On low-speed roads there is no significant autocorrelation, at least under low flow rates. The moving probability curve, however, goes down to significant values near flow rate q = 1,000 veh/hr. The combined significance for high-speed samples is about $3 \cdot 10^{-22}$ and for low-speed samples 0.04.

These results suggest that the renewal hypothesis should be rejected on high-speed roads, especially under flow rates above 500 veh/hr. On low-speed roads the possibility of autocorrelation should be considered at least under flow rates greater than 1,000 veh/hr.

Dunne et al. (*14*) also studied the autocorrelation of trend-free samples. The combined probability of their nine data sets (Lag 1) is 0.702 (one-sided test), which is consistent with the renewal hypothesis. The two-sided test gives 0.284, which is also nonsignificant. Breiman et al. (*28*) found in one of eight data sets (three-lane unidirectional section of John Lodge Expressway in Detroit) significant autocorrelation (Lag 1) at the 0.05 level. The hypothesis of independent intervals was not rejected. The combined significance is 0.23. Testing for positive autocorrelation only (one-sided test), the combined significance is 0.05, suggesting possible positive autocorrelation. Cowan (*19*) studied 1,324 successive headways. The re-

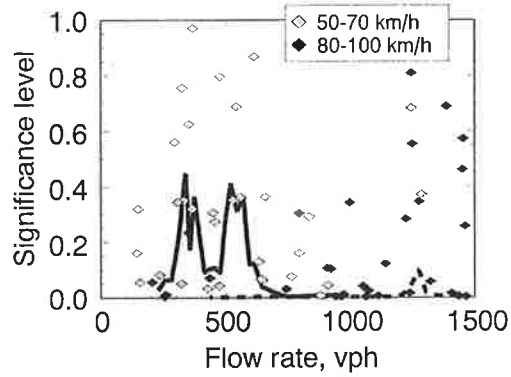


FIGURE 10 Significance of runs tests. Probability of fewer runs above or below median. Seven-point moving probabilities for low-speed (solid curve) and high-speed (dashed curve) roads.

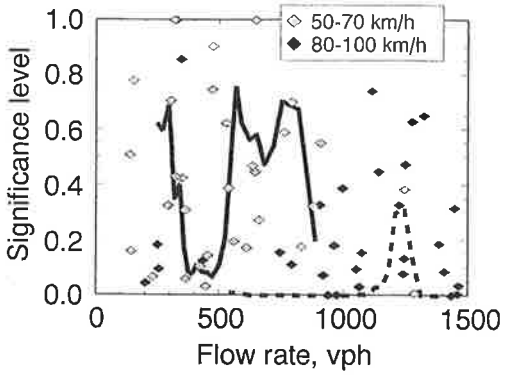


FIGURE 11 Significance of bunch size tests for geometric distribution. Nine-point moving probabilities for low-speed (solid curve) and high-speed (dashed curve) roads.

newal hypothesis was not rejected. Chrissikopoulos et al. (*20*) studied six samples. The result of their unspecified tests was that the headways are independently distributed. However, the combined significance level of their data sets (Lag 1) is 0.012 using two-sided test and 0.0015 when testing for positive autocorrelation. This result disagrees with their conclusion. (Combining further the three combined probabilities above yields 0.008 for one-sided and 0.027 for two-sided tests.) Breiman et al. (*13*) allow for the possibility of small positive autocorrelation in freeway traffic.

Previous studies have so far supported the renewal hypothesis. Further analysis of this material has now cast some doubt on the conclusion. Also, the new data presented here show that the possibility of positive autocorrelation between consecutive headways should be taken seriously, although the magnitude of autocorrelation is small (about 0.1).

Randomness

Randomness of the headway data was tested using the Wald and Wolfowitz (*29*) runs test. The test is performed to determine whether long and short headways are randomly distributed or whether short headways are clustered. Testing runs above and below the median is appropriate here (*28*). Clustering reduces the number of runs. The number of runs is also reduced by trends in the data. Therefore, it is important to have trendless data.

The number of runs is assumed to be normally distributed with mean and variance equal to

$$\text{mean} = 2r(n - r)/n + 1$$

$$\text{variance} = 2r(n - r)\,[2r(n - r) - n]/[n^2(n - 1)] \qquad (9)$$

where n is the total number of observations and r is the number of observations below the median (*29*). Observations equal to the median are ignored. One-sided test is used to find the probability of fewer runs.

Figure 10 shows the moving probabilities for high- and low-speed roads. On high-speed roads the test gives significant values nearly everywhere. On low-speed roads nonrandomness is significant under flow rates > 700 veh/hr. The combined significance for high-speed roads is $1.6 \cdot 10^{-20}$ and for low-speed roads is 0.001. These results suggest that the arrival process in road traffic is not totally random, but clustered.

Breiman et al. (*28*) found only one significant value in eight runs tests on their data sets. The combined probability (0.49) is also nonsignificant. This is in contrast to the preceding results.

Bunching

Vehicle i is a follower (0) if it has headway at most s, otherwise it is a leader (1). The status of a vehicle is accordingly defined as

$$X_i = \begin{cases} 0 & \text{if } T_i \le s \\ 1 & \text{if } T_i > s \end{cases} \qquad (10)$$

The difference in speed is ignored. If the headways are independent and identically distributed (i.i.d.), the probability of Vehicle i being a follower is

$$p = Pr\{T_i \le s\} \qquad (11)$$

The number of vehicles in a bunch is the number of consecutive headways $\le s$ (followers) plus 1 (leader). The bunch is of size n if $X_1 = 1, X_2 = 0, \ldots, X_n = 0, X_{n+1} = 1$. If the headways are i.i.d., the bunch sizes are geometrically distributed (*30*) and the probability of bunch size k is

$$p_k = p^{k-1}(1 - p) \qquad k \ge 1 \qquad (12)$$

Now the renewal hypothesis (i.i.d. headways) can be tested using the null hypothesis

$$H_0: p_k = p^{k-1}(1 - p) \qquad (13)$$

against

$$H_1: p_k \ne p^{k-1}(1 - p) \qquad (14)$$

The chi-squared test was performed with $m-2$ degrees of freedom, where m is the number of classes (different bunch sizes) in the sample. One degree of freedom was lost, because p was estimated from the sample. The threshold for leaders was set to $s = 5$ sec. Bunch sizes 1, 2, . . . , 20 and > 20 were separated into distinct classes. They were combined so that the expectation for each class was ≥ 5, except for the last, which was ≥ 1. On high-speed roads two samples (having flow rates of 1,837 and 1,457 veh/hr) were left out of the analysis, because after combining classes there were no degrees of freedom left.

Figure 11 shows the results of the chi-squared tests. The combined probability is 0.13 on low-speed roads. On high-speed roads the combined probability is $1.9 \cdot 10^{-9}$. So, the hypothesis of geometric bunch size distribution should be rejected on high-speed roads, at least under flow rates above 500 veh/hr. On low-speed roads the hypothesis of geometric distribution cannot be rejected. Chrissikopoulos et al. (*20*) and Taylor et al. (*31*) discard the geometric distribution as a bunching model.

CONCLUSIONS

There is a considerable difference between high-standard and low-standard roads. The headway distributions on high-standard roads have higher peak values and higher coefficient of variation. That is, at a given flow level there are more small headways on high-standard roads. The vehicles are also more clustered and there is a small positive autocorrelation between consecutive headways. On low-standard roads there is some indication of possible positive autocorrelation under high flow rates. On high-standard roads the autocorrelation is statistically significant but too small to be helpful, for example, in traffic control applications. In simulation studies and bunching models the stochastic structure of the arrival process should be fully considered.

The examination of relative speeds and the tail of the headway distribution supports the view that drivers become affected by the vehicle ahead when the headway is less than 8 to 9 sec. At larger headways the standard deviation of relative speeds is rather constant and headways are exponentially distributed. At smaller headways the standard deviation of relative speeds decreases and the hypothesis of exponentiality must be rejected. The proportion of trailing vehicles on high- and low-speed roads is approximately the same as the proportion of headways ≤ 3.1 sec and ≤ 5.0 sec, respectively.

Local conditions, such as road category, speed limit, and flow rate, have a considerable effect on the statistical properties of headways. The effect of opposing traffic, especially, deserves further research. But the statistical analysis of vehicle headways requires very extensive data sets and the application of powerful statistical techniques.

ACKNOWLEDGMENTS

This research was partly supported by the Henry Ford Foundation in Finland. The cooperation of the Traffic and Transportation Laboratory at Helsinki University of Technology is particularly acknowledged. The comments and suggestions of Matti Pursula and the referees are deeply appreciated.

REFERENCES

1. M. Pursula and H. Sainio. *Basic Characteristics of Traffic Flow on Two-Lane Rural Roads in Finland* (in Finnish). Finnish National Roads Administration, TVH741824. Helsinki, Finland, 1985.
2. M. Pursula and Å. Enberg. Characteristics and Level of Service Estimation of Traffic Flow on Two-Lane Rural Roads in Finland. Presented at 70th Annual Meeting of the Transportation Research Board, Washington, D.C., 1991.
3. D. J. Buckley. A Semi-Poisson Model of Traffic Flow. *Transportation Science*, Vol. 2, No. 2, 1968, pp. 107–133.
4. P. Wasielewski. Car-Following Headways on Freeways Interpreted by the Semi-Poisson Headway Distribution Model. *Transportation Science*, Vol. 13, No. 1, 1979, pp. 36–55.
5. D. R. Cox and A. Stuart. Some Quick Sign Tests for Trend in Location and Dispersion. *Biometrika*, Vol. 42, 1955, pp. 80–95.
6. M. Kendall, A. Stuart, and J. K. Ord. *The Advanced Theory of Statistics. Volume 3: Design and Analysis, and Time-Series* (4th edition). Charles Griffin & Co. Ltd., London, 1983.
7. D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events*. Methuen & Co. Ltd., London, 1966.
8. A. Stuart. The Efficiencies of Tests and Randomness Against Normal Regression. *Journal of the American Statistical Association*, Vol. 51, 1956, pp. 285–287.
9. A. Stuart and J. K. Ord. *Kendall's Advanced Theory of Statistics. Volume 1: Distribution Theory*. Charles Griffin & Company Ltd., London, 1987.
10. A. D. May. Gap Availability Studies. In *Highway Research Record 72*, HRB, National Research Council, Washington, D.C., 1965, pp. 101–136.
11. A. D. May. *Traffic Flow Fundamentals*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1990.
12. L. Breiman. The Poisson Tendency in Traffic Distribution. *The Annals of Mathematical Statistics*, Vol. 32, 1963, pp. 308–311.
13. L. Breiman, R. Lawrence, D. Goodwin, and B. Bailey. The Statistical Properties of Freeway Traffic. *Transportation Research*, Vol. 11, 1977, pp. 221–228.
14. M. C. Dunne, R. W. Rothery, and R. B. Potts. A Discrete Markov Model of Vehicular Traffic. *Transportation Science*, Vol. 2, No. 3, 1968, pp. 233–251.
15. J. K. Cochran and C.-S. Cheng. Automating the Procedure for Analyzing Univariate Statistics in Computer Simulations Contexts. *Transactions of the Society for Computer Simulation*, Vol. 6, No. 3, 1989, pp. 173–188.
16. A. J. Miller. A Queueing Model for Road Traffic Flow. *J. Roy. Statist. Soc. Ser. B.* Vol. 23, No. 1, 1961, pp. 64–75.
17. R. F. Dawson. The Hyperlang Probability Distribution—A Generalized Traffic Headway Model. In *Beiträge zur Theorie des Verkehrsflusses* (W. Leutzbach and P. Baron, eds.). Referate anlässlich des IV. Internationalen Symposiums über die Theorie des Verkehrsflusses in Karlsruhe im Juni 1968. Strassenbau und Strassenverkehrstechnik, Heft 86, 1969, pp. 30–36.
18. D. Branston. Models of Single Lane Time Headway Distributions. *Transportation Science*, Vol. 10, No. 2, 1976, pp. 125–148.
19. R. J. Cowan. Useful Headway Models. *Transportation Research*, Vol. 9, No. 6, 1975, pp. 371–775.
20. V. Chrissikopoulos, J. Darzentas, and M. R. C. McDowell. Aspects of Headway Distributions and Platooning on Major Roads. *Traffic Engineering & Control*, May 1982, pp. 268–271.
21. R. B. D'Agostino and M. A. Stephens. *Goodness-of-Fit Techniques*. Marcel Dekker, Inc., New York, 1986.
22. H. W. Lilliefors. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *American Statistical Association Journal*, Vol. 62, 1967, pp. 399–402.
23. H. W. Lilliefors. On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *American Statistical Association Journal*, Vol. 64, 1969, pp. 387–389.
24. R. T. Luttinen. Testing Goodness of Fit for 3-Parameter Gamma Distribution. In *Proceedings of the Fourth IMSL User Group Europe Conference*, IMSL and CRPE-CNET/CNRS, Paris, 1991, pp. B10/1–13.
25. R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1938.
26. *Highway Capacity Manual*. Bureau of Public Roads, Washington, D.C., 1950.
27. D. Branston. A Method of Estimating the Free Speed Distribution for a Road. *Transportation Science*, Vol. 13, No. 2, 1979, pp. 130–145.
28. L. Breiman, A. V. Gafarian, R. Lichtenstein, and V. K. Murthy. An Experimental Analysis of Single-Lane Time Headways in Freely Flowing Traffic. In *Beiträge zur Theorie des Verkehrsflusses* (W. Leutzbach and P. Baron, eds.). Referate anlässlich des IV. Internationalen Symposiums über die Theorie des Verkehrsflusses in Karlsruhe im Juni 1968. Strassenbau und Strassenverkehrstechnik, Heft 86, 1969, pp. 22–29.
29. A. Wald and J. Wolfowitz. On a Test Whether Two Samples Are from the Same Population. *Annals of Mathematical Statistics*, Vol. 2, 1940, pp. 147–162.
30. R. T. Luttinen. *Introduction to the Theory of Headway Distributions* (in Finnish). Helsinki University of Technology, Traffic and Transportation, Publication 71, Otaniemi, 1990.
31. M. A. P. Taylor, A. J. Miller, and K. W. Ogden. A Comparison of Some Bunching Models for Rural Traffic Flow. *Transportation Research*, Vol. 8, 1974, pp. 1–9.

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

Modeling Queued Driver Behavior at Signalized Junctions

JAMES A. BONNESON

Some of the findings from a recent study of the queue discharge headway process are summarized. One outcome of the study was the development of a model of discharge headway at signalized junctions. The model is based on vehicle and driver capabilities, including driver reaction time, driver acceleration, and vehicle speed. To calibrate the model, data were collected at five signalized junctions. The discharge headway model developed in this research indicates that the minimum discharge headway of a traffic movement is not reached until the eighth or higher queue positions. Application of the model suggests that the minimum discharge headway of a traffic movement under ideal conditions may be shorter than 2.0 sec/veh and that its corresponding start-up lost time may be longer than 2.0 sec.

Some of the findings from a recent study of the queue discharge headway process at single-point urban interchanges (SPUIs) (*1*) are summarized. One outcome of the study was the development of a model of discharge headway at signalized junctions. This model is based on vehicle and driver capabilities such as driver reaction time, driver acceleration, and vehicle speed.

BACKGROUND

Discharge Headway

Average vehicle headways by queue position have been the subject of several past studies (*2–5*). The headways reported in these studies for passenger car through movements are shown in Figure 1, which indicates that the discharge rate varies during the initial portion of the green interval. The variation reflects the reaction time of the first driver responding to the change in signal indication and the steady acceleration of the first few vehicles in queue. Eventually, the headways stabilize at a relatively constant value, which is called the minimum discharge headway. In recognition of this trend toward convergence after the first few vehicles, the 1985 *Highway Capacity Manual* (HCM) (*6*, Chapter 9) recommends that the headways of the fifth and subsequent queued vehicles be averaged to estimate the minimum discharge headway.

Under ideal operating conditions (i.e., 12-ft lanes, all through vehicles, all passenger cars, no parking, flat grade, and no pedestrian activity), the 1985 HCM recommends 1,800 vphgpl as the saturation flow rate of a traffic lane at a signalized intersection. This value corresponds to a minimum discharge headway of 2.0 sec/veh. More recent research, such as that by Lee and Chen (*5*) and Zegeer (*7*), suggests that the ideal minimum discharge headway may be shorter than 2.0 sec/veh. Although Lee and Chen do not specifically calculate a minimum discharge headway for their data set, the average of the 5th through 10th headways that they reported is 1.97 sec/veh. Similarly, Zegeer (*7*) found an ideal minimum discharge headway of 1.92 sec/veh.

Headway Models

The discharge headway between successive vehicles has been described by Drew (*8*) in terms of a time-space diagram, as shown in Figure 2. The curved lines in Figure 2 represent the trajectories of individual vehicles as they travel through the intersection. The curved portions of each trajectory represent the acceleration or deceleration of the individual vehicles. As each successive vehicle crosses the stop line, its speed increases and its headway decreases. At a point after the fourth or fifth vehicle, the speed of each vehicle crossing the stop line becomes constant and, as a result, so do the headways between vehicles.

A deterministic model of the headway process based on the trajectories shown in Figure 2 has been described by Briggs (*9*). His model, which is based on the assumption that queued vehicles accelerate at a constant rate, has the following form.

If $n * d < d_{max}$, then

$$h_n = T + \sqrt{\frac{2 * d * n}{A}} - \sqrt{\frac{2 * d * (n - 1)}{A}} \quad (1)$$

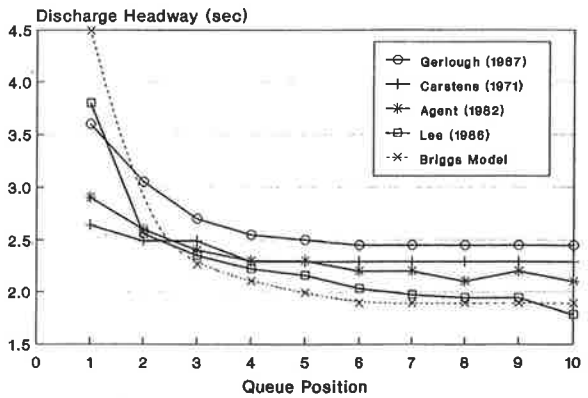


FIGURE 1 Comparison of past studies of queue discharge headway.

Civil Engineering Department, University of Nebraska-Lincoln, Lincoln, Nebr. 68588-0531.

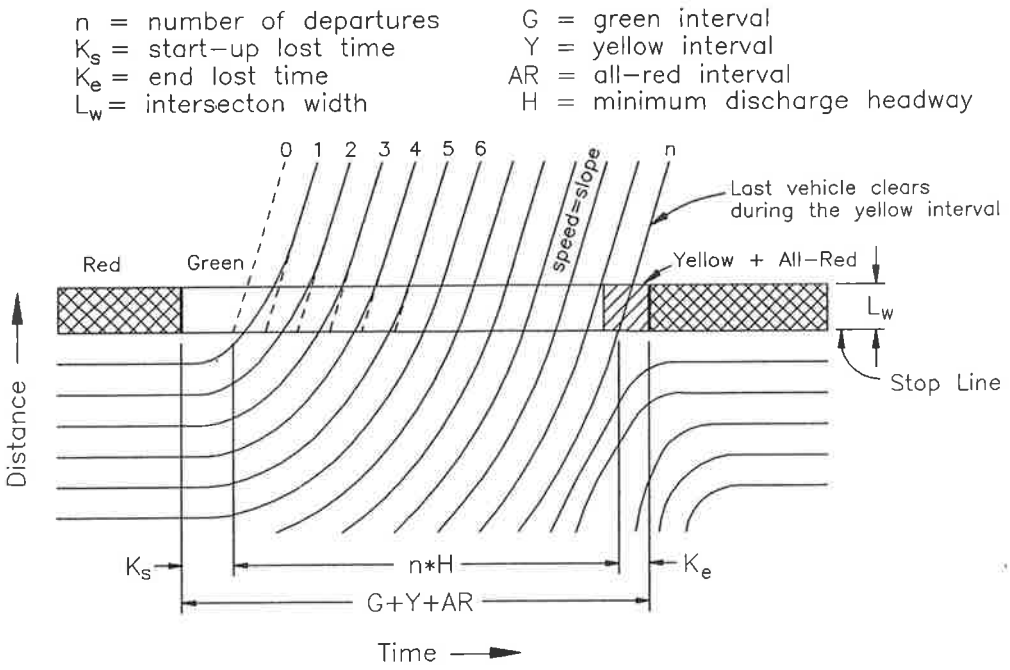


FIGURE 2 Time-space relationship of the queue discharge process.

Otherwise

$$h_n = T + \frac{d}{V_q} \tag{2}$$

with

$$d_{\max} = \frac{V_q^2}{2 * A} \tag{3}$$

where

h_n = headway of the nth queued vehicle (sec),
n = queue position (n = 1, 2, 3, . . .),
d = distance between vehicles in a stopped queue (ft),
V_q = desired speed of queued traffic (fps),
$d_{\max}$ = distance traveled to reach speed V_q (ft),
T = driver starting response time (sec), and
A = constant acceleration of queued vehicles (fpss).

Briggs calibrated his model using data from five previous studies conducted in the United States and Germany by other researchers. The parameters yielding the best fit were T = 1.22 sec, A = 3.67 fpss, d = 19.65 ft for each queued vehicle, and V_q = 29.4 fps.

The model has two parts. The part to use depends on whether the vehicle speed at the stop line has reached the desired speed (V_q). For the first few queue positions, vehicle speed is less than V_q and headway is a function of acceleration and queue position. However, after vehicles reach the desired speed (i.e., $n * d \geq d_{\max}$), headways become dependent only on driver response time and desired speed. Thus, this model indicates that headways become essentially constant after the desired speed is reached.

The predictive ability of Briggs's headway model is compared with the results of previous headway studies in Figure 1. As this figure indicates, Briggs's model yields a relatively good fit to the data and appears to explain the trend toward decreasing headways with queue position.

Starting Response Time and Distance Between Queued Vehicles

Driver starting response time and the distance between vehicles in a stopped queue at signalized intersections have been the subject of several previous studies (*10–12*). Messer and Fambro (*10*) found that driver response was fairly constant at 1.0 sec, regardless of queue position. The only exception was with the driver in the first queue position, who had an additional delay of 2.0 sec. The shorter response time of the second and subsequent queued drivers is probably due to their ability to anticipate the time to initiate motion by seeing the signal change or the movement of vehicles ahead, or both. Messer and Fambro also found that the average length of roadway occupied by each queue position is about 25 ft.

Another study of driver response time was conducted by George and Heroy (*11*). They found driver response to be relatively constant at about 1.3 sec for all queue positions. However, further examination of their data suggests that the first driver's response time was slightly longer, at about 1.5 to 2.0 sec.

Response times in the preceding studies were all measured at the start of vehicle motion. Herman et al. (*12*) found that

driver response to disturbance (including the start of motion) remained fairly constant as the platoon of queued vehicles increased its speed. In particular, they found that the speed of propagation of the response wave was relatively constant at about 26 fps up to platoon speeds of 30 fps. Beyond this speed, the response wave began to slow down as speeds neared the final cruising speed. A constant speed of propagation implies that all vehicles in the queue have the same trajectory, which supports a fundamental premise of Briggs's headway model. The authors also found the average distance between stopped vehicles to be 25.9 ft. Using this value, the starting response time can be calculated as 1.0 sec (= 25.9/26).

MODEL DEVELOPMENT

The most restrictive assumption in Briggs's model is that of constant acceleration. Experience suggests that drivers vary their acceleration as they increase their speed. Thus, the objective of this section is to develop an alternative headway model based on a nonconstant acceleration behavior.

Driver Acceleration Model

An early study of driver acceleration characteristics on freeway on-ramps was conducted by Buhr et al. (*13*). On the basis of their study of passenger cars undergoing "normal" acceleration from a stopped condition at the ramp entrance, the authors determined that acceleration decreased linearly with increasing speed. The model they proposed was

$$a = A_{max} * \left(1 - \frac{V}{V_{max}}\right) \quad (4)$$

where

a = instantaneous acceleration (fpss),
V = velocity of vehicle (fps),
A_{max} = maximum acceleration (fpss), and
V_{max} = maximum speed corresponding to zero acceleration (fps).

For on-ramps on level terrain, the authors found that parameter values of A_{max} equal to 15 fpss and V_{max} equal to 60 fps would describe the acceleration behavior of ramp drivers.

A study of the acceleration and speed characteristics of queued drivers departing from a stop line was conducted by Evans and Rothery (*14*). The average speed and acceleration for each queue position as observed by Evans and Rothery are shown in Figure 3. As the figure indicates, the average speed found for each queue position increases exponentially to an average "desired" speed of about 50 fps.

The apparent desired speed of 50 fps is somewhat lower than the reported speed limit of 45 mph (66 fps). The reason for this difference is not clear from the authors' paper; however, it is probably attributable to the constraining effect of queued flow conditions. Under this assumption, higher speeds would be reached only as vehicles begin to increase their spatial separation and transition from a queued flow regime to a free flow regime downstream of the intersection.

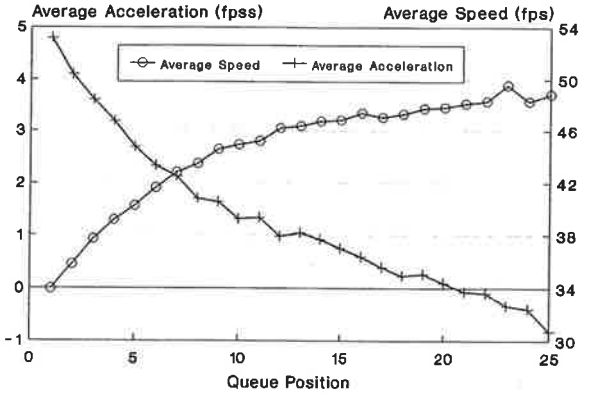


FIGURE 3 Acceleration and speed of queued vehicles at an intersection as observed by Evans and Rothery.

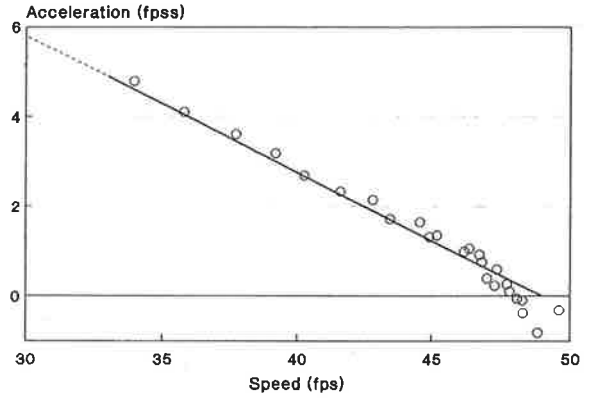


FIGURE 4 Acceleration versus speed relationship of queued vehicles as observed by Evans and Rothery.

The relationship between acceleration and speed for each queue position is shown in Figure 4, which indicates a strong linear relationship between acceleration and speed. The negative accelerations (i.e., decelerations) accompanying the higher speeds are probably a manifestation of a speed correction effect that stems from queued flow behavior, as observed by Herman et al. in an earlier study (*12*).

The acceleration model in Equation 4 can be rewritten in terms of a differential equation of motion to derive other relationships between speed, acceleration, distance, and time. In particular, integral calculus can be used to determine the following speed-time and distance-time relationships:

$$a = \frac{\delta V}{\delta t} = A_{max} * \left(1 - \frac{V}{V_{max}}\right) = A_{max} - B * V \quad (5)$$

$$V = \frac{A_{max}}{B} * (1 - e^{-B*t}) + V_0 * e^{-B*t} \quad (6)$$

$$x = \frac{A_{max}}{B} * t - \frac{A_{max}}{B^2} * (1 - e^{-B*t}) + \frac{V_0}{B} * (1 - e^{-B*t}) \quad (7)$$

where

a = instantaneous acceleration (fpss),
V = velocity of vehicle (fps),
t = time of acceleration (sec),
A_{max} = maximum acceleration (fpss),
V_{max} = maximum speed (fps),
$B = A_{max}/V_{max}$
V_0 = initial velocity at time $t = 0$ (fps), and
x = acceleration distance (ft).

Discharge Headway Model

In general, a vehicle's time of arrival at the stop line is composed of two time increments. The first increment is the time measured from the beginning of the signal phase to the instant when the driver first begins motion. This increment will be called the cumulative starting response time, and, as previous studies have indicated (*10–12*), it can be estimated for each queue position (n) as $\tau + n * T$, where T is the starting response time for an individual driver and τ is the additional response time of the first driver.

The second increment represents the time needed to accelerate over the distance between the stopped vehicle and the stop line (t_a). As in the approach of Briggs, the distance occupied by each queued vehicle is assumed equal to d ft. Thus, the back axle of the first vehicle is d ft back from the stop line, the second axle is $2 * d$ ft back, and the nth vehicle is $n * d$ ft back. The time for the nth queued vehicle to reach the stop line once it starts (i.e., $V_0 = 0$) can then be expressed using Equation 7 as

$$x_n = n * d = \frac{A_{max}}{B} * t_{a(n)} - \frac{A_{max}}{B^2} * (1 - e^{-B*t_{a(n)}}) \quad (8)$$

Examination of Equation 8 indicates that a closed-form solution for $t_{a(n)}$ is not obtainable. However, substitution of Equation 6 into Equation 8 will replace the exponential term with a term representing stop line speed (V_{sl}) and thereby yield a solution for $t_{a(n)}$:

$$t_{a(n)} = \frac{n * d}{V_{max}} + \frac{V_{sl(n)}}{A_{max}} \quad (9)$$

The discharge headway (h) between the nth and $(n-1)$th vehicles can be calculated as the difference between their stop line arrival times:

$$h_n = \tau * N_1 + (n * T + t_{a(n)}) - [(n - 1) * T + t_{a(n-1)}] \quad (10)$$

Finally, substituting Equation 9 into Equation 10 and simplifying leads to the proposed headway model:

$$h_n = \tau * N_1 + T + \frac{d}{V_{max}} + \frac{V_{sl(n)} - V_{sl(n-1)}}{A_{max}} \quad (11)$$

where

h_n = headway of the nth queued vehicle (sec),
T = driver starting response time (sec),
τ = additional response time of the first queued driver (sec),
N_1 = 1 if $n = 1$ or 0 if $n > 1$,
d = distance between vehicles in a stopped queue (ft),
$V_{sl(n)}$ = stop line speed of the nth queued vehicle (fps),
V_{max} = maximum speed (fps), and
A_{max} = maximum acceleration (fpss).

Comparing this model to that proposed by Briggs (Equations 1–3) reveals an important similarity. That is, as the speed of queued vehicles becomes constant ($n \rightarrow \infty$), the time headway between successive vehicles converges to the minimum discharge headway $H = T + d/V$, where V represents the ultimate speed of queued flow for both models. This similarity supports the argument that the V_{max} parameter of the acceleration model represents the average queued driver's desired speed.

Stop Line Speed Model

As discussed previously, Equation 6 can be used to relate stop line speed to the time of acceleration. Recognizing that time spent accelerating to the stop line (t_a) is dependent on queue position, a more useful model of stop line speed can be derived by substituting n for $t_{a(n)}$ and setting $V_0 = 0.0$, yielding the following result:

$$V_{sl(n)} = V_{max} * (1 - e^{-n*k}) \quad (12)$$

where

$V_{sl(n)}$ = stop line speed of the nth queued vehicle (fps),
V_{max} = common desired speed of queued traffic (fps),
$k = \beta/V_{max}$, and
β = empirical calibration constant.

The usefulness of this model stems from its empirical formulation, which allows it to be easily calibrated using stop line speed versus queue position data. The exponential form of the model suggests that the traffic queue never reaches V_{max}; however, this is more of a theoretical anomaly than a practical limitation.

Start-Up Lost Time Model

As indicated by Figure 1, the first few vehicles in a traffic queue experience headways in excess of the minimum discharge headway. Any discharge time in excess of the minimum headway is essentially unused, or lost, time. Moreover, the sum of the lost times for the first few queue positions is called start-up lost time (K_s), which can be calculated as

$$K_s = \sum_{n=1}^{N} (h_n - H) \quad (13)$$

where N is the number of vehicles crossing the stop line that have speeds less than V_{max}. Substitution of Equation 11 for h_n and $T + d/V_{max}$ for H and elimination of the summation yield the proposed start-up lost time model:

$$K_s = \tau + \frac{V_{max}}{A_{max}} \quad (14)$$

The start-up lost time predicted by Equation 14 represents a limiting value as $N \rightarrow \infty$. Because the queue theoretically never reaches V_{max}, this suggests that all queue positions incur some lost time. This is in contrast to the method suggested by the 1985 HCM, wherein only the first four vehicles are assumed to have headways in excess of the minimum discharge headway. Thus, Equation 14 is not limited by an a priori assumption as to which queue position first achieves the minimum discharge headway.

EXPERIMENTAL DESIGN

Study Sites

Three SPUIs and two at-grade intersections (AGIs) were identified as candidate study sites. All three SPUIs are located on an 8-mi section of US-19 in the Tampa, Florida, area. Two of the SPUIs have been in operation for more than 15 years each, and the third has been open to traffic for about 6 months (at the time of the study).

Because one objective of this research was to identify the effect of SPUI geometry on selected traffic characteristics, it was decided that the field study should also include two typical, high-type AGIs for statistical control. One of the two AGIs selected for study is located about 1 mi from the other SPUIs (at SR-60 and Belcher Road). The second AGI selected for study is located on Wellborn Road in College Station, Texas.

All four Florida sites have actuated control, whereas the Texas site has pretimed timing plans downloaded from a centralized computer by time of day. Cycle lengths at the SR-60 SPUI, SR-694 SPUI, and the Wellborn AGI ranged from 90 to 100 sec; cycle lengths at the SR-686 SPUI and the Belcher AGI ranged from 120 to 140 sec. Traffic demands were well below capacity, as indicated by the volume-to-capacity ratios, which ranged from 0.30 to 0.70.

Data Collection System

The computerized data collection system used for the field studies relied on a series of sensors located around the junction. One type of sensor used to monitor vehicular motion was the tapeswitch. Another type of sensor used was the photocell. Photocells were connected to the load switch LEDs inside the traffic signal controller cabinet and used to monitor the status of the signal indications.

These sensors were monitored by an Environmental Computer (EC) manufactured by the Golden River Corporation. The EC has the capability to check the status of each sensor every 1/600 sec.

Field Study Procedure

The data collection system was used to record discharge headway, speed, and acceleration data for the cross road through movement at the SPUIs and for the major road through movement at the AGIs. Left-turn movements were also studied for this research (*1*); however, only the findings for the through movements will be discussed here.

To collect the headway data for this study, a tapeswitch was located just past the point on the interchange approach where vehicles most frequently stopped (usually the stop line). A second tapeswitch was installed a known distance from the stop line tapeswitch along the study movement's travel path. The distance to this switch ranged from 50 to 150 ft, depending on the particular site.

Data Reduction Procedure

The technique used to calculate stop line speed and acceleration was based on a procedure described by Evans and Rothery (*14*). In general, this technique assumes that a vehicle's acceleration is constant between the two tapeswitches, which form a "trap" of known length (D). The assumption of a constant acceleration allows the vehicle trajectory to be modeled by a second-order equation of distance (x) as a function of time (t). In fact, this assumption allows two equations to be written, one for each axle of the vehicle. This relationship is shown in Figure 5 in terms of the vehicle's trajectory in time and space.

As shown in Figure 5, the trajectories of each axle are identical but separated by a distance equal to the vehicle wheelbase (L). When a vehicle crosses the trap, it causes four event times to be recorded: t_1, the time the first axle hits the first switch; t_2, the time the first axle hits the second switch; t_3, the time the second axle hits the first switch; and t_4, the time the second axle hits the second switch. Subtracting t_1 from each of these event times yields the relative travel time

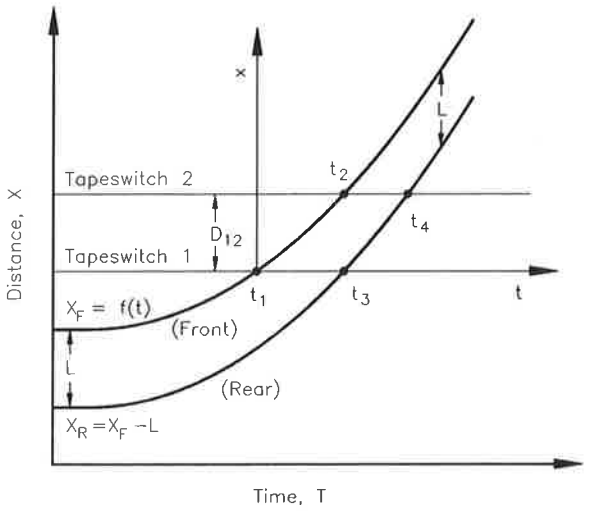


FIGURE 5 Time-space trajectory of a vehicle as it traverses the tapeswitch trap.

of the front and back axles through the trap. By setting $x = 0$ when $t = 0$ for the transformed data, the trajectory of the front and back axles can be described by the following second-order equation:

$$x = b_1 * t + b_2 * t^2 \quad (15)$$

where x and t correspond to the remaining three time-event pairs: $(x = D, t = t_2)$, $(x = L, t = t_3)$, and $(x = D + L, t = t_4)$. Using these values of x and t, three equations can be written to solve for the three unknowns b_1, b_2, and L. Solving the three equations for b_1, b_2, and L yields

$$b_2 = \frac{D * (t_2 - t_4 + t_3)}{t_2 * (t_4 - t_3) * (t_4 + t_3 - t_2)} \quad (16)$$

$$b_1 = \frac{D}{t_2} - b_2 * t_2 \quad (17)$$

$$L = b_1 * t_3 + b_2 * t_3^2 \quad (18)$$

Finally, differentiating the second-order equation gives the stop line speed and acceleration as $V_{sl} = b_1$ and $A = 2 * b_2$, respectively.

STATISTICAL ANALYSIS AND MODEL CALIBRATION

Regression techniques were used to calibrate the proposed models. For the acceleration, stop line speed, start-up lost time, and discharge headway models, sufficient data were collected to use half for model calibration and the other half for validation. The regression models were based on the models developed in preceding section; however, covariates found to be significant from an analysis of variance (ANOVA) were also included in the final model form. The statistical analysis was conducted using the SAS system (*15*).

Driver Acceleration Model

To calibrate the acceleration model, passenger car acceleration and speed data were collected for one through movement at each of the five study sites. All total, acceleration and speed data were collected for 4,820 through vehicles. The relationship between speed and acceleration is shown in Figure 6. The data points represent the observed acceleration and speed averaged by queue position.

As Figure 6 indicates, a relatively strong linear relationship exists between acceleration and speed. The strength of the linear relationship between speed and acceleration was tested using least-squares linear regression techniques (*1*). This analysis also indicated that each site had a maximum acceleration (A_{max}) in the relatively narrow range of 6.0 to 8.0 fpss. Further statistical analysis indicated that these maximum accelerations were not significantly different from their average value of 6.63 fpss ($p = 0.10$). This average value is similar to the 6.0 fpss found by Evans and Rothery (*14*) (see Figure 4).

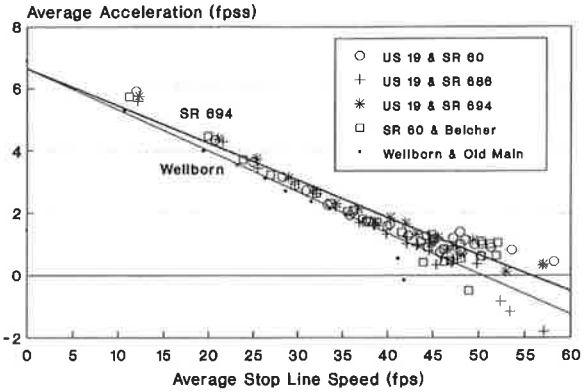


FIGURE 6 Acceleration as a function of stop line speed.

Stop Line Speed Model

The relationship between speed and queue position is shown in Figure 7, which indicates that the speeds of the first few queue positions increase rapidly but tend to reach a maximum at later positions. This trend is consistent with the exponential form of the stop line speed model (Equation 12). The maximum speed obtained from this analysis represents the best estimate of the common desired speed of queued traffic (V_{max}), as required by the discharge headway model.

Closer examination of the desired speeds in Figure 7 indicates a relatively constant value for four of the five through movements. In particular, the desired speed at these four sites is in the relatively narrow range of 46.7 to 51.0 fps with a median value of about 49 fps. The desired speed of 39.9 fps found at the Wellborn AGI is well below this range. A possible explanation for the lower desired speed at this AGI is the driver's awareness of the relatively near (about 1,000 ft) downstream signalized intersection and of the likelihood of encountering stopped, turning, or weaving vehicles before reaching a higher speed. The nearest downstream intersection

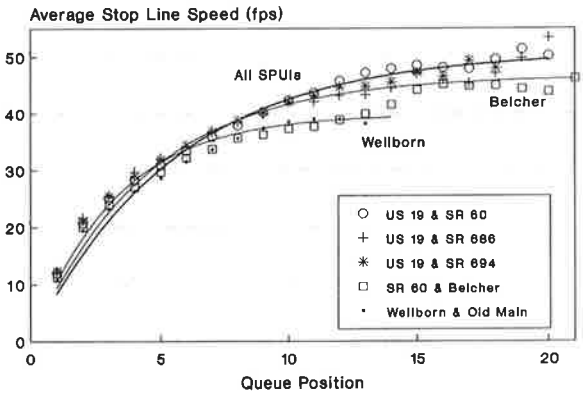


FIGURE 7 Stop line speed as a function of queue position.

at the other study sites was much more distant than at the Wellborn AGI.

Although intuition suggests that there should be some correlation between speed limit and the desired speed of through traffic, a significant relationship was not found at the five study sites ($p = 0.32$). The lack of significance found in this analysis may be partly attributed to other influential factors (e.g., downstream effects) and to the relatively small sample size. As a result, it cannot be concluded from this analysis that desired speed is unaffected by speed limit at all interchanges and intersections. This analysis will only support the use of 49 fps as the best estimate of the desired speed of a through movement. However, V_{max} was found to vary as a function of radius for the left-turn movements that were also studied for this research (*1*).

As a result of the calibration process, the stop line speed model is modified slightly from that originally described in Equation 12. The revised, calibrated model is

$$V_{sl(n)} = V_{max} * (1 - e^{-n*k}) \quad (19)$$

where

$V_{sl(n)}$ = stop line speed of the nth queued vehicle (fps),
$k = -0.290 + 24.0/V_{max}$,
n = queue position ($n = 1, 2, 3, \ldots$), and
V_{max} = common desired speed of queued traffic (fps).

The ability of this model to predict the observed speeds (as averaged by queue position) is shown in Figure 7. As the figure indicates, the model was a relatively good predictor of stop line speed ($R^2 = 0.9$) (*1*).

Discharge Headway Model

To calibrate the discharge headway model, passenger car headways were collected for one through movement at each of the five study sites. Headways were collected for 12,053 through vehicles at the five study sites. The average headway for each site, movement studied, and queue position is provided elsewhere (*1*).

Calibration of the discharge headway model was based on a linear regression of discharge headways averaged by site and queue position. Because the number of headways recorded varied widely among these factors, the headway data were averaged to remove the bias that an unequal sample size would have on model parameters. Because of this technique, the statistics used to assess model fit to the data (i.e., standard deviation and R^2) do not reflect the total variability in individual driver headways. Rather, the statistics indicate the ability of the model to predict the average discharge headway by queue position.

The calibrated discharge headway model for through movements is

$$h_n = \tau' * N_1 + T' + \frac{d'}{V_{max}} + b_3 * \left(\frac{V_{sl(n)} - V_{sl(n-1)}}{A_{max}}\right) + b_4 * v + b_5 * AGI \quad (20)$$

The model parameters are

τ' = regressed additional response time of the first queued driver (sec),
T' = regressed driver starting response time (sec),
d' = regressed distance between vehicles in a stopped queue (ft),
$V_{sl(n)}$ = stop line speed of the nth queued vehicle (fps),
V_{max} = common desired speed of queued traffic (fps), and
A_{max} = maximum acceleration (fpss).

The model variables are

h_n = headway of the nth queued vehicle (sec),
n = queue position (N = 1, 2, 3, . . .),
v = traffic pressure (veh/cycle/lane),
N_1 = indicator variable (1 for first queue position; 0 for all others), and
AGI = indicator variable (1 for AGI; 0 for SPUI).

Parameter	*Parameter Value*	t-*statistic*
$\tau'(b_0)$	1.03	17.7
$T'(b_1)$	1.57	4.6
$d'(b_2)$	25.25	1.7
b_3	0.357	8.2
b_4	−0.0086	1.6
b_5	−0.23	4.3

Variable	*Minimum Value*	*Maximum Value*
n	1	18
v	0.0	16.8
h	1.6	3.8

Observations: 164
Std. Deviation: 0.16
R^2: 0.88

In general, the parameter values for b_0, b_1, and b_2 are consistent with the definitions of the theoretical model parameters to which they correspond (i.e., τ, T, and d, respectively). As these values do not, however, represent actual measurements, a prime symbol (′) has been added to each model parameter to denote that its value was established using regression analysis and that the relationship between this value and its definition may be distorted.

The ANOVA of individual headways revealed that traffic pressure, as measured by lane volume per cycle, was significant in reducing discharge headway ($p = 0.001$). In using this component of the headway model, a one-to-one relationship between the duration of the volume average and the predicted average headway must be maintained. In other words, a total lane volume representing a 1-hr average should be used in the regression model to predict discharge headways during the same 1-hr period.

In general, the significance of the calibrated parameter values combined with the theoretical basis of the discharge headway model suggests that the model adequately describes the headway process of queued vehicles. This ability is demonstrated in Figure 8, in which the calibrated model is compared with the data for two movements.

The exponential form of the stop line speed model implies that V_{max} (and the minimum discharge headway) will, theoretically, never be reached. An examination of Figure 8, how-

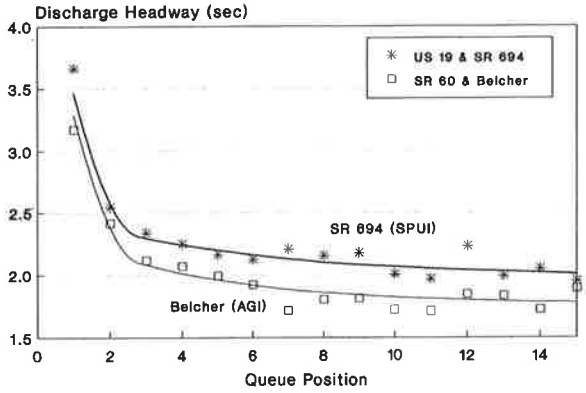


FIGURE 8 Discharge headway as a function of queue position.

ever, indicates that the predicted discharge headway does reach a relatively constant, minimum value after the eighth or ninth queue position. The likelihood that a minimum value is not reached until at least the eighth or ninth queue position suggests that the 1985 HCM's method for estimating the minimum discharge headway (i.e., average headways of the fifth through last queue positions) may be biased because it includes queue positions that have not achieved a minimum headway.

This potential bias-by-queue-position in the 1985 HCM method has implications for capacity evaluation and statistical analysis of cause and effect. Application of the HCM method will probably result in an estimate of minimum headway that is longer than that ultimately achieved by the traffic queue. Thus, the capacity of high-demand movements may be underestimated using the HCM method. More important, the trend in headway studies toward observing more headways for the lower queue positions tends to magnify any bias-by-queue-position. Thus, a statistical analysis of cause and effect (e.g., lane width, grade) may be clouded, and perhaps misdirected, by the added variance introduced by unequal numbers of headways observed at each queue position among the study sites.

Minimum Discharge Headway Model

According to the headway model (Equation 11), a minimum discharge headway (H) is not reached until the queue reaches its desired speed (V_{max}). At this point, the difference in stop line speed of successive vehicles is zero and the minimum discharge headway becomes

$$H = T + \frac{d}{V_{max}} \tag{21}$$

On the basis of the regression results, the calibrated minimum discharge headway model is

$$H = 1.57 + \frac{25.25}{V_{max}} - 0.0086 * v - 0.23 * \text{AGI} \tag{22}$$

where

H = minimum discharge headway for a through movement (sec/veh),
V_{max} = common desired speed of queued traffic (fps),
v = traffic pressure (veh/cycle/lane), and
AGI = 1 if the movement is at an AGI and 0 if it is at a SPUI.

Equation 22 implies that an AGI with a common desired speed of 49 fps and a nominal traffic pressure of 5.0 veh/cycle/lane would have a minimum discharge headway of 1.81 sec/veh. This value suggests that the ideal minimum headway may actually be shorter than the 2.0 sec/veh recommended by the 1985 HCM.

Start-Up Lost Time Model

The theoretical start-up lost time is calculated from the calibrated minimum discharge headway model and Equation 14 as

$$K_s = 1.03 + 0.357 * \frac{V_{max}}{A_{max}} \tag{23}$$

where K_s is start-up lost time for a through movement (sec/phase) and A_{max} is maximum acceleration (equal to 6.63 fpss). The start-up lost time for a typical through movement with V_{max} of 49 fps and A_{max} of 6.63 fpss is 3.67 sec. This value suggests that start-up lost time for lengthy traffic queues may be greater than the 2.0 sec suggested by the 1985 HCM (*6*, Chapter 2).

In theory, the values for H and K_s from Equations 22 and 23, respectively, represent limiting values for infinitely large queues. Alternative forms of these equations could be derived to predict the average headway and lost time for queues of a more practical size. However, the comparative use of Equation 20 (h_n) versus Equations 22 (H) and 23 (K_s) to predict service time (T_N) (i.e., $T_N = \Sigma_{n=1}^{N} h_n$ versus $T_N = N * H + K_s$) suggests that this added rigor is not necessary. In general, using $N * H + K_s$ yields service times that are only about 0.7 sec (at $N = 6$) and 0.2 sec (at $N = 12$) longer than would be determined by summing the individual headways. Thus, it appears that there is little to be gained by using more complicated model forms, such as $T_N = \Sigma h_n$, to predict service time or phase capacity.

CONCLUSIONS

The examination of driver acceleration indicated a linear trend toward decreasing acceleration with increasing vehicle speed. Both the initial acceleration and ultimate desired speed of these drivers were essentially constant among sites (i.e., $A_{max} = 6.63$ fpss and $V_{max} = 49$ fps) for through traffic movements.

A linear relationship between speed and acceleration implies an exponential increase in vehicle speed with time. For this research, the exponential speed-time relationship was extended to the modeling of queued vehicle speed at the stop line. On the basis of the results of this research, the speed of

queued vehicles was found to agree closely with the exponential model form. The calibrated stop line speed model indicated that drivers do not reach a practical maximum speed until the eighth or higher queue positions (although, theoretically, the exponential form implies that the maximum speed is never attained).

The minimum discharge headway of a traffic movement is a complex process that is dependent on driver response time, desired speed, and traffic pressure. The discharge headway model developed in this research indicates that practical values of the minimum discharge headway of a traffic movement are not reached until the eighth or higher queue positions. Application of this model suggests that the minimum discharge headway of a traffic movement under ideal conditions may be shorter than 2.0 sec/veh and that its corresponding start-up lost time may be longer than 2.0 sec.

ACKNOWLEDGMENTS

The author recognizes the agencies responsible for sponsoring the research project that provided the data for this study. In particular, the data were collected as part of a study titled Single Point Urban Interchange Design and Operations conducted by the Texas Transportation Institute and sponsored by the National Cooperative Highway Research Program, Washington, D.C. The author is particularly grateful to the faculty members at Texas A&M University who served on his Ph.D. committee: Carroll Messer, Daniel Fambro, Raymond Krammes, and Martin Wortman.

REFERENCES

1. J. A. Bonneson. *Operational Characteristics of the Single-Point Urban Interchange*. Ph.D. dissertation. Texas A&M University, College Station, Tex., 1990.
2. D. L. Gerlough and F. A. Wagner. *NCHRP 32: Improved Criteria for Traffic Signals at Individual Intersections*. HRB, National Research Council, Washington, D.C., 1967.
3. R. L. Carstens. Some Traffic Parameters at Signalized Intersections. *Traffic Engineering*, Aug. 1971, pp. 33–36.
4. K. R. Agent and J. D. Crabtree. *Analysis of Saturation Flow at Signalized Intersections*. Report UKTRP-82-8. Kentucky Transportation Research Program, University of Kentucky, Lexington, Ky., 1982.
5. J. Lee and R. L. Chen. Entering Headway at Signalized Intersections in a Small Metropolitan Area. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986, pp. 117–126.
6. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
7. J. D. Zegeer. Field Validation of Intersection Capacity Factors. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986, pp. 67–77.
8. D. R. Drew. Design and Signalization of High-Type Facilities. *Traffic Engineering*, July 1963, pp. 17–25.
9. T. Briggs. Time Headways on Crossing the Stop Line After Queueing at Traffic Lights. *Traffic Engineering & Control*, May 1977, pp. 264–265.
10. C. J. Messer and D. B. Fambro. Effects of Signal Phasing and Length of Left-Turn Bay on Capacity. In *Transportation Research Record 644*, TRB, National Research Council, Washington, D.C., 1977, pp. 95–101.
11. E. T. George and F. M. Heroy. Starting Response of Traffic at Signalized Intersections. *Traffic Engineering*, July 1966, pp. 39–43.
12. R. Herman, T. Lam, and R. W. Rothery. The Starting Characteristics of Automobile Platoons. *Proc., 5th International Symposium on the Theory of Traffic Flow and Transportation*, American Elsevier Publishing Co., New York, 1971, pp. 1–17.
13. J. H. Buhr, R. H. Whitson, K. A. Brewer, and D. R. Drew. Traffic Characteristics for Implementation and Calibration of Freeway Merging Control Systems. In *Highway Research Record 279*, HRB, National Research Council, Washington, D.C., 1969, pp. 87–106.
14. L. Evans and R. W. Rothery. Influence of Vehicle Size and Performance on Intersection Saturation Flow. *Proc., 8th International Symposium on Transportation and Traffic Theory*, University of Toronto Press, Toronto, Ontario, Canada, 1981, pp. 193–222.
15. *SAS/STAT User's Guide*. Release 6.03 edition. SAS Institute Inc., Cary, N.C., 1988.

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

Signal Timing Determination Using Genetic Algorithms

MARK D. FOY, RAHIM F. BENEKOHAL, AND DAVID E. GOLDBERG

The implementation of a genetic algorithm (GA) (an artificial intelligence technique) to produce optimal or near-optimal intersection traffic signal timing strategies is described. The focus is on examining this application within a simple traffic situation, giving the reader a clear understanding of how the genetic algorithm is used. The problem involves finding a signal timing strategy that produces the smoothest traffic flow with the least average automobile delay. The problem domain has many tentative solutions. Therefore, signal timing design is expected to benefit from the parallel, global, and robust search characteristics of GAs. This gain is realized on a simulated four-intersection traffic network in the current implementation. The GA, by considering how traffic moves among multiple intersections (through simulation), can find a logical, near-optimal timing configuration. When this timing configuration is used in the corresponding real-world traffic situation, minimal total automobile delay is expected.

Many motorists are frustrated with traffic signal timings and believe that they can be greatly improved to allow better traffic flow. This is where computers can be useful in the traffic environment. Computers can improve signal timings and therefore improve travel times, personal attitudes, fuel efficiency, pollution, and safety.

This paper presents ideas relating to the use of computers in an automobile traffic environment, specifically ideas to achieve demand-responsive control. The focus is on the use of a genetic algorithm (GA) to control traffic signals and the benefits that can be attained from its use. The implementation discussed in this paper, which uses a GA to control traffic signals, will be called the Traffic GA.

The goal of the Traffic GA is to find near-optimal traffic signal-timing strategies. To achieve this goal, a simplistic traffic flow simulation model was used. The traffic simulation model is sufficient for the purposes of this study; however, it is not intended to be immediately ready for real-world use (i.e., capacity analysis, comparison of actual versus computed delay, etc.). On the other hand, it is possible to improve the current simulation or insert another, more realistic simulation model into the existing GA and use this system to find near-optimal timing strategies by the techniques described in this paper. By making the simulation model more realistic, it would be possible to compare actual traffic conditions with the Traffic GA's simulation module. A simple simulation model was used because the focus of this research was on the application of a GA to improve traffic flow, not the design of a new, more realistic simulation system. Efforts to make the simulation more realistic are essential in model calibration. However, this paper does not deal with these issues.

University of Illinois at Urbana-Champaign, Urbana, Ill. 61801.

MOTIVATIONS

Traffic control today is in need of an intuitive, robust system to continually optimize traffic signal timings. Intelligent computer traffic control systems are needed to dynamically handle changing traffic conditions.

In pretimed controllers, traffic signal timings are fixed at what is determined to be the most effective timing strategy. Timing determination involves either extensive analysis of traffic data or observations of traffic trends, making it a fairly time-consuming task. Because of the time constraints, timing determinations are done infrequently, making the pretimed control method a static model. Therefore, with this method, signal cycle times and offset times are calculated once, for current conditions, and are then set into the individual traffic signals for an extended period of time (i.e., months). The signal cycle timings do not change with demand. This static characteristic is a clear disadvantage and motivates the development of more dynamic methods.

Demand-responsive controllers offer more flexibility than pretimed controllers because traffic signals can have their timings adjusted on the basis of current demand. In addition, flexibility is gained from the capability of each signal of gathering data continually and automatically, allowing continuous analysis of current situations. In the network of demand responsive intersections, a central computer is necessary to (*a*) read traffic data continuously from the entire network, (*b*) process the network data to produce traffic signal timings for all network intersections, and (*c*) operate the traffic signals in a demand-responsive mode. These applications allow an intersection's traffic signal times to be calculated using data from that intersection as well as from adjacent intersections since all data are aggregated in the central computer.

The commonly known full-actuated and semiactuated traffic controllers, as well as the traffic-adaptive control approaches suggested by Gartner (*1*) and Lin (*2*), are all grouped in the demand-responsive category. In this category of traffic controllers, the signal timings are changed depending on the demand, although the nature of these changes is different. The actuated controllers and the traffic-adaptive approach are based on the traditional programming methods, whereas other approaches, like the one described in this paper, use artificial intelligence (AI) techniques.

One of the advantages of AI techniques is that they can be easily designed to perform demand-responsive control on a network of intersections. This paper concentrates on a genetic optimization search algorithm, called a GA. Other applications of AI to intersection traffic control are given elsewhere (*3*,*4*).

When considering a large number of multiphase traffic signals, the number of possible traffic signal-timing strategies can be very large. For example, for a network of 100 intersections, with a cycle length varying from 30 to 150 sec, the number of phases varying from 2 to 5, and the green time allocations varying at increments of 1 sec, the number of possible signal settings is enormous. If a search for the best timing strategy is repeated every few minutes to update the signal settings and a blind search method were used, the number of computations could easily become prohibitive. On the other hand, an intelligent search and optimization system should be able to avoid nonoptimal regions and learn from its past experiences. This should reduce the number solutions searched and allow the system to converge to a near-optimal solution in much less time. In addition, such a system can be put online to overcome some of the limitations of traditional signal optimization techniques.

The question now is whether GAs can find near-optimal signal-timing strategies that improve traffic flow. An answer to this question will be given for a small test problem consisting of a four-intersection street network, but first a more detailed description of GAs will be given.

DESCRIPTION OF GAs

GAs are algorithms that search by manipulating populations of structures (i.e., binary strings representing data structures that symbolize possible solutions to a problem) into new solution populations using operators patterned after natural genetic operations. These operators may include reproduction, crossover, mutation, and others. The three simple GA operators will be discussed later.

GA components can be split into two parts: application-dependent components and application-independent components (such as the GA operators described later). GAs only require two application-dependent components: a procedure to encode bit strings (chromosomes) into solutions to the problem and an evaluation function that will accept a solution to a problem and evaluate its fitness or rating (this function is often called a black box because the GA does not need to know anything specific about this function). The evaluation function, which is also called the fitness function, is similar to the objective function in traditional search problems. Its purpose is to give the GA a numerical evaluation of a possible solution in the same way that an objective function gives a numerical evaluation of a point in space. A GA uses an evaluation function to locate an optimal solution.

GA Evolutionary Process

GAs begin with a population of randomly generated members. The GA then requests that each individual member in the population have its fitness evaluated. The evaluation is done in the fitness function, and the fitness value is returned to the GA. Once a GA has a completely evaluated population, the GA operates on these members to form a new population. This can be thought of as a generation of parents producing a generation of children. Although the new population contains characteristics of the old population, all the new members are different from the members of the last population, so all of its new members must now be evaluated. As this process continues with fitness evaluation and execution of GA operators, new generations of members are created. The new populations are generally more fit (that is, they have higher fitness values) than earlier populations because evolution favors stronger, more fit individuals. This characteristic can be better understood by examining the three basic GA operators.

The Three Simple GA Operators

The GA used in the project discussed here is a simple genetic algorithm consisting of the three basic GA operators.

First, reproduction is responsible for choosing the members that will be allowed to reproduce during the current generation. These members are selected on the basis of their fitness values. All reproduction operators are biased to choose higher-fitness members over lower-fitness members, so high fitness characteristics are passed on to future generations. After the required number of population members has been selected for reproduction (some duplicates in this selection probably will exist), the next operator, crossover, can proceed.

The crossover operator randomly selects two members (i.e., bit strings) from the new subpopulation. Then a location within these two bit strings is selected at random. The location is used as the swapping point for the two strings, that is, all bits to the right of this location on the first string are exchanged with all bits to the right of this location on the second string. For example, suppose the two following strings were selected for reproduction: String A = 00000000 and String B = 11111111. Then suppose the random bit location was selected as 5, causing the two strings to split after Bit 5. This would result in two new strings, String C = 00000111 and String D = 11111000. After the new population has been filled with crossed-over members, mutation can take place.

The mutation operator is simple: with a small probability, a bit will be selected within a string, and it will be flipped (i.e., a 0 would become a 1 and a 1 would become a 0). Then these final members make up the new population, and all old members, from before reproduction, are thrown out. Because we now have a new population with new members, each member must have its fitness evaluated so this evolutionary process can continue.

These are the three basic GA operators, but many variations on these and other operators exist. A description of other operators and further details about GAs are given elsewhere (*5–8*).

PROBLEM DESCRIPTION

The problem addressed in this study entails finding a near-optimal traffic signal timing configuration at all intersections given the current intersection characteristics. The current characteristics consist of the current number of cars at each lane of each intersection and the external arrival volumes. It is anticipated that after the GA converges, the output will be a near-optimal timing configuration for north/south green phase

and east/west green phase for the current conditions for all the intersections in the network. As will be discussed later in this paper, the Traffic GA can be run repeatedly (e.g., every 10 min), where each run takes the newest traffic data and produces new traffic signal timings that are better suited to the current traffic conditions. First, the inputs and outputs for running the Traffic GA will be defined.

Input and Output

The preceding perspective allows the problem of traffic control to be considered a function of two vectors. The first is

$$(\text{input.1}) = \begin{bmatrix} n_{111} \\ \cdot \\ \cdot \\ n_{ijk} \\ \cdot \\ \cdot \\ n_{443} \end{bmatrix} \quad (1)$$

where n_{ijk} is the number of cars on Lane k of Approach j of Intersection i (in this example, $i = 1$ to 4, $j = 1$ to 4, and $k = 1$ to 3. The second is

$$(\text{input.2}) = \begin{bmatrix} v_{11} \\ \cdot \\ \cdot \\ v_{ij} \\ \cdot \\ \cdot \\ v_{44} \end{bmatrix} \quad (2)$$

where v_{ij} is the arrival volume on Approach j of Intersection i. In this example, $i = 1$ to 4 and $j = 1$ to 4.

The result of evaluating these two vectors through the Traffic GA will be one integer value, one binary vector, and one real number vector.

$$(\text{output.1}) = [\text{tgt}] \quad (3)$$

where tgt is total green time given to each intersection for one full cycle. The same total green time is used for all intersections in the current Traffic GA, but there is no reason this cannot be changed.

$$(\text{output.2}) = \begin{bmatrix} d_1 \\ \cdot \\ \cdot \\ d_i \\ \cdot \\ \cdot \\ d_4 \end{bmatrix} \quad (4)$$

where d_i is the direction in which the first green phase will allow traffic to flow at Intersection i, either north and south or east and west. In this example, $i = 1$ to 4.

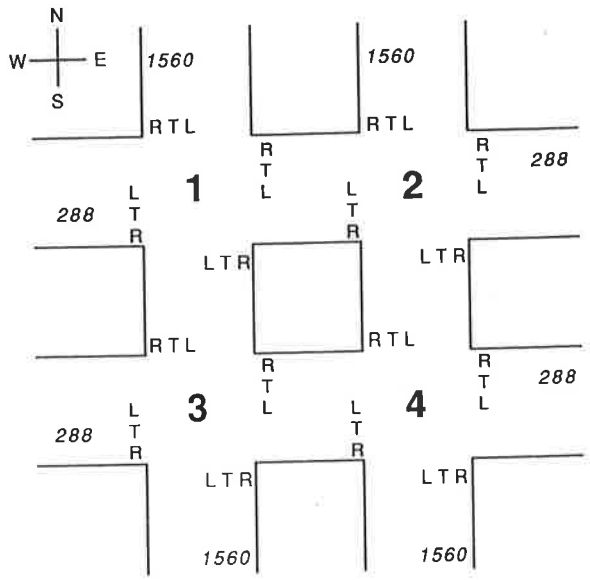


FIGURE 1 Street network configuration (intersections numbered from 1 to 4; L = turning left, R = turning right, T = going straight through). Numbers appearing at external components are arrival volumes in cars per hour.

$$(\text{output.3}) = \begin{bmatrix} \text{nsgt}_1 \\ \cdot \\ \cdot \\ \text{nsgt}_i \\ \cdot \\ \cdot \\ \text{nsgt}_4 \end{bmatrix} \quad (5)$$

where nsgt_i is the proportion of total green time (output.1) that will be allocated for the north/south green phase at Intersection i. In this example, $i = 1$ to 4.

Test Domain

To facilitate simulation and understanding, a typical traffic situation was constructed that was both manageable and comprehensive. The street network has four intersections shaped in a square configuration, each intersection being connected to two other intersections by perpendicular roadways. All GA simulations discussed in this paper were performed using this configuration, shown in Figure 1.

IMPLEMENTATION

The first stage of implementation involved developing a simulation program that could accept both traffic conditions (input.1 and input.2) and a proposed signal-timing strategy (output value output.1 and output vectors output.2 and output.3) and produce an evaluation of this signal-timing strategy under the given traffic conditions. This simulation is needed

by the Traffic GA—it is the fitness evaluation black box. This simulation executes cars through the network (see Figure 1).

The simulation is done on a micrograined scale, where all cars are considered separate entities. A car's actions are individually considered at every simulation time step (approximately 3 sec of traffic time), leading to a more accurate real-world representation and increased computational effort. This simulation has limited capabilities and is used only to illustrate the potential that GAs have in locating near-optimal timing strategies. Other simulation models, such as TRAFNETSIM, are much more complex and can handle a much more diverse set of roadway conditions (*9*).

The simulation module of the Traffic GA at this point should not be compared with other simulation models because the purpose of the Traffic GA is to show how this optimization technique is applied to a traffic situation. The simulation model used here is simplistic at this stage and may not provide simulations more realistic than existing traffic simulation models. Those models have been field tested and validated to replicate real-world traffic conditions, but the Traffic GA simulation has not yet been tested. However, the Traffic GA has a different purpose: to show that a GA can be successfully applied to a traffic timing situation, even with a simplistic simulation model.

The simulator has a number of aspects involving the generation of random events. First, the arrival volumes are specified by the probability of receiving input for any single simulation time step and the bounds on the number of cars coming into the network. The simulator chooses to add input based on the probability and then selects an equally distributed random real number between the given bounds. The integer part of this real number is added as input, and one additional car is added with probability equal to the decimal part of the real number. Alternatively, the simulation could be easily modified to accept single arrival volume values, and the simulator could choose to add input based on the probabilities related to these volumes. Second, the destination of cars is decided at random based on a probability distribution of which lane a car will choose: the left lane (to turn left) (0.15 probability), the middle lane (to go straight through) (0.70), or the right lane (to turn right) (0.15).

Optimization Criteria

The simulation output consists of a value of merit describing how well the cars were able to move through the street network using the given signal-timing strategy under the given traffic conditions. Many different values of merit could have been selected (individually or in combination), including total delay, total number of stops, total linear combination of delay and the number of stops, total cost of losses, total fuel consumption, total person delay, and sum of the squares of the queue lengths (*10*). These criteria options are optimal when they are minimized.

To consider multiple values of merit, an expression that arithmetically combines a number of the individual values of merit could be defined. For example, total delay and total number of stops could be used to define the final value of merit through an expression like

$$m = (k1)(td) + (k2)(ts) \tag{6}$$

where

m = final value of merit,
td = total delay,
ts = total number of stops, and
k1, k2 = specified constants.

In the Traffic GA, total delay, or what we called total average wait time of a car in the street network, was chosen as the preferred evaluation criterion because it is relatively easy to calculate in the Traffic GA's simulation module.

In general, computing automobile delay is a complex process. This process is well documented in the *Highway Capacity Manual* (*11*). The process used to compute delay in the simulation discussed here is simple. However, this procedure is sufficient for the purpose of this study—to examine the application of a GA to traffic signal optimization. The GA can function in the same manner with more complex delay equations. For this study, the delay is computed by counting the total number of cars involved in the simulation and summing the number of cars that were not moving for each simulation time step. The expression for "total average wait time per car" is

$$\frac{\sum_{i=1}^{i=\mathrm{TTS}} w_i}{\mathrm{TC}} \tag{7}$$

where

TTS = total number of time steps executed in a complete simulation,
w_i = number of cars waiting at Time Step i, and
TC = total number of cars in the network.

This expression indicates how long, on the average, a car will be delayed between the time it enters the street network and the time it exits the street network.

This evaluation expression needs to be modified slightly so a GA can use it during reproduction. The GA's only requirement from the simulation module is availability to an objective function (that will produce a fitness value). This function needs to be optimal at maximum values. Therefore, since the evaluation criteria we chose above relates to a minimization problem, it needs to be converted to a maximization problem. This was done by using the inverse of the total average wait time per car. Therefore, since we want to minimize the wait time per car, we'll need to maximize the inverse of this wait time.

Decision Variables

Specifically for the current implementation of the Traffic GA there are nine decision variables: one global variable (total green time) and two local variables for each of the four intersections. The two local variables are (*a*) the directions in

which the first green phase will allow traffic to flow (that is the north/south traffic will be allowed to move first = 1 or the east/west traffic will be allowed to move first = 0) and (*b*) the proportion of the total green time allocated to the north/south green phase (a real value between 0.0 and 1.0). This results in a cycle consisting of two phases, a north/south phase and an east/west phase where left-turning cars proceed during traffic gaps (i.e., permitted). The directions of flow for the first phase are determined from the variables above, and the directions of flow for the second phase are assumed to be the directions perpendicular to the first phase's directions (e.g., if the first direction is east/west then the second phase's direction is north/south). Therefore, the GA will not have the opportunity to change the alternating nature of the traffic signals but will be allowed to change which directions get the green phase first.

Note that this choice of decision variables is not fixed. Because of the adaptive nature of GA applications, other decision variables could be easily implemented in the future. For example, if a user wanted to add more than two phases per intersection cycle or wanted to include offsets as decision variables, only the bit string and the simulation would have to be altered. The GA's overall structure would not have to be changed.

Constraints

These decision variables have been established so that almost no external constraints are needed. The only constraints on the variables are the strict limitations on the range of values they may use. First, the direction can only take on binary values because there are only two phases implemented in the current traffic simulation module. Second, the individual green phase times may not be less than 6 sec because times less than this would barely allow any cars to get through an intersection.

Bit String Coding

A direct coding of these nine decision variables was chosen. The global variable, total green time, was coded into a four-bit string mapped between the minimum total green time (24 sec) and the maximum total green time (2 min). The first phase directions are coded directly from a single bit as stated above. The variables that represent the proportion of total green time allocated for north/south green phase are coded into four-bit strings. The four bits are converted to an actual time value by transforming to an integer value between 0 and 15, dividing the number by 15 (to get the number between 0.0 and 1.0), and then mapping it to an integer between minimum green time (6 sec) and total green time (calculated above) − minimum green time (6 sec).

The nine decision variables result in a 4 + (1 + 4) * 4 = 24 bit string. This string is ordered as follows: the total green time and then the two variables for each intersection are grouped together, and then strung together from Intersections 1 to 4, as shown in Figure 2. This ordering was selected so that intersection characteristics would be adequately near one another, so the GA would have a higher probability of developing tight linkage between the relevant bits (5).

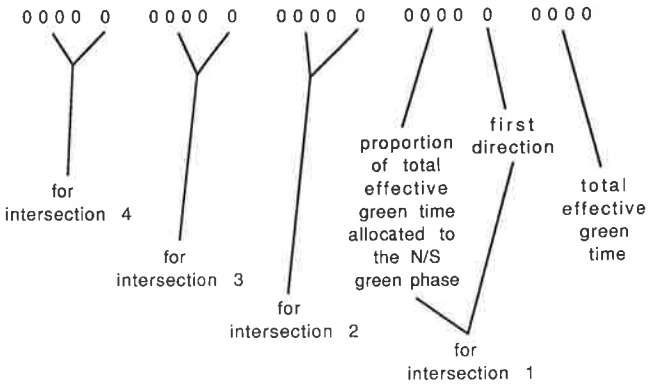


FIGURE 2 Bit string mapping (read right to left).

Traffic GA—Step by Step

As discussed earlier, most simple GAs operate similarly to what is described here, with only the fitness function varying from application to application. Figure 3 is a flow chart of the steps executed by the Traffic GA to find a near-optimal traffic signal-timing configuration for given traffic conditions.

The three main steps involved in the Traffic GA are shown in Figure 3. First, the traffic simulation and the GA are initialized. The initialization of the traffic simulation involves establishing the street configuration and the traffic conditions within the computer program. The simulation need not be reinitialized later in this procedure because all simulations start at these same common conditions. The initialization of the GA involves establishing an initial, completely random population of bit strings. The bit strings symbolize traffic signal timing strategies as described earlier.

The second main step in the Traffic GA is the fitness computation. This involves taking each GA population member and executing a simulation using the timing strategy represented by this member. The fitness evaluation step is executed many times because new population members are continually being generated by the GA. Fitness evaluation is usually continued until the GA has converged; this point is generally defined by the user.

The last main step is the evolution of the GA population. This involves manipulations on the bit strings (i.e., operations on the population members). The three manipulations, or operators, used in the Traffic GA are reproduction, crossover, and mutation, which were described earlier.

The Traffic GA may be run either off-line or on-line. If it is run off-line, the Traffic GA finds a near-optimal signal-timing strategy for any given traffic condition. If it is run on-line, traffic information is continuously gathered from detectors placed on all approaches to all intersections, and the Traffic GA is periodically executed. It is possible to specify very short time intervals between execution, but this would probably not be desirable. To execute, the Traffic GA would be given the most recent traffic data, and then it would be expected to find a near-optimal signal-timing strategy that promoted smooth traffic flow. The new signal-timing strategy would be used in the real traffic signals until the Traffic GA was executed again with new, updated traffic information.

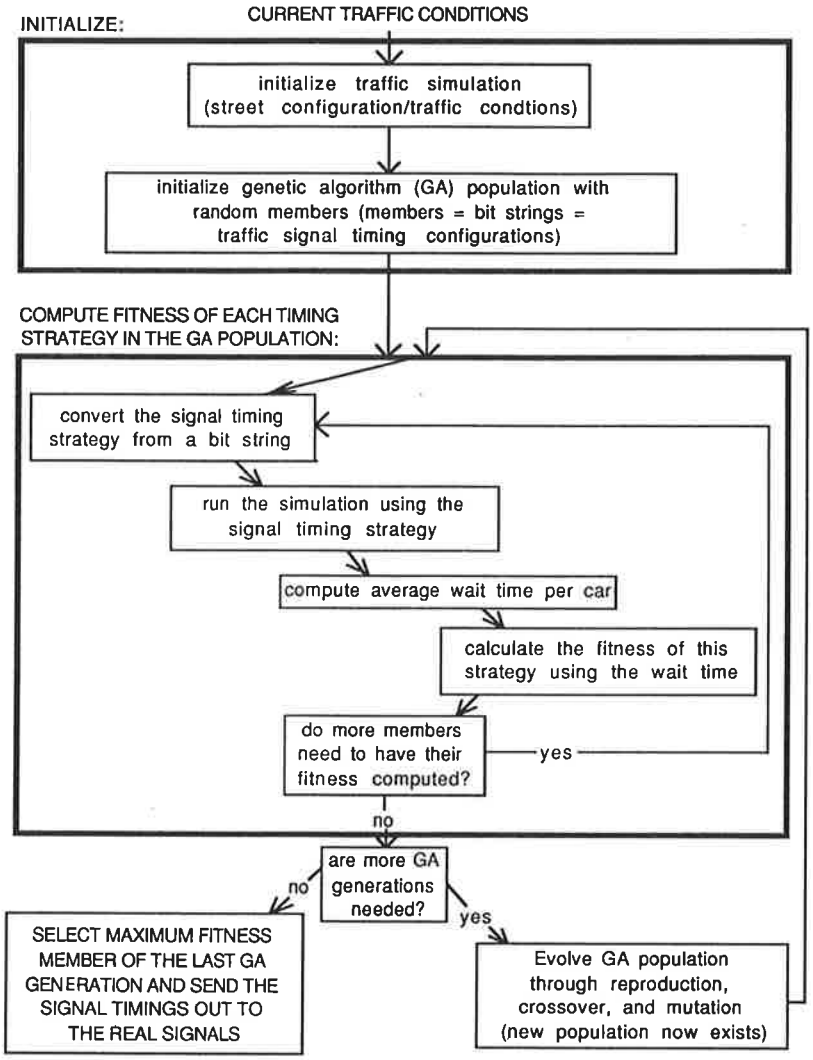


FIGURE 3 Traffic GA procedural flowchart.

COMPUTATIONAL RESULTS

The results of running this GA on typical traffic situations can vary depending on the simulation settings used. All runs performed during the writing of this paper show steady improvement in the average population fitness as the GA population evolves from generation to generation.

Simulator Parameter Settings

In the case examined here, the GA simulations used typical parameter settings: four intersections configured in a square (see Figure 1); yellow time of 3 sec; the probability of a car going straight = 0.70, left = 0.15, and right = 0.15. The average rate at which cars enter an intersection on a green phase was as follows: for cars going straight, one car every 2 sec, right, 1 car every 2 sec, and left, one car every 6 or 12 sec (permitted to enter depending on the arrival volumes of the opposing traffic). The length of time for a car to get from one intersection to another was 24 sec (translates into a distance between all adjacent intersections of 1,000 ft and a constant traveling speed of 28 to 30 mph). The simulation time was 5 min (equal to 100 simulation time steps). The minimum green phase time was 6 sec; the maximum green phase time was 114 sec. The minimum cycle time was 30 sec, and the maximum cycle time was about 126 sec.

Traffic Environment

A common traffic environment was used for all GA runs discussed in this paper. The input.1 vector, the number of cars at all locations, was initialized with typical numbers. This situation was initialized with typical numbers since no particular real-world situation was involved. The task of modifying the code to read in current traffic conditions from detectors, so that real-world problems could be solved, would be very simple. The second input vector, specifying arrival volumes, was set to the values corresponding to the volumes given in Table 1. The north and south approaches were given 5 to 6 times as much traffic as east and west directions. This situation is common in street networks where two opposing directions have considerably higher traffic volumes than the perpendic-

TABLE 1 Arrival Volumes for the Traffic GA Test Run

Approaching	Arrival Volume (cars/hour)
from North to Intersection 1	1560
from West to Intersection 1	288
from North to Intersection 2	1560
from East to Intersection 2	288
from South to Intersection 3	1560
from West to Intersection 3	288
from South to Intersection 4	1560
from East to Intersection 4	288

ular directions because of a busy central business district. For example, if this block of intersections is directly north of a shopping mall, the highest traffic volumes will occur for cars traveling south, into the mall area, and north, out of the mall area.

Traffic GA Results

To obtain a stable, unbiased, average result for this report, five Traffic GA runs with different initial GA populations were executed. Each run was executed for 60 GA generations with a GA population of 50. This means that for each run, the fitness function (the 5-min-of-traffic-time simulation) was executed 3,000 times. At each generation, the average fitness of the generation is calculated and the member in the population with the best fitness value (that is, the shortest wait time) is identified. To produce Figure 4, the average fitnesses of each GA generation (from Generation 0 to 60) of each of the five Traffic GA runs were then averaged together. This produced the "average wait time of population" line. The "minimum wait time of population" was produced in the same manner by averaging the fitnesses of the best-of-generation members for each of the five Traffic GA runs. Figure 4 shows how the GA starts with bad solutions (that is, solutions that produce on-average high wait times) and locates good solutions (that is, solutions that produce on-average short wait times).

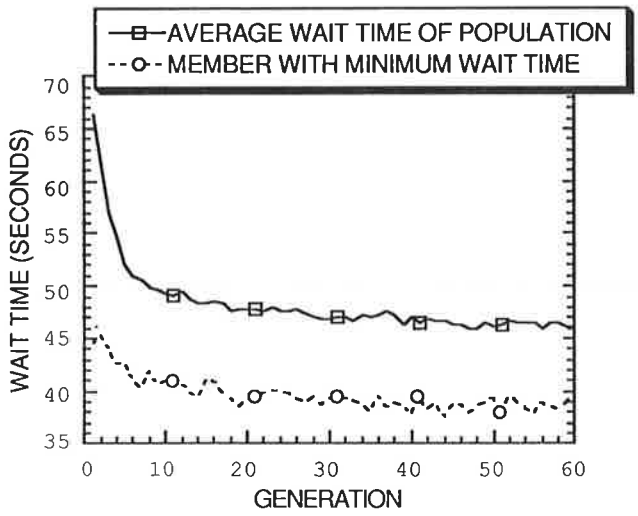


FIGURE 4 Average of five traffic GA runs [best-of-generation (minimum wait time) and generation average (average wait time) results].

The graph shows that the population seems to converge to the optimum or near-optimum member by the 20th or 30th generation. Therefore, it is possible to terminate the GA after 20 generations instead of after 60 generations and still obtain a near-optimal solution. This reduced-generation scenario would reduce the number of simulations from 3,000 to 1,000. The graph also shows that typical minimum wait time values were around 40 sec for these traffic conditions. After the last generation, which in this case was the 60th, the member with the maximum fitness (minimum wait time) can be selected as the best signal-timing configuration and called the solution from the Traffic GA. Then, if this Traffic GA run was performed using real traffic data, the solution could be used to time the real traffic signals to promote smooth traffic flow.

A typical maximal fitness member, actually found by one of the GA runs executed on the traffic environment described in Table 1, is given in Table 2. Note that for all intersections, the green phase time for the north/south (N/S) directions was considerably longer than for the east/west (E/W) directions. Observe that total cycle times are equal because the bit string contains only one field to represent total green time, but again the bit string and simulation could be easily modified to allow different total cycle times. The Traffic GA selected a total cycle time of 60 by itself; this number is not programmed into the GA. The GA found a strategy that used very similar green phase times for the north/south directions and also for the east/west directions. We expect this behavior because similar green phase times often allow the best flow of traffic because cars can move through the network with fewer stops if there is some type of synchronized cycle time (*10*). Finally, the GA could have given green phase times up to 114 sec but only went as high as 45 sec. This is because the GA was searching for a strategy that would minimize wait time, and if it were to allocate more green phase time to north and south directions, the wait times for cars coming from the east and west would increase too dramatically to make this beneficial. This solution provides a 33-sec green band for northbound traffic of Intersections 1 and 3 and another 33-sec green band for southbound traffic of Intersections 2 and 4.

Run Time

One entire GA run, which amounts to a total of 3,000 simulations and 60 generations of a GA, took 2.0358 system CPU sec on a supercomputer (Cray 2 with four processors). This is a reasonably long job time considering that this time would be greater on more readily available processors. On the other

TABLE 2 GA's Maximum Fitness (Minimum Wait Time) Timing Strategy

Intersection Number	First Direction	Green Time (sec)	Second Direction	Green Time (sec)	Total Cycle Time (sec)[a]
1	E/W	12	N/S	42	60
2	N/S	36	E/W	18	60
3	N/S	42	E/W	12	60
4	E/W	9	N/S	45	60

[a] Total Cycle Times have two yellow phases of 3 seconds each added, in addition to the two green times.

hand, if the number of generations were cut by two-thirds to 20 as suggested earlier, the CPU time would be cut by two-thirds because the simulation takes up almost all of the CPU time, whereas the GA operators use very little. Depending on how often a user wants to recalculate a signal-timing strategy and for how many intersections, the required processing time may increase or decrease. It is possible that the required computational effort could be too large, preventing use of the Traffic GA to calculate signal timings in very short time intervals.

Performance

Though this run is only one case, it is expected that the Traffic GA will always converge to a reasonable timing strategy.

Reasonable timing strategies have been found in many different cases not reported in this paper. For example, when arrival volumes were increased to a point of oversaturation, the GA responded by finding signal-timing strategies with longer cycle times, something a traffic engineer also would do if it were possible to have constant human monitoring of traffic signals.

Furthermore, most GA researchers agree that the theory of convergence for simple GAs has become fairly well developed, indicating that the performance reported earlier is typical of GA behavior. The critical components of this theory focus on building blocks (*5,12*), building block growth (*12*), the possibility of being misled by building blocks (*13–15*), and mixing and statistical decision making (*16*).

Therefore, overall Traffic GA results (including the cases not reported here) and the theory of convergence indicate that GAs may be able to solve more difficult problems than traditional control strategies and search methods. GAs seem to be better on both accuracy and convergence time. Finally, the advantages of demand-responsive control over other forms of traffic control include the capacity to constantly examine situations and respond to them with no traffic knowledge and no human attention.

CONCLUSIONS

This paper reported on an application of a genetic algorithm to produce near-optimal traffic signal-timing strategies for a network of intersections. Examples and simulation parameters were included for illustrative purposes and to demonstrate the roll a GA could play in signal-timing determination. The Traffic GA produced reasonable traffic signal-timing plans. The results suggest that this method of searching for an optimal signal-timing strategy has the potential to improve existing traffic control techniques. It is especially encouraging that the GA could find balanced conditions of green phase times and a reasonable cycle length as a function of traffic demand.

The Traffic GA produced logical signal timings using simple GA operators and a simple simulation model. Changing the GA may be warranted if this problem were scaled up to handle many more intersections. Future work on the simulation would be required to make the Traffic GA more realistic and capable of handling more complex intersection flow conditions.

Computer traffic control deserves attention because of the possible benefits from improving traffic flow. An adequate solution to this problem would increase roadway efficiency, reduce travel time, make travel time more predictable, improve safety, cut down on harmful emissions, decrease fuel consumption, and increase driver comfort.

ACKNOWLEDGMENTS

Mark Foy would like to thank Chad Hall for the initial motivation to examine traffic flow and traffic control strategies. Thanks go to the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign for providing the computer time to run the Traffic GA.

David Goldberg acknowledges support by the National Science Foundation.

REFERENCES

1. N. H. Gartner. OPAC: A Demand-Responsive Strategy for Signal Control. In *Transportation Research Record 906*, TRB, National Research Council, Washington, D.C., 1983, pp. 75–81.
2. F.-B. Lin and S. Vijayakumar. Adaptive Signal Control at Isolated Intersections. *Journal of Transportation Engineering*, Vol. 114, No. 5, Sept. 1988, pp. 555–573.
3. J. S. Linkenheld, R. F. Benekohal, and J. H. Garrett, Jr. A Knowledge-Based System for the Design of Signalized Intersections. *ASCE Journal of Transportation Engineering*, Vol. 118, No. 2, March 1992, pp. 241–257.
4. D. P. Mital. An Intelligent Urban Traffic Network Controller and Simulator. *IETE Technical Review*, Vol. 7, No. 1, Jan. 1990, pp. 52–62.
5. D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, Mass., 1989.
6. *Proc., International Conference on Genetic Algorithms and Their Applications* (John J. Grefenstette, ed.). Carnegie-Mellon University, 1985.
7. *Proc., Second International Conference on Genetic Algorithms* (John J. Grefenstette, ed.). Massachusetts Institute of Technology, 1987.
8. *Proc., Third International Conference on Genetic Algorithms* (J. David Schaffer, ed.). George Mason University, 1989.
9. *TRAF-NETSIM User's Manual*. Federal Highway Administration, U.S. Department of Transportation, 1989.
10. S. Reljic. TRAFSIG: A Computer Program for Signal Settings at an Isolated, Under- or Oversaturated, Fixed-Time Controlled Intersection. *Traffic Engineering and Control*, Vol. 29, No. 11, Nov. 1988, pp. 562–566.
11. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
12. J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.
13. D. E. Goldberg. Simple Genetic Algorithms and the Minimal, Deceptive Problem. In *Genetic Algorithms and Simulated Annealing* (L. Davis, ed.), Morgan Kaufmann, Los Altos, Calif., 1987, pp. 74–88.
14. D. E. Goldberg. Genetic Algorithms and Walsh Functions: Part I. A Gentle Introduction. *Complex Systems*, Vol. 3, No. 2, 1989, pp. 129–152.
15. D. E. Goldberg. Genetic Algorithms and Walsh Functions: Part II. Deception and its Analysis. *Complex Systems*, Vol. 3, No. 2, 1989, pp. 153–171.
16. D. E. Goldberg, K. Deb, and B. Korb. Messy Genetic Algorithms Revisited: Studies in Mixed Size and Scale. *Complex Systems*, Vol. 4, No. 4, 1990, pp. 415–444.

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

Investigation of the Impacts of Ramp Metering on Traffic Flow With and Without Diversion

SALAMEH A. NSOUR, S. L. COHEN, J. EDWIN CLARK, AND A. J. SANTIAGO

The effect of various levels of ramp metering on traffic flow in a 7-mi-long urban corridor consisting of a freeway, two parallel surface arterials, and seven perpendicular connecting surface arterials was evaluated. The study was conducted both with and without traffic diversion from on-ramps to surface streets using the INTRAS freeway corridor simulation model. Three levels of ramp metering were analyzed to determine how much each would reduce the effects on traffic of an incident on the freeway. For each level of metering, several traffic diversion schemes were introduced in an incremental manner. The diversion of vehicles was implemented for each level of ramp metering until the remaining number of vehicles behind the meters was the same for each level and was less than the storage capacity of the ramp behind the meter. The main conclusion is that, whereas ramp metering improves the traffic flow on the freeway, it adversely affects the total system because of the overflow queues behind the meters, which spill back onto the surface streets. The only metering level that does not do this (in the absence of diversion) is one in which the metering rates are adjusted so that the overflow queues do not occur (equivalent to a queue detector at the upstream end of a ramp overriding the meter when a queue is detected). This level of metering, however, is rarely sufficient to overcome the capacity reduction resulting from an incident. To minimize the adverse effects of ramp metering, an appropriate traffic diversion plan for the implemented ramp metering strategy is required. The impacts on the total system under the optimum ramp metering and the best diversion plan consist of only a 4.1 percent increase in speed and a 10.5 percent decrease in delay.

Ramp metering controls the number of vehicles entering a freeway from on-ramps by subjecting the entry to a fixed-time or a traffic-responsive control similar to the conventional control provided by traffic signals. The purpose of such a control is to reduce the traffic demand onto the freeway to maintain adequate freeway traffic flow and to prevent the development of congestion on the freeway. The impacts of ramp metering on freeway traffic flow include increasing the traffic speed and the rate of mainline flow in terms of miles per hour and vehicles per hour per lane, respectively. If the demand at the metered ramp is higher than the discharge rate, a queue behind the signal will build up and eventually overflow onto the adjacent surface street. The occurrence, duration, and extent of overflow queues also depend on the available storage length between the meter and the surface street. Depending on the duration and the magnitude of overflow queues, delay and congestion will be introduced onto the surface street system. Another impact of ramp metering is that it induces traffic route diversion. This is because some motorists wishing to use metered ramps will divert either to unmetered on-ramps or to less restrictive metered ramps or will not travel on the freeway and make their trips exclusively on surface streets. Since the delays due to waiting behind a ramp meter are proportionally greater for short trips (i.e., one to three interchanges), these are the trips most likely to be diverted to surface streets.

The diversion of vehicles from on-ramps and freeways to surface streets is a major institutional concern for both city traffic officials and freeway agencies. Whereas the diverted vehicles help to reduce the delay due to queuing behind the meters on ramps and help to reduce overflow queues, they increase the traffic volume on the surface streets. This type of diversion, in the opinion of some transportation agencies, tends to "greatly reduce the delay in the corridor by simply removing some of the vehicles from the freeway" (*1*).

When vehicles divert in this manner, there is no documentation available that quantifies this delay reduction and compares it to the delay in the same case of ramp metering but without traffic diversion taking place. However, several studies and reports have documented the benefits of ramp metering with diversion relative to the "before" condition of no ramp metering. However, previous studies in this area do not treat the parallel surface streets that will handle the diverted traffic with the same level of detail as the freeway. For instance, many of them use a freeway simulation model together with a module that treats the alternate route only in terms of a given average speed with no consideration of the impacts of the diverted traffic. Unlike the previous studies, this study considers the surface street arterials with the same level of detail as the freeway. The purpose of this study is to investigate, through simulation, the impacts of ramp metering with and without diversion on traffic flow in an urban corridor and to examine, quantitatively, the effects of the most likely type of ramp metering–induced diversion. The study considers different levels of diversion induced by three levels of ramp metering operating at six ramps in an urban corridor.

S. A. Nsour, Civil Engineering Department, Santa Clara University, Santa Clara, Calif. 95053. S. L. Cohen and A. J. Santiago, Turner-Fairbank Highway Research Center, HSR-10, Federal Highway Administration, 6300 Georgetown Pike, McLean, Va. 22101-2296. J. E. Clark, Civil Engineering Department, Clemson University, 110 Lowry Hall, Clemson, S.C. 29634-0911.

METHODOLOGY

A review of the literature on ramp metering and traffic diversion was undertaken. The subject of traffic diversion in ramp metering operations was not discussed in every acquired report of ramp metering experiences. However, some of these reports included information relevant to diversion, such as storage behind meters, ramp delays, and queues. Several reports discussed the changes in the traffic flow on surface streets in terms of volumes and speeds. Only a few studies (*2–4*) of ramp metering experiences included a quantitative comparison of measures of effectiveness (MOEs) between "after" conditions with diversions occurring and "before" conditions prior to the implementation of ramp metering.

Two studies (*5,6*) involved the use of simulation as a tool in evaluating the effect of diversion in general terms. One study (*5*) included simulation of a diversion strategy only from NY-495 to another freeway without ramp metering. The other study (*6*) of I-25 in Denver, Colorado, used a simulation technique to select the best alternative from several schemes involving ramp metering on different sections and diversion of different numbers of vehicles. Most of the reports that calculated the impacts of ramp metering considered only the freeway MOEs.

To achieve the purpose and objectives of the study, a real-world corridor was selected for simulating traffic flow under the various conditions. A review of the available simulation models indicated that the INTRAS (*7,8*) simulation model was the most appropriate for this type of study for the following reasons:

1. INTRAS is a microscopic model that has been validated for freeway traffic, including simulation of incidents, and has a microscopic surface street component equivalent to the NETSIM model that has been validated for simulating surface streets (*9*).
2. The INTRAS model simulates ramp metering and the effect of overflow queues at the interface between the links behind a ramp meter and the upstream surface streets.
3. The INTRAS model simulates a surveillance system and has a module for computing detector point processing, which can be used to follow the formation of queues behind the incident on the freeway.
4. Because the INTRAS model is microscopic, quantities such as capacity are an output rather than an input, which is the case with macroscopic models. This is an absolute requirement if we are to use the results of the model to determine the number of vehicles that must be metered to alleviate the effects of an incident lane blockage.
5. INTRAS explicitly models the origin-destination pattern on the freeway by assigning individual vehicles to destinations as they enter the freeway. Thus, the paths and strategies they must take to achieve the assigned destination are explicitly modeled. This is not true of any other model.

No other simulation model currently available satisfies these requirements.

The simulation approach, rather than actual field studies, was chosen for several reasons. Some of the MOEs needed in this study cannot be measured in field studies precisely or even adequately within reasonable time and cost constraints. These include the total travel time in vehicle hours and queuing delays. However, since INTRAS is a microscopic simulation model, it is possible to obtain precisely those MOEs that are difficult to collect in the field.

Another reason for using simulation is that, to achieve the objectives of this study, different schemes of traffic diversions involving variable numbers of diverted vehicles and different origin-destination patterns are required. Experimentation with various combinations of numbers of diverted vehicles, ramp metering rates, and origins and destinations of trips are impractical in field studies, especially during incident-caused congestion like the one considered in this study. Further, in a ramp metering system, diversion might take place immediately after implementation. Therefore, no time period involving metering without diversion of the same system will exist, and accordingly, no MOEs for this condition can be observed and measured for comparison.

The sequence of steps in this study's approach was as follows:

1. Select a real-world network (made up of a freeway with parallel and intersecting surface arterial streets) that provides closely spaced interchanges with both metered and unmetered on-ramps.
2. Code the selected network according to the INTRAS model coding procedures.
3. Revise the signal phase timings for all signalized intersections using the Webster method.
4. Perform the basic simulation, which represents the "before" condition to which all other cases will be referenced and compared.
5. Develop three levels of ramp metering at six on-ramps to reduce traffic congestion on the freeway.
6. For each level of ramp metering, develop a diversion scheme or schemes to help reduce any overflow queues resulting from ramp metering.
7. Investigate the impacts of each metering level diversion scheme.

THE NETWORK

To satisfy the method and purpose of this study, a real-world network was chosen. A 7-mi stretch of Route 22, the Garden Grove freeway, in Orange County, California, was selected for the study with a boundary encompassing two parallel and seven intersecting arterial surface streets with 28 signalized intersections. The two-way freeway section includes seven interchanges with 14 on-ramps and 14 off-ramps and has generally three through lanes in each direction. The nonincident level of service for the freeway was C and for the surface streets was in the range C to D, depending on the particular intersection. Figure 1 represents the total areawide network. This network was chosen for the following reasons: data were available (*10*), the network included both a freeway and surface streets, and the surface streets provided good alternate routes for diverting traffic.

No claim is made that this network is representative of freeway corridors around the country. On the contrary, the

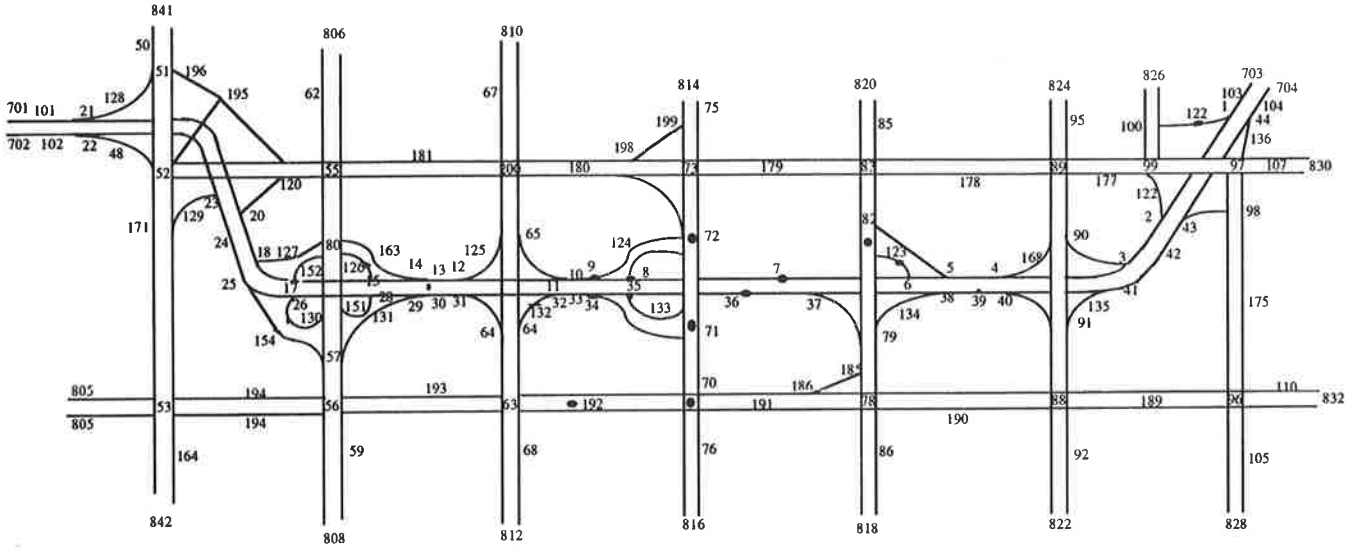


FIGURE 1 Garden Grove freeway network, Orange County, California.

situation relative to alternative routes is substantially better than most freeway corridors in the country. Therefore, it might be said that this network provides a "best case" opportunity for ramp metering with diversion. If no benefits can be obtained here, it is unlikely that they can be obtained anywhere else.

DESIGN OF SIMULATIONS

Basic traffic parameters had to be established and input into all simulations. These include the desired free-flow speed in each of the subsystems of the network and the mean queue discharge headway from signalized intersection approaches. The desired speed or the free-flow speed of the freeway was set at 65 mph, which represents the observed operating speed on the Garden Grove freeway, and was set at a speed of 35 mph for all entrance and exit ramps, which represents the observed operating speed on those facilities. The desired speed for all surface streets was set at 40 mph, representing the observed operating speed on the surface streets. For traffic discharging from a queue on off-ramps and surface streets, at signalized intersections, the mean queue discharge headway was set to 1.8 sec, on the basis of physical observations made when the data were originally collected (*10*). Finally, the duration of simulation was set at 90 min to allow the introduction of the incident, analysis of perturbations, system recovery, and other factors as discussed in a later section.

A no-incident simulation was established to represent the traffic flow in the network with the timings for the signal phases of intersections adjusted to produce the minimum possible delay. This case represents the "before" condition.

The various after conditions included an incident in the westbound direction. The location of the incident was determined so that the number of vehicles entering from the on-ramps and passing through the link where the incident occurred would be maximized. In reference to Figure 1, the incident, in which the right lane was blocked, was placed on Link 41-42. The trip origin-destination distribution from the no-incident simulation was used to calculate the number of vehicles destined downstream of the incident.

The characteristics of the incident, such as length of time, extent, type, and time of onset, were chosen to produce the congestion on the freeway that required restrictive ramp metering at several on-ramps. The recovery of the freeway traffic flow conditions to the preincident state was then evaluated, and on that basis the final basic simulation was established. Thus, a one-lane blockage lasting for 45 min of simulation and starting 15 min into simulation was selected.

Three levels of ramp metering were then designed for simulation. For each one, an appropriate number of diversion schemes were established.

Metering Level I: Restrictive Metering

The restrictive ramp metering plan was designed to reduce the demand at the incident site to the observed capacity (as measured by the no-metering INTRAS simulation) at the incident site. The metering plan was designed as follows:

1. From the origin-destination table in the output of the basic simulation, the demand on the incident link from each on-ramp was calculated.
2. The total number of vehicles per hour from each ramp to be metered passing through the incident site was obtained. (Some vehicles from upstream on-ramps will exit the freeway before the incident site. These vehicles must be accounted for in developing the metering plan.)
3. Each metered on-ramp contributes to reducing the demand as follows:

$$\text{Reduction in demand at each ramp} = (V_1/V_2)(V_3)$$

where

V_1 = total number of vehicles per hour to be reduced,
V_2 = vehicles per hour passing through incident link, and

V_3 = volume passing through incident link and originating from the on-ramp.

4. Subtracting the reduction in demand as calculated above at each metered on-ramp from its demand yields the volume that should be metered (discharged) from that ramp. Accordingly, the metering rate at each ramp was obtained in terms of release headways between vehicles (e.g., a metering rate of 600 vph is input as a 6-sec release headway in the model).

5. The constraints that affected the final determination of the ramp metering rate were (*a*) the INTRAS requirement that metering headways should be expressed in integers; (*b*) the maximum acceptable release headway, which was fixed as 18 sec for each metered vehicle, and thus the minimum metering rate was 200 vehicles per hour per lane; and (*c*) the minimum acceptable release headway, which was fixed as 5 sec (*11*) for each metered vehicle, so that the maximum metering rate was 720 vehicles per hour per lane.

First Diversion Scheme in Metering Level I

The first diversion scheme in Metering Level I was based on the amount of overflow queue observed under the no-diversion Metering Level I simulation. The first scheme diverts either all the short trips or sufficient short trips to eliminate the overflow queue at each on-ramp, whichever is smaller. Here a short trip is defined as a trip that starts at a given on-ramp, enters the freeway, and exits at the first downstream off-ramp. For example, in Figure 1, a vehicle on Link 63-64 that is destined to traverse a path 63-64-132-32-33-34-71-72-73-75 would divert to the path 63-64-65-209-180-73-75.

To reflect the diversion on the traffic flow in the network, the following steps were taken:

1. At each on-ramp, determine the original routes of those trips that are destined for the next interchange (referred to as one-interchange trips) and determine the most probable alternative arterial routes.
2. Determine the original routes and alternate routes link by link for each on-ramp.
3. On the basis of the amount of diverted traffic and the identification of the alternate routes, revised turning percentages and O-D assignments were calculated for each affected link.

Second Diversion Scheme in Metering Level I

In order to eliminate overflow queue conditions that remained after the initial diversion, a certain number of trips destined for the second interchange from each on-ramp (two-interchange trips) were diverted to surface streets through an alternate route. The exact number of vehicles to be diverted was based on the overflow queue computation.

Third Diversion Scheme in Metering Level I

A third diversion scheme was devised for Metering Level I in which more two-interchange trips from ramps that still have these types of trips were diverted to reduce the number of stored vehicles at each ramp to short queues relative to the available storage capacity.

The speeds on the freeway tend to fall as more vehicles making short trips are diverted from the metered ramps. This is because the metering rates are held fixed but the short trips (one and two interchanges) are diverted. Thus, a greater percentage of the vehicles that are allowed through the meters are long trips destined downstream of the incident site. Another source for variations in speed is the stochastic variation in the assignment of individual vehicles by the O-D trip distribution. To offset such variations, the demand on the freeway was reduced by reducing the metering rate at one ramp.

Fourth Diversion Scheme in Metering Level I

The results of the previous simulation with a reduced metering rate at one ramp caused only a very minor increase in the average speed on the freeway at the completion of Subinterval 2. This situation occurred in spite of the reduction of 154 vehicles in the 45-min incident period while keeping the diverted vehicles as in the previous simulation. To eliminate the resulting overflow queue on this ramp, 40 two-interchange trips in 45 min were diverted.

Design of Metering Level II: More Restrictive Metering

In this strategy, more restrictive ramp metering was used so that the total hourly demand on the incident link would be further reduced by 400 vehicles compared with the demand in the Level I restrictive metering. This reduction in demand was made to offset the effect of the stochastic variations in the on-ramp origin-destination assignment. The new reduction in demand was made higher than any likely increase in demand on the freeway link with the incident due to these stochastic variations. This is, in fact, what is done in actual systems that use time of day metering plans. A real-time traffic-responsive metering strategy would be able to respond to these stochastic fluctuations and thus, on the average, meter less heavily.

First Diversion Scheme in Metering Level II

All the single-interchange trips originating from the six metered on-ramps were diverted as in the first diversion scheme of Metering Level I.

Second Diversion Scheme in Metering Level II

A sufficient number of two-interchange trips were diverted to eliminate the overflow queues at each metered ramp. However, at one on-ramp, diverting all two-interchange trips was not sufficient to eliminate the overflow queues.

TABLE 1 Summary of Delay and Number of Diverted Vehicles in All Cases

Case	Number of Diverted Trips				Adjusted Delay (Vehicle-Minutes)			Unadjusted Delay
	1-Interch	2-interch	3-Interch	Total	Total System without No Incident Freeway Direction	Surf. Streets Without Ramps & Preceding Links	Freeway Incident Direction	For Ramps and Preceding Links (Veh-Min)
Basic Simulation	0	0	0	0	89837	65992	23395	452
Metering Level I	0	0	0	0	103236	81860	11182.00	10429
1st Diversion	288	0	0	288	90457	68029	14500.00	7824
2nd Diversion	288	106	0	394	85784	65664	13779.00	6038
3rd Diversion	288	174	0	462	83253	64394	14550.00	3918
3rd Diversion *	288	174	0	462	85075	64636	15596.00	4618
4th Diversion	288	214	0	502	80451	63589	11717.00	4663
Metering Level II	0	0	0	0	118376	97622	10137.00	10788
1st Diversion	306	0	0	306	95073	76913	9756.00	8183
2nd Diversion	306	282	0	588	80936	63661	10731.00	5830
3rd Diversion	306	282	87	675	84210	64612	14650.00	4381
Metering Level III	0	0	0	0	88651	63404	19506.00	5841
1st Diversion	82	22	0	104	86579	64339	18918.00	3385

* This third diversion is the same as the preceding one except that less metering was introduced at one ramp

Third Diversion Scheme in Metering Level II

At this ramp, then, a number of three-interchange trips had to be diverted.

Design of Metering Level III: Less Restrictive Metering

This strategy was based on metering rates that did not result in any overflow queues from the ramp meters. It is equivalent to a scenario in which metering rates are increased when the overflow queue is detected by a presence detector at the upstream end of a ramp. There was only one stage of diversion needed in Strategy III. The number of vehicles diverted at one ramp was based on the previously stated requirement that there should be the same number of queued vehicles behind a meter in the final diversion scheme in each level of metering. Table 1 includes a summary of the number of diverted vehicles in the period from 4:15 to 5:00 p.m. by type of trip (i.e., one-interchange, two-interchange, or three-interchange).

ANALYSIS OF RESULTS

The total corridor was divided into the following subsystems:

1. Portion of the freeway in the direction with the incident,
2. Metered on-ramps and their preceding links,
3. Surface street subsystem, and
4. Total system without no-incident direction of the freeway.

The INTRAS MOEs output includes results on a link-by-link basis, which permits separating the MOEs in the subsystems mentioned. All the MOEs are output cumulative at the end of every 15-min period in the 90 min of simulated traffic flow. Delay, as defined in this study, is the difference between the actual travel time that vehicles take to traverse the links and the time vehicles would take to travel the links at the free-flow speed. Thus, the comparison of delay in two separate cases in the same system or subsystem provides a comparison of each metering level/diversion scheme combination.

Analysis of Delay

Total System

The changes in the delay (given as vehicle minutes) in the total system (without the no-incident freeway direction) were evaluated at the end of the 90-min simulation period. The values for total delay were adjusted to reflect the fact that the total number of vehicle miles was not constant between simulation runs. The adjustment was done by factoring all values to reflect the no-metering simulation (this is equivalent to using delay per vehicle mile as the MOE). Table 2 presents the MOEs for the total system including adjusted delay and the changes in the adjusted delay in all cases with and without diversion. Hence, on the basis of delay reduction in the total system, Metering Level III is the best strategy considering metering only without diversion. It is, however, the least effective for alleviating congestion on the freeway.

Freeway Subsystem

Table 3 presents the MOEs of the freeway subsystem. It indicates that the most restrictive Metering Level II is the best strategy for alleviating the effects of the incident. However, Table 3 indicates that delays on the incident direction for various cases of diversions vary within the same ramp metering level because of short trips being diverted and stochastic variability of OD assignments, as discussed previously.

TABLE 2 Summary of MOEs in Total System Without No-Incident Direction on the Freeway

Case	Delay (Veh-Min)	Adjusted Delay (Veh-Min)	Change in Adj. Delay (%)	Travel (Veh-Mile)	Change in Travel (%)	Travel Time (TT) (Veh-Min)	Adjusted TT (Veh-Min)	Change in Adj. TT (%)	Average Speed (mph)	Change in Av.Sp (%)
Basic Simulation	89837	89837	0	110000	0	232031	232031	0	28.44	0
Metering Level I	102771	103236	14.91	109505	-0.45	244227	245331	5.73	26.90	-5.42
1st Diversion	90367	90457	0.69	109890	-0.10	232432	232665	0.27	28.37	-0.27
2nd Diversion	85733	85784	-4.51	109934	-0.06	227619	227756	-1.84	28.98	1.88
3rd Diversion	83026	83253	-7.33	109700	-0.27	224983	225598	-2.77	29.26	2.85
3rd Diversion*	85158	85075	-5.30	110107	0.10	227338	227117	-2.12	29.06	2.17
4th Diversion	80592	80451	-10.45	110193	0.18	223289	222898	-3.94	29.61	4.10
Metering Level II	116902	118376	31.77	108630	-1.25	257404	260650	12.33	25.32	-10.98
1st Diversion	94866	95073	5.83	109760	-0.22	236802	237320	2.28	27.81	-2.23
2nd Diversion	80599	80936	-9.91	109542	-0.42	222429	223359	-3.74	29.55	3.88
3rd Diversion	84447	84210	-6.26	110310	0.28	227344	226705	-2.30	29.11	2.35
Metering Level III	88659	88651	-1.32	110010	0.01	231034	231013	-0.44	28.57	0.44
1st Diversion	86664	86579	-3.63	110108	0.10	228988	228763	-1.41	28.85	1.43

* This third diversion is the same as the preceding one except that less metering was introduced at one ramp

TABLE 3 Summary of MOEs in Freeway Subsystem Without No-Incident Direction on the Freeway

Case	Delay (Veh-Min)	Adjusted Delay (Veh-Min)	Change In Adjusted Delay (%)	Travel (Veh-Miles)	Change in Travel (%)	Travel Time (TT) (Veh-Min)	Adjusted TT (Veh-Min)	Change in Adj. TT (%)	Aver.Speed (mph)	Change in Av.Sp (%)
Basic Simulation	23395	23395	0	46719	0	67502	67502	0	41.53	0
Metering Level I	11198	11182	-52.20	46784	0.14	55146	55069	-18.42	50.9	22.56
1st Diversion	14473	14500	-38.02	46632	-0.19	58137	58245	-13.71	48.13	15.89
2nd Diversion	13690	13779	-41.10	46417	-0.65	57512	57886	-14.25	48.43	16.61
3rd Diversion	14397	14550	-37.81	46227	-1.05	57913	58529	-13.29	47.89	15.31
3rd Diversion *	15559	15596	-33.33	46607	-0.24	59262	59404	-12.00	47.19	13.63
4th Diversion	11616	11717	-49.92	46317	-0.86	54992	55469	-17.83	50.53	21.67
Metering Level II	10063	10137	-56.67	46377	-0.73	53481	53875	-20.19	52.03	25.28
1st Diversion	9712	9756	-58.30	46507	-0.45	53624	53868	-20.20	52.04	25.31
2nd Diversion	10527	10731	-54.13	45830	-1.90	53541	54580	-19.14	51.36	23.67
3rd Diversion	14509	14650	-37.38	46269	-0.96	57795	58357	-13.55	48.03	15.65
Metering Level III	19582	19506	-16.62	46902	0.39	63800	63551	-5.85	44.11	6.21
1st Diversion	19002	18918	-19.14	46927	0.45	63244	62964	-6.72	44.52	7.20

* This third diversion is the same as the preceding one except that less metering was introduced at one ramp

Thus, as far as the reduction in the delay on the freeway (incident direction) is concerned, simulation of the final diversion scheme in Metering Level I with 502 diverted trips and simulation of the second diversion scheme in Metering Level II with 588 diverted trips represent the best cases having almost equal reductions in delay (49.9 percent and 54.1 percent, respectively).

Ramp and Preceding Link Delay

On the basis of the delays on ramps and preceding links (e.g., referring to Figure 1, Link 134-38 is a ramp link and Link 79-134 is a preceding link) presented in Table 1, the increases in delay for the metering cases (compared with the no-metering basic case) were very drastic as expected. On the other hand, when the diversion schemes were invoked, these delays were reduced proportionally to the number of diverted vehicles. In the final cases of diversions in Metering Levels I, II, and III, the reduction in delay was 55.3, 59.4, and 72 percent, respectively, relative to the delay in the no-diversion metering case for each metering level. The maximum encountered average delay per vehicle of 737 sec (approximately 12 min) falls within the acceptable limits of on-ramp delays reported in many experiences.

Surface Street Delay

Table 4 presents the MOEs for the surface street subsystem (not including the links immediately upstream of the meters) for cases with and without diversion. The changes in the cases without diversion for Metering Levels I, II, and III were calculated relative to the delay in the no-metering basic simulation case and were 24.1, 47.9, and 3.9 percent, respectively. The large increases in delay for the two most restrictive levels

TABLE 4 Summary of MOEs on the Surface Street Subsystem Without Ramps and Preceding Links

Case	Delay (Veh-Min)	Adjusted Delay (Veh-Min)	Change in Adjusted Delay (%)	Travel (Veh-Miles)	Change in Travel (%)	Travel Time (TT) (Veh-Min)	Adj. TT (Veh-Min)	Change in Adj. TT (%)	Aver. Speed (mph)	Change in Av.Sp (%)
Basic Simulation	65992	65992	0	62311	0	162363	162363	0	23.03	0
Metering Level I	81144	81860	24.05	61766	-0.87	176981	178543	9.97	20.94	-9.08
1st Diversion	68070	68029	3.09	62349	0.06	164878	164778	1.49	22.69	-1.48
2nd Diversion	66004	65664	-0.50	62634	0.52	162521	161683	-0.42	23.12	0.39
3rd Diversion	64711	64394	-2.42	62618	0.49	161668	160875	-0.92	23.24	0.91
3rd Diversion *	64981	64636	-2.06	62644	0.53	161961	161100	-0.78	23.21	0.78
4th Diversion	64315	63589	-3.64	63022	1.14	162144	160315	-1.26	23.32	1.26
Metering Level II	96051	97622	47.93	61308	-1.61	191503	194636	19.88	19.21	-16.59
1st Diversion	76971	76913	16.55	62358	0.08	173425	173294	6.73	21.57	-6.34
2nd Diversion	64242	63661	-3.53	62880	0.91	161604	160142	-1.37	23.35	1.39
3rd Diversion	65557	64612	-2.09	63222	1.46	163732	161373	-0.61	23.17	0.61
Metering Level III	63236	63404	-3.92	62146	-0.26	159699	160123	-1.38	23.35	1.39
1st Diversion	64277	64339	-2.50	62251	-0.10	160721	160876	-0.92	23.24	0.91

* This third diversion is the same as the preceding one except that less metering was introduced at one ramp

of metering were due to the overflow queues blocking the links feeding the links upstream of the meters.

Table 4 indicates that delay decreased when the various traffic diversion schemes were invoked for the two more restrictive levels of metering. This is because the added delay due to the increase in demand on the surface street network was more than compensated for by the decrease in delay due to the overflow queues.

The MOE space mean speed reflects both the travel miles and travel time. Tables 1 through 4 also present the speed changes for all simulations. Figures 2 through 4 show the results graphically.

Analysis of Rerouted Trips

To complete the evaluation, we must consider the changes in vehicle miles traveled and vehicle minutes of travel for diverted vehicles as compared with the original freeway routes. To accomplish this, the vehicle-minutes and vehicle-miles for each link of each route during Subinterval 2 were extracted from the cumulative values for all diversion schemes. From those data, an average travel time per mile on each such link was calculated for that period and then the actual travel times of trips on the diversion routes were obtained.

The incremental schemes for diverting vehicles permitted the comparison of the alternate routes of the last diverted trips with the original routes in the previous simulation. The total length of all original routes for all diversion schemes is not very different from the total length of all alternate routes (3,071 vehicle-mi versus 3,024 vehicle-mi). However, the total travel time involved in using the original routes, for all diversion schemes, is about twice that of the alternate routes (262 vehicle-hr versus 137 vehicle-hr). This indicates that the diversion schemes are valid according to Wardrop's principal,

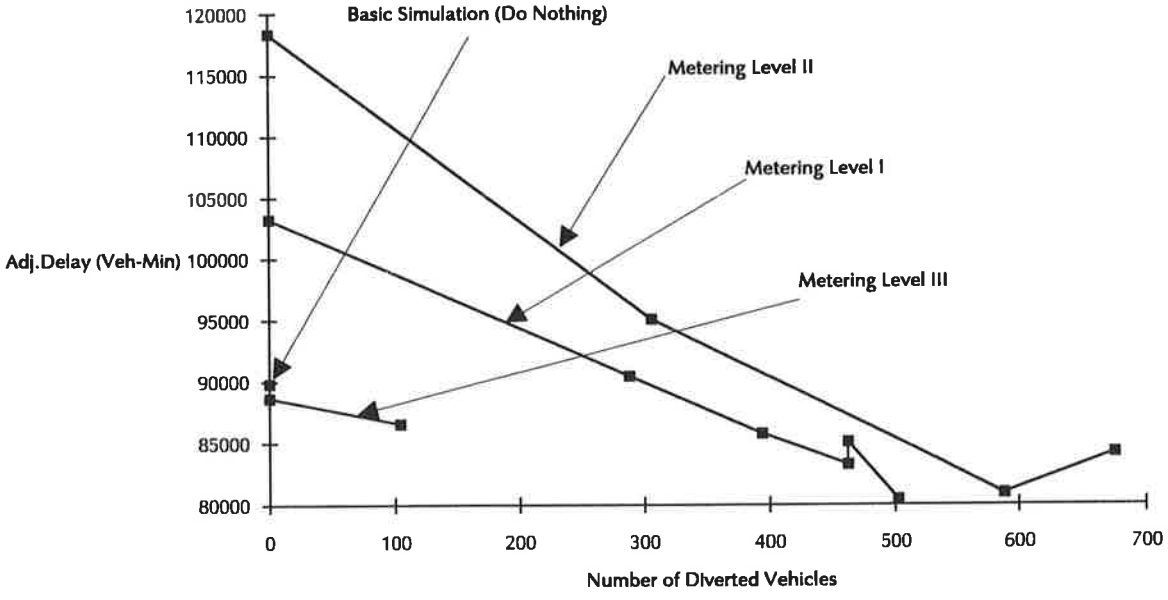


FIGURE 2 Adjusted delay in the total system without no-incident direction on the freeway.

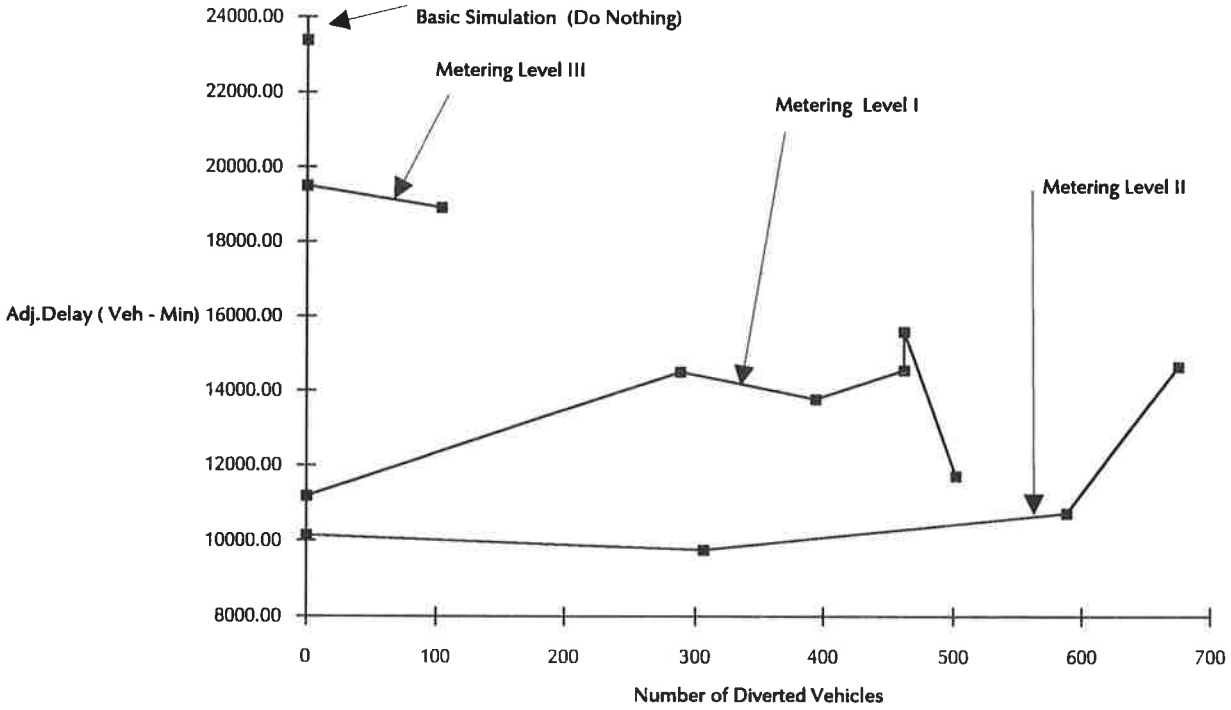


FIGURE 3 Adjusted delay in freeway in the direction of incident only.

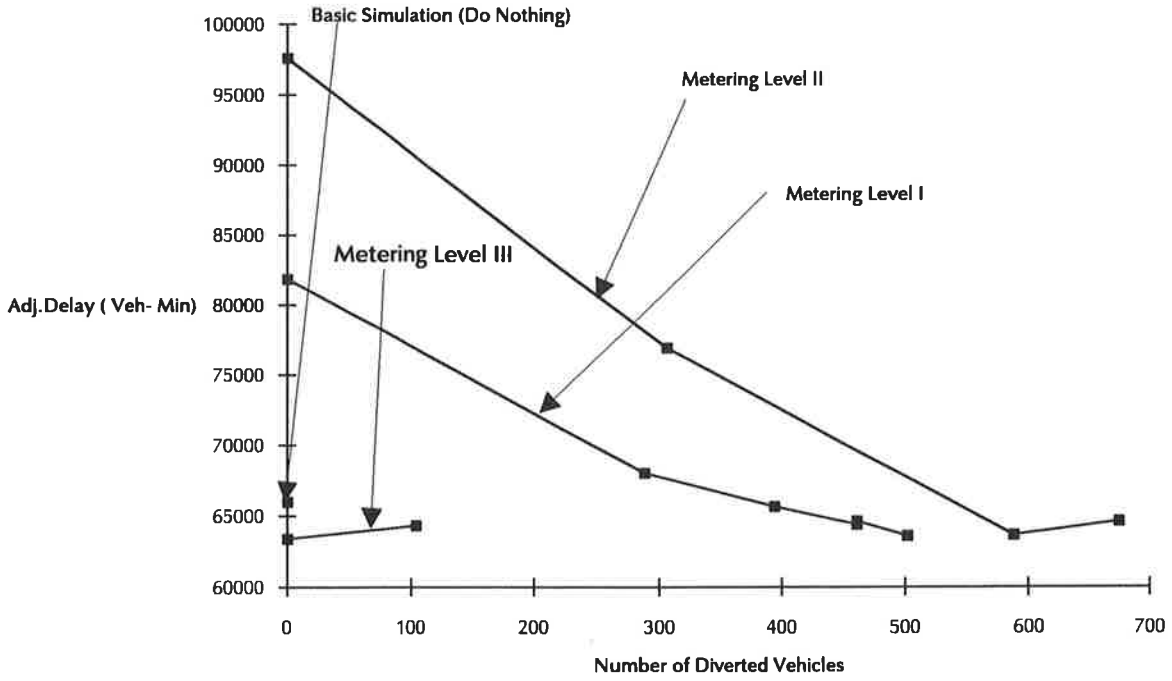


FIGURE 4 Adjusted delay in surface streets without preceding links to metered ramps.

which states that drivers will attempt to use their perceived shortest travel time route.

CONCLUSIONS

The most important conclusions of this study are summarized in this section. The conclusions show that the purpose of the study and its objectives have been achieved.

1. This study demonstrates that, to improve overall network performance by ramp metering, significant diversion from metered ramps is required. This in turn requires that good alternative routes exist. The best results were achieved when all overflow queues behind the meters were alleviated by diverting short trips to alternate routes. However, the improvements were relatively modest compared with some previous results that predict 40 to 50 percent improvements. The previous results were obtained by ignoring the details of the alternate routes, unlike this work, in which these details were included. It is unlikely that even the improvements shown in this study will be obtained in networks with poorer alternate routes (which probably occur in the majority of freeway corridors).
2. Metering Level II with the maximum diversion scheme yielded improvements comparable with those obtained in Metering Level I with the maximum diversion scheme, but it required more vehicles to be diverted to reach the same level of performance. On the other hand, the performance of the freeway was substantially better during Level II than during Level I. In Metering Level III, the improvement in freeway congestion is substantially less than in the other two levels, although no vehicles need be diverted to avoid overflow queues behind the meters. Therefore, this strategy is less effective in reducing freeway congestion.
3. The best case achieved in this study increased the average speed in the total system by 4.1 percent and reduced the total delay by 10.5 percent. This is equivalent to a reduction of 154 vehicle-hr of delay during a 1.5-hr period in which an incident took place for 45 min and closed one lane. Moreover, the significance of this case is that it achieved these total system benefits while addressing the basic direct need of reducing the adverse effects of an incident on the freeway traffic flow in a manner allowing the restrictive Metering Level II. Also, the resulting benefits were achieved by diverting 502 vehicles to alternate routes that took less travel time than they would have for their original routes had they traveled on the freeway after waiting in queues behind the meters. Finally, the users of the surface street subsystem did not incur any significant increase in travel delay because of the diverted vehicles (probably because the diverted flow was very small compared with the existing demand on the surface streets).
4. The INTRAS model was shown to be a powerful tool for analyzing ramp metering and associated traffic diversion. No previous study was able to analyze the alternate surface streets routes to the same level of detail as is available in the INTRAS model.

RECOMMENDATIONS

Following are the major recommendations based on the conclusions and procedures of this study:

1. A capability of assigning alternate routes for diverting vehicles away from metered ramps and automatically adjusting turning fractions and freeway OD matrices would be helpful in performing analyses like those done in this study. This had to be done by hand in this study. Modifying the input headways for metering to tenths of seconds would also help in adjusting metering plans with more precision.
2. Further studies involving ramp metering with diversion are needed to examine the effect of extending the metering period beyond the end of the incident. This would allow faster clearance of the incident on the freeway and would allow one to study the effect of gradually increasing metering rates as the incident clears. Currently, the INTRAS model does not have the capability of changing metering rates within a simulation run, although work is under way to add such a capability.
3. A more refined method is needed to address the fact that vehicles start to divert after a queue has formed behind a meter. In this study, it was assumed that diversion begins immediately after metering begins.
4. A further refinement that should be introduced in similar studies is to consider diverted trips that may reenter the freeway downstream of the incident. This potential for diverting trips was ignored in this study, which only diverted trips that do not return to the freeway.

REFERENCES

1. *Traffic Management System: District 11 San Diego, California.* California Department of Transportation.
2. E. D. Arnold. *Changes in Travel in the Shirley Highway Corridor 1983–1986.* Virginia Transportation Research Council, 1987.
3. *An Evaluation of Fixed Time Ramp Control on the San Diego Freeway in the South Bay Area.* Freeways Operations Branch, California Department of Transportation, 1973.
4. D. Owens and M. J. Shofield. Access Control on the M6 Motorway: Britain's First Ramp Metering Scheme. *Traffic Engineering and Control*, 1988, pp. 616–623.
5. J. F. Torres et al. *Freeway Control and Management for Energy Conservation.* Washington Department of Transportation, 1982.
6. *I-25 TSM Study: Ramp Metering Feasibility Analysis.* Colorado Department of Highways, 1979.
7. B. J. Andrews and D. A. Wicks. *Development and Testing of INTRAS, A Microscopic Freeway Simulation Model—Volume 2. User's Manual.* U.S. Department of Transportation, 1980.
8. D. A. Wicks and E. Lieberman. *Development and Testing of INTRAS, A Microscopic Freeway Simulation Model—Volume 1. Program Design, Parameter Calibration and Freeway Dynamics of Component Development.* U.S. Department of Transportation, 1980.
9. R. B. Goldblatt. *Development and Testing of INTRAS, A Microscopic Freeway Simulation Model—Volume 3. Validation and Application.* U.S. Department of Transportation, 1980.
10. W. W. Recker et al. *Engineering Strategies for Major Reconstruction of Urban Highways.* California Department of Transportation, Sacramento, Calif.
11. *Traffic Control Systems Handbook.* Federal Highway Administration, U.S. Department of Transportation.

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

Development of an Improved High-Order Continuum Traffic Flow Model

PANOS G. MICHALOPOULOS, PING YI, AND ANASTASIOS S. LYRINTZIS

Widespread use of continuum traffic models in practical applications has not been realized since their introduction in the early 1960s and 1970s. This is because some improvements are necessary for wide acceptance of these models, especially in congested and interrupted flow situations. A new high-order formulation capable of describing traffic dynamics under these conditions is introduced and implemented. The formulation does not contain an equilibrium speed-density relationship and therefore requires less calibration effort in field applications. Satisfactory results are obtained when the model is tested using field data representing flows at pipeline freeway and freeway junctions with entrance and exit ramps.

Advanced traffic management and control schemes as well as simulation require reasonably accurate description of flow dynamics, especially in congested situations. Conventional input/output models in general are adequate for simulating traffic in a coarse sense and can be used for planning purposes. However, they are not sufficiently sophisticated for use in the development of high-performance freeway surveillance and control systems or real-time applications. Continuum models are more suitable for representing the short-term traffic behavior because they include both time and space in the state equations and take compressibility into account. The most widely known continuum formulations can be characterized as being either simple order (*1*) or high-order (*2–4*).

The simple continuum formulation is based only on the mass conservation equation, which is supplemented by an equilibrium equation of state. In the high-order continuum formulation, a momentum equation is added to the mass conservation to achieve conservation of momentum as well. In spite of the conceptual appeal of continuum models, they have not been widely used, partly because of our inexperience in implementing them in practical situations and partly because of some needed improvements in their formulation. For instance, the simple continuum model is known for its simplicity, but in principle it is not suitable for describing nonequilibrium traffic dynamics because it does not take into account acceleration and inertia effects. However, it is not clear how important these effects really are, especially in congested conditions. On the other hand, although the existing high-order continuum models consider acceleration and inertia, they appear problematic at congested and interrupted flows (*5–7*). Motivated by these considerations, in this paper we introduce a new formulation based on the existing high-order continuum models. This formulation does not contain an equilibrium speed-density relationship as in most existing continuum models and thus is more attractive in field applications. In addition, it includes a friction term to address the effect of ramp flows especially at congested flows. The proposed model is implemented numerically through finite difference methods. A number of such methods have been tried, and in this paper only the most successful one is presented. A stability analysis has also been performed to determine appropriate mesh sizes in space and time and to bound the parameter values of the model. Subsequently, both qualitative and quantitative tests have been performed to test and validate the model. Whereas the qualitative tests are focused on its physical behavior at congested flows, the quantitative tests of the model evaluate its ability to estimate real-world traffic at congested and interrupted flows. To understand the effect of different discretizations, a number of mesh sizes have been applied (within stability limit), and the findings are discussed. Test results from a simple continuum model are also presented and compared with those from the proposed model.

P. G. Michalopoulos and P. Yi, Department of Civil and Mineral Engineering, and A. S. Lyrintzis, Department of Aerospace Engineering and Mechanics, University of Minnesota, Minneapolis, Minn. 55455.

SIMPLE AND HIGH-ORDER CONTINUUM FORMULATIONS

Continuum models are needed not only for better understanding the collective behavior of traffic, but also for analyzing flow conditions in a dynamic fashion in devising efficient control strategies, simulation, and assessing the effects of geometric improvements. According to the simple continuum model, flow can be described by the conservation equation, which has the following general form:

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = g(x, t) \tag{1}$$

where

$q = q(k) = uk$ is the flow rate of the traffic stream (veh/hr);
k = density (veh/mile);
u = speed (mi/hr);
t = time;
x = space; and
$g(x, t)$ = generation rate, which is equal to zero in freeway sections without entrances or exits.

In Equation 1 speed is related to density through an equilibrium relationship:

$$u = u_e(k) \tag{2}$$

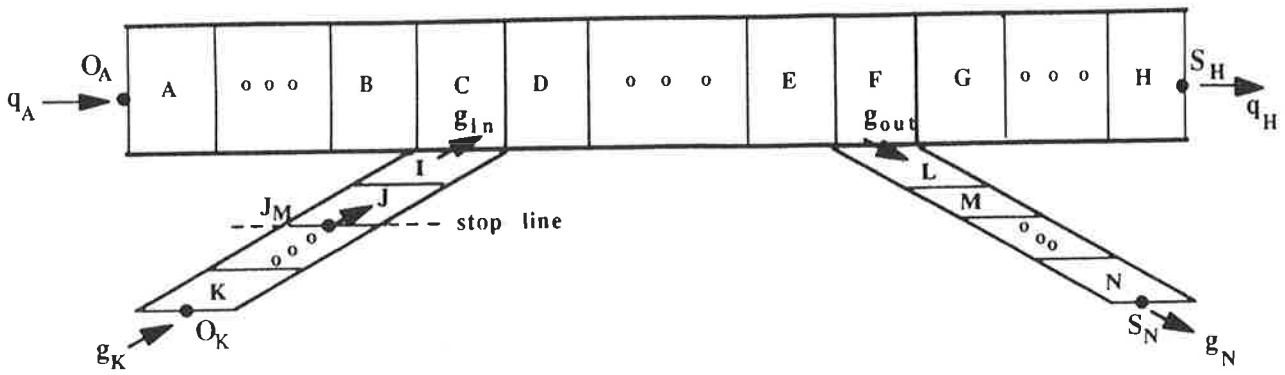


FIGURE 1 Space discretization of a typical freeway section.

metering rate, Link J is treated the same as the other links in Item 2; when g_K^{n+1} > metering rate, Link J is treated as an internal boundary and g_J^{n+1} = metering rate.

If the entrance ramp is not metered, Link J is treated the same as the links in Item 2.

4. Link C: Link C represents a freeway junction with an entrance ramp. Since there is a merging flow, Equations 6 through 11 need to be modified. A generation term g_{in} will be added to the right-hand side of Equations 6 and 9. $g_{in} = q_{in}/\Delta x_c$, where $q_{in} > 0$ is the merging flow from Link I to Link C. In addition, the viscosity term in Equation 13 will be added to the right-hand side of Equations 7 and 10.

Determination of g_{in} is explained in Item 5.

Equations 8 and 11 remain unchanged.

5. Link I: Link I connects the ramp to the freeway. On one hand, this link provides ramp volume q_{in} to Link C. On the other, it serves as the downstream boundary of the ramp being considered as a pipeline. In general merging volume q_{in} is not equal to ramp demand q_K, especially when freeway is congested and there is a waiting queue at the ramp. Since we consider freeway and ramp as being connected at a single point, it is not necessary to use the same treatment of merging dynamics used in our earlier work (*12*), in which Link I and Link C are further discretized into small Δx's. However, in seeking the relationship of the maximum merging flow and the freeway mainline flow, the fundamental rule that merging value is governed by the gap availability on mainline freeway is still followed. Figure 2 shows the merging volume versus the Lane 1 (the rightmost lane) volume (q_1) of freeway from our field data collected at eight entrance ramps (on two- and three-lane freeways) in 14 peak periods. The volume is measured by loop detectors (installed after the metering stop line) when its corresponding occupancy is so high that it suggests that there are excessive cars at Links J and I waiting to merge pending the gaps in the mainline flow. In order to curve-fit the merging capacity C_g with the freeway Lane 1 volume, we have modified one of the merging capacity equations by Adams (*13,14*) and applied the least-square technique to fit the upper half of the curve in Figure 2, which applies to uncongested mainline flow situations. When freeway becomes congested, density increases, resulting in reduction of mainline flow rate and available space for vehicles to merge from ramps. Merging capacity therefore decreases with the reduction of mainline flow. Unfortunately, we have not been able to obtain enough data to estimate the relationship of merging

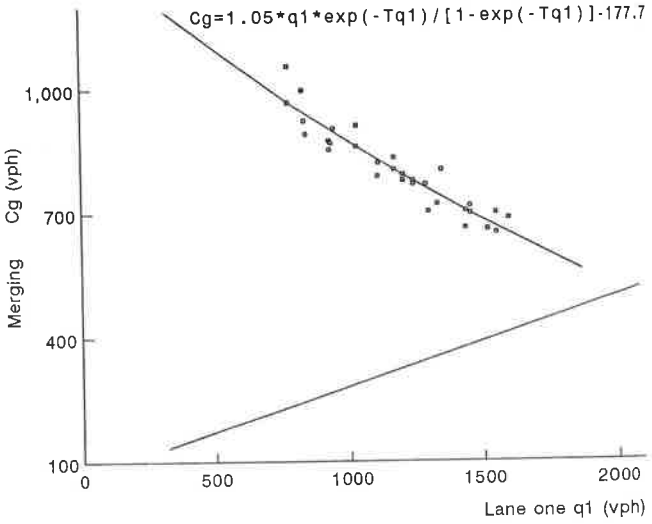


FIGURE 2 Merging capacity at Link I.

capacity and merging volume at congested flows. Experimentally, a straight line is used as an approximation to this relationship on the basis of traffic operation practices of the Traffic Management Center of the Minnesota Department of Transportation. Consequently, q_{in} is determined as follows:

$$q_{in} = \min \{q_1, C_g, (Cap_C - q_C)\}$$

where q_{in} is the flow rate on Link I and C_g is determined from the two empirical curves in Figure 2. Cap_C is the capacity of Link C and q_C is the total mainline flow before merging takes place.

6. Links L and F: Similar to Link I, Link L receives diverging volumes from freeway to surface streets, and it serves as the upstream boundary of the exit ramp. All exiting demand is assumed to leave freeway through Link L.

The exiting volume from Link F to Link L is in general not greater than the capacity at the downstream end of the exit ramp. Specifically,

$$q_{out}^{n+1} = \min \{q_{exit}^{n+1}, q_M^n, Cap_L\}$$

where q_{exit}^{n+1} is the exiting demand, g_M^n is the flow rate on Link M at the nth time step, and Cap_L is the capacity of Link L.

Congestion may spill back from Link L to Link F when $q_{exit} > q_{out}$. In this case, the through capacity of Link F is reduced. First, the cumulative exiting demand at mainline freeway is determined as follows:

$$(ST)^{n+1} = (ST)^n + (g_{exit}^{n+1} - g_{out}^{n+1}) * \Delta t/3{,}600 \qquad (ST)^{n+1} \geq 0$$

where ST is the cumulative exiting demand remaining on the mainline freeway. Second, the through capacity TC of Link F becomes

$$TC_F = Cap_F * (LN_F - 1)/LN_F$$

where Cap_F is the capacity of Link F under normal conditions and LN_F is the number of lanes at Link F.

Once q_{out}^{n+1} is obtained, Link L is treated as a pipeline section and the associated equations discussed before for pipeline freeways will be used. Link F is treated the same as Link C (except now $g_{out} < 0$) if there is no congestion spillback and is treated as a pipeline section with reduced capacity for the through traffic if such a spillback prevails.

The modeling presented allows simultaneous treatment of the freeway and its ramps in an integrated fashion. This feature is especially important in future development of this model. For example, in modeling a freeway corridor, both surface streets and the freeway, which are connected through the ramps, will be processed in a uniform and integrated manner. The numerical scheme used to implement the model provides a stable difference approximation with first order accuracy in both Δx and Δt. The model does not include a $u_e(k)$-k relationship, and the arrival and departure patterns are the only inputs to the model that can be in any form, including stochastic ones.

TESTING AND FIELD VALIDATION

To evaluate the effectiveness of the model implementation, we present the results of our model in both qualitative and quantitative tests. Whereas the former is done by applying the model to a hypothetical situation, the latter is based on field data involving basic freeway segments and entrance and exit ramps.

Qualitative Testing

An effort was first made to examine whether the model is physically reasonable in representing traffic dynamics, especially at congested flows. The test scheme involves a 2-mi, three-lane pipeline freeway with demand at upstream boundary and capacity at downstream shown in Figure 3. Starting from $t = T_a = 4$ min 0 sec, downstream capacity is reduced by about one-half because of an incident. A bottleneck is therefore created and remained until $t = T_b = 7$ min 0 sec, when the incident is cleared and the initial capacity at the downstream boundary is restored. Figure 4 shows the density distribution given by the model at four time instants during and after the incident. It can be observed that congestion

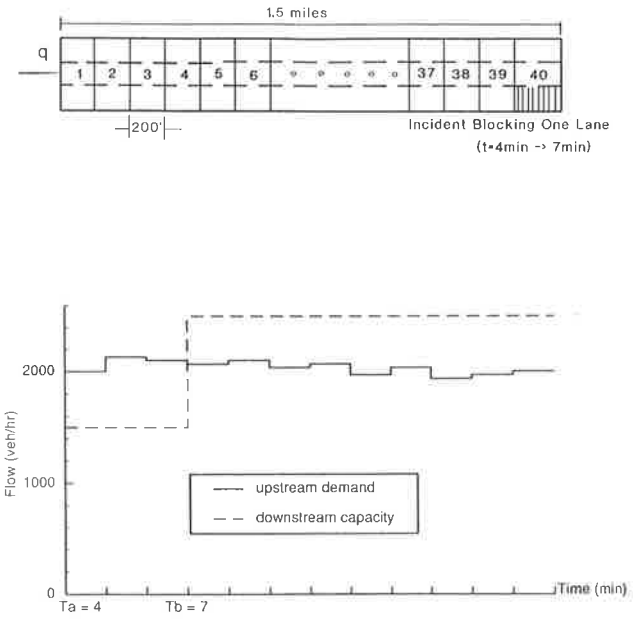


FIGURE 3 Geometry and boundary conditions for qualitative testing.

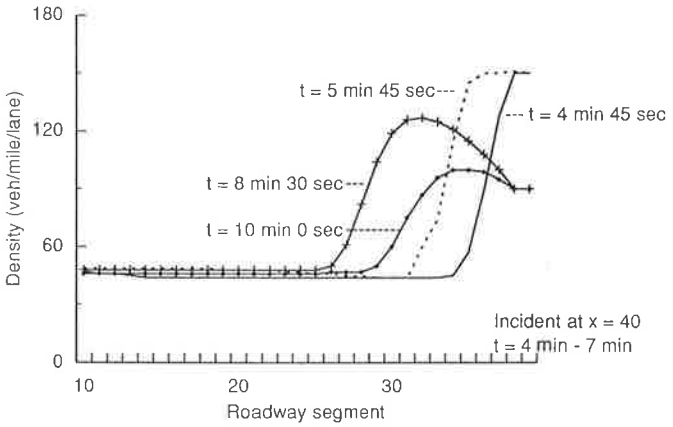


FIGURE 4 Queue propagation (during incident) and dissipation (after incident).

propagates to the upstream freeway at $t_1 = 4$ min 45 sec and $t_2 = 5$ min 45 sec, and is gradually smoothed out toward the downstream freeway at $t_3 = 8$ min 30 sec and $t_4 = 10$ min 0 sec in the dissipation process.

Testing on Pipeline Freeway

Field validation of the model is done by using data collected from the I-35W in Minneapolis. In a recent research project funded by the Minnesota Department of Transportation, an 8-mi section of the freeway was selected that connects downtown Minneapolis to the southern suburban areas and contains a variety of geometric types, such as entrances, exits, weaving areas, and so forth. The test scenario (Case 1) involves a four-lane freeway close to downtown Minneapolis

that carries southbound traffic from 4:00 to 6:00 p.m. Congestion starts at 4:10 p.m. at the downstream boundary and reaches the upstream boundary by 4:15 p.m. The freeway remains congested until 6:00 p.m., when congestion dissipates through the downstream boundary. The arrival and departure traffic patterns (boundary conditions) are shown in Figure 5. The initial traffic condition of the system is obtained from data at the time interval before the start of simulation. There are three mainline detection stations at the test site. Two stations at the upstream and downstream boundaries provide boundary conditions at every 5-min interval. The remaining station ("check station" in Figure 5) is located between the boundaries, and it provides measurements to compare with the simulation results.

To evaluate the quantitative effectiveness, we included in the testing three models [simple continuum, Payne, and Papageorgiou (*15*)] together with our proposed model (upwind version), and they are referred to as Models 1, 2, 3, and 4, respectively. Whereas the simple continuum model was implemented through the Lax (*16*) scheme, Models 2 and 3 were discretized by using the Euler (*3,15*) scheme as recommended by the original authors. A different implementation of our model (Lax version) was also included. Our model was properly calibrated by using field data, and in both versions $\Delta x = 200$ ft and $\Delta t = 1$ sec were used. On the basis of the deviations of the estimated results from the field observations, the following statistics are calculated:

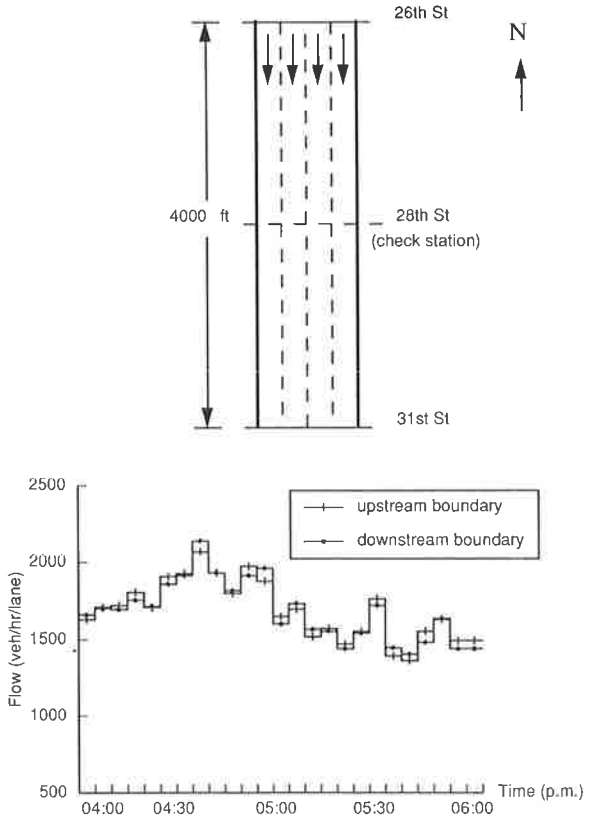


FIGURE 5 Geometry and arrival and departure patterns for Case 1.

$$\text{Mean absolute error (MAE)} = (\sum_{1}^{N}|\text{observed} - \text{estimated}|)/N$$

$$\text{MAE (\%)} = \sum_{1}^{N}(|\text{observed} - \text{estimated}|/\text{observed})/N$$

$$\text{Mean square error (MSE)} = [\sum_{1}^{N}(\text{observed} - \text{estimated})^2]/N$$

$$\text{St. deviation} = \{[\sum_{1}^{N}(\text{observed} - \text{estimated})^2]/(N - 1)\}^{1/2}$$

where N is the number of observations.

Computational efficiency is also compared among all the included models. Computer execution time was measured from an IBM PC (results could be machine dependent), and an index (Comp Index) was used to indicate the ratio of computation time with respect to Model 1.

Test results from Case 1 are summarized in Table 1, where only the error indices for volumes are presented since speed data at the check station were not available. Table 1 indicates the following:

1. When there is downstream congestion, all high-order models included performed substantially more accurately than the simple continuum model.
2. Model 3 was more accurate than Model 2, both of which are implemented with the same method (Euler).
3. Model 4 (upwind version) was the best overall in terms of accuracy, and it was faster than Models 2 and 3 by 20 to 25 percent and faster than Model 1 by 14 percent.
4. The Lax version of Model 4 was not as good as its upwind version because it not only produced larger errors but also required more computation time.
5. Model 4 is faster than Model 1 even for the same numerical method (Lax), mainly because of the absence of a u_e-k table to look up.

Testing with Entrance/Exit Ramps

Two additional test cases are presented in this section that involve merging/diverging flows at freeway junctions with on-

TABLE 1 Error Indices and Computational Efficiency for Case 1

Models	1	2	3	4 (upwind)	4 (Lax)
MAE (a)	82	59	46	23	34
MAE(%)	15.26	8.10	7.84	4.46	6.43
MSE (b)	5871	4662	3725	808	1470
Std. Dev.	90.02	74.80	66.85	31.15	42.00
Comp Index(c) (Time Improv)	1.00	1.07 (-7%)	1.10 (-10%)	0.86 (+14%)	0.97 (+3%)

(a) MAE, veh/5 minutes;
(b) MSE, veh²/5 minutes; Std. Deviation, veh/5 minutes;
(c) Computer execution time index, compared with Model 1 and based on IBM-PC 386-25Mhz machine.

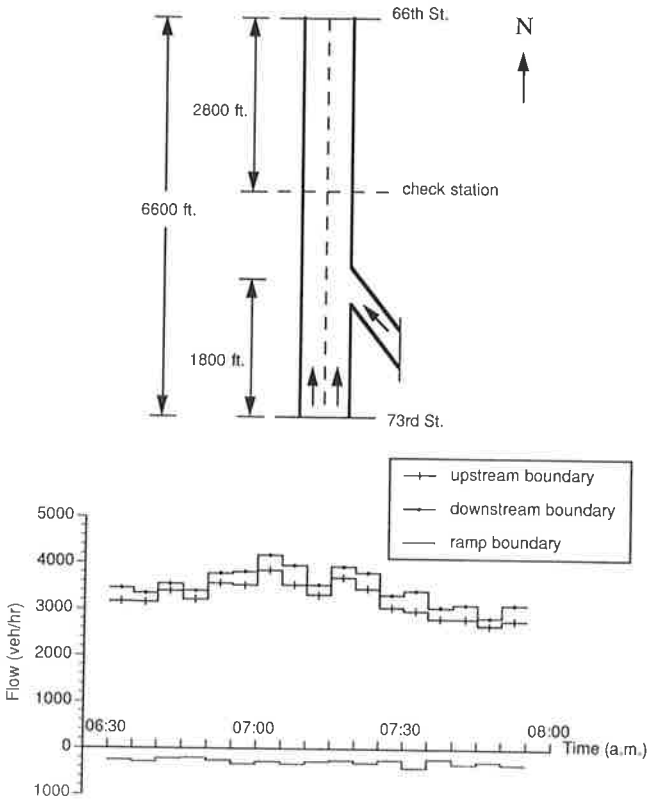


FIGURE 6 Geometry and arrival and departure patterns for Case 2.

trance/exit ramps. Case 2 is based on a two-lane freeway with morning traffic from 6:30 to 8:00 a.m. Case 3 is also a two-lane freeway with an exit ramp carrying northbound traffic from 7:00 p.m. to 8:20 a.m. The roadway geometry and traffic patterns at boundaries and the demand at the entrance ramp for Case 2 are shown in Figure 6.

Since it is not clear how other high-order models are implemented at freeway ramp junctions, we have included in addition only the simple continuum model in the testing. Furthermore, since similar error statistics are obtained for Cases 2 and 3, only those from Case 2 are summarized in Table 2 because of space limitation. To understand the effect of variation in the mesh size, results from the application of several Δx and Δt values have been presented. The results are as follows:

1. In Cases 2 and 3, the high-order model is able to estimate traffic volumes at a relatively high level of accuracy. Among the volume errors, 2.03 to 2.05 percent was produced in Case 2 and 3.51 to 3.54 percent in Case 3.
2. The simple continuum model does not have the same accuracy as the high-order model: 9.7 to 11.9 percent of volume error was produced in Case 2 and 10.4 to 13.5 percent in Case 3.
3. Speed estimations in both models produced larger errors. Whereas the high-order model generated 3.7 percent of errors in Case 2 and 4.09 to 4.1 percent in Case 3, the simple continuum yielded 9.3 to 10.3 percent in Case 2 and 13.3 to 13.6 percent in Case 3.
4. The simple continuum model seems to be sensitive to the changes in Δx and Δt sizes. With the increase of Δx, the errors in traffic volumes (MAE, MSE, and Std. Dev.) increase.
5. For the high-order model, changes in Δx and Δt sizes have no obvious effect on the errors in both volumes and speeds (when $\Delta x \leq 500$ ft). All MAE, MSE, and Std. Dev. remained about the same. This feature produces flexibility to choose the level of discretization according to the purpose of the application while operating the model at about the same error level.

CONCLUSIONS

An improved high-order continuum model was proposed and implemented with very encouraging results. A traffic friction term is included to handle the effect of vehicular interactions at ramp junctions. The use of an explicit $u_e(k)$-k relationship is not needed. This feature saves a significant amount of effort for acquiring such a relationship, making our model more practical for field applications. Qualitative testing showed good capability of describing queue propagation and dissipation properties.

For a pipeline situation our model was compared with a simple continuum model and other high-order models and produced lower error. In general, high-order models seem to be significantly better than the simple continuum model in

TABLE 2 Error Indices for Case 2

Tests(a)	dx=100, dt=1		dx=200, dt=1		dx=300, dt=1		dx=400, dt=2		dx=500, dt=3	
	vol	spd	vol	spd	vol	spd	vol	spd	vol	spd
MAE(b)	6 (28)	2.1 (5.6)	6 (33)	2.1 (5.0)	6 (34)	2.1 (5.0)	6	2	6	2
MAE (%)	2.03 (9.7)	3.72 (10.3)	2.04 (11.7)	3.72 (9.3)	2.05 (11.9)	3.72 (9.3)	2.05	3.72	2.04	3.71
MSE(c)	46 (1061)	5.9 (34.7)	46 (1455)	5.9 (29.1)	47 (1482)	5.9 (29.2)	46	5.9	46	5.9
Std. Dev.(d)	7.01 (33.6)	2.51 (6.1)	7.00 (39.4)	2.51 (5.6)	7.04 (39.8)	2.51 (5.6)	7.02	2.51	6.98	2.51

(a) Numbers in brackets are from simple continuum model.
(b) MAE, veh/5 minutes for volumes, mile/hr for speeds.
(c) MSE, veh^2/5 minutes for volumes, (mile/hr)^2 for speeds.
(d) Std. Dev. veh/5 minutes for volumes, mile/hr for speeds.

congested situations, although further testing is needed before such a generalized conclusion can be made. Finally, the choice of the numerical method (e.g., upwind versus Lax) seems to have substantial impact on accuracy and efficiency.

A simplified modeling methodology suitable for freeway pipeline plus entrance/exit ramps was used. Two test cases were shown, and our model produced lower errors compared with the simple continuum model.

The overall performance of our model was very promising. However, more field testing/validation is needed to improve our model on more complicated geometrics and by using more detailed traffic measurements (shorter time and space increments). As a future possibility, traffic data can be collected from AUTOSCOPE (*17*), which is currently being developed at the University of Minnesota and will include 38 video detectors by 1993 along a 2.5-mi section of the I-394 freeway in Minneapolis (*17*). This freeway section will serve as a laboratory for collecting and studying traffic characteristics and testing and validating traffic flow models, including the one presented in this paper.

ACKNOWLEDGMENTS

Financial support for this research was provided by Minnesota Department of Transportation.

REFERENCES

1. M. H. Lighthill and G. B. Whitham. On Kinematic Waves: II. A Theory of Traffic Flow on Long Crowded Roads. *Proc. R. Soc. London*, Ser. A. 229, 1955, pp. 317–345.
2. H. J. Payne. Models of Freeway Traffic and Control. In G. A. Bekey, Mathematical Models of Public Systems, Simulation Counc. Proc. Ser., 1, 1971, pp. 51–61.
3. H J. Payne. FREFLO: A Macroscopic Simulation Model of Freeway Traffic. In *Transportation Research Record 772*, TRB, National Research Council, Washington, D.C., 1979, pp. 68–75.
4. W. F. Phillips. *A New Continuum Model for Traffic Flow*. Report DOT-RC-82018. Utah State University, Logan, 1979.
5. N. A. Derzko, A. J. Ugge, and E. R. Case. Evaluation of a Dynamic Freeway Model Using Field Data. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 52–60.
6. A. K. Rathi, E. B. Lieberman, and M. Yedlin. Enhanced FREFLO Program: Simulation of Congested Environments. In *Transportation Research Record 1112*, TRB, National Research Council, Washington, D.C., 1987, pp. 61–71.
7. P. Ross. Traffic Dynamics. *Transp. Res. B*, Vol. 22B, No. 6, 1988, pp. 421–435.
8. P. G. Michalopoulos, P. Yi, D. E. Beskos, and A. S. Lyrintzis. Continuum Modelling of Traffic Dynamics. *Proc., Second International Conference on Applications of Advanced Technologies in Transportation Engineering*, ASCE, Minneapolis, Minn., 1991.
9. G. J. Forbes and F. L. Hall. The Applicability of Catastrophe Theory in Modelling Freeway Traffic Operations. *Transpn. Res. A*. Vol. 24A, No. 5, 1990.
10. C. Hirch. *Numerical Computation of Internal and External Flows*. Vol. 1. John Wiley and Sons, 1990.
11. I. Prigogine and R. Herman. *Kinetic Theory of Vehicular Traffic*. American Elsevier, New York, 1971.
12. P. G. Michalopoulos, E. Kwon, and J. G. Kang. Enhancement and Field Testing of a Dynamic Freeway Simulation Program. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991.
13. W. F. Adams. Road Traffic Considered as a Random Series. *J. Inst. Civil Eng.*, Vol. 4, 1936, pp. 121–130.
14. D. L. Gerlough and M. J. Huber. *Special Rept 165: Traffic Flow Theory*. TRB, National Research Council, Washington, D.C., 1975.
15. M. Papageorgiou, J. M. Blosseville, and H. Hadj-Salem. Macroscopic Modelling of Traffic Flow on the Boulevard Peripherique in Paris. *Transpn. Res. B*, Vol. 23B, 1989, pp. 29–47.
16. P. D. Lax. Weak Solution of Non-Linear Hyperbolic Equations and Their Numerical Computations. *Communs. Pure Appl. Math.*, Vol. 7, 1954, pp. 159–173.
17. P. G. Michalopoulos, B. Wolf, and R. Benke. Testing and Field Implementation of the Minnesota Video Detection System. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.

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

Variance Reduction Applied to Urban Network Traffic Simulation

AJAY K. RATHI AND MOHAN M. VENIGALLA

The effectiveness of variance reduction techniques that users can apply to improve the efficiency and reliability of simulation experiments with the TRAF-NETSIM simulation model is described and illustrated. The two variance reduction techniques, antithetic variates and common random numbers, reduce the variance of simulation output by replacing the original sampling procedure by a new procedure that yields the same parameter estimate but with a smaller variance. Thus, the users can obtain greater statistical accuracy for the same number of simulation runs. A recent modification of the stochastic sampling process has made the TRAF-NETSIM model amenable to these variance reduction techniques and allows the users to apply these techniques with minimal additional effort. The effectiveness of these techniques is evaluated through an analysis of simulation output data from a TRAF-NETSIM case study. The estimated values and variances are computed for some representative measures of effectiveness after 10, 20, and 30 replications. The results indicate that both techniques are effective in reducing variance of the model output. By using the variance reduction techniques, the variance of parameter estimates is reduced on the average by 65 percent in the 24 comparisons that are made. The common random numbers strategy is more effective than the antithetic variates procedure. Over 50 percent reduction in variance is obtained using the common random numbers strategy in all comparisons and 80 percent or more in 6 of the 12 comparisons. In all cases studied, better statistical precision is obtained by making two-thirds fewer simulations than under conventional multiple replications–based experimentation.

Simulation models can be helpful in many contexts. By creating and executing the model, analysts and system managers can get a better understanding of the complex process that is being simulated. That is, one can get answers to questions such as, How does the system respond to changes in input data or operating rules? Often, the simulation models are used to evaluate alternative system designs or operating policies. Simulation models can also be used to develop optimal or near-optimal system designs or policies when used in conjunction with appropriate analytical methods. Above all, simulation eliminates the need for costly, time-consuming, risky, or sometimes infeasible field experiments and evaluations before real-life implementation on a new system management or control strategy.

Given these attractions of simulation, it is not surprising that there has been a considerable increase in the uses of simulation as a decision support tool in transportation applications during the past few years. Models such as TRAF-NETSIM (*1–3*) are now used routinely to solve a wide range of transportation problems. Such increases may also be attributed to greater complexity of problems under consideration; greater accessibility and reduced computing costs resulting from the availability of simulation models on microcomputers and workstations; improvements in simulation models; and availability of graphical animation and other support utilities, which result in greater understanding and use of these models by engineers, planners, and analysts.

The widespread use of simulation models, in turn, raises serious questions about their proper usage. Is the model output understood and analyzed properly? Are the conclusions valid? Are the model predictions properly qualified? Is the design of experiments robust? Such questions become critical in the case of microscopic, stochastic models such as TRAF-NETSIM, which could exhibit considerable variability between replications. The concern about misuses of simulation models is heightened by the "black box" treatment of the models by the users—one that frequently implies the lack of understanding of either the stochastic simulation process or the statistical aspects of simulation experimentation. The model builders, on the other hand, are too often more concerned with building an appropriate model than with addressing the issues related to the analysis of output data that the model generates.

This paper deals with the analysis of output data generated by the TRAF-NETSIM model and contains information that should assist the model users in performing the simulation experiments intelligently. The paper presents methods that the users can use to improve the efficiency and accuracy of the simulation experimental process. These methods, known as variance reduction techniques, reduce the variance of a parameter estimate (increase accuracy and thus credibility), or equivalently allow the parameter estimation with same variance using fewer runs (increase efficiency), by controlling the random number seeds used to drive the simulation model. The methodology presented here is applicable to stochastic simulation models using Monte Carlo procedures where the random behavior observed in the simulation experiments is completely under the control of the user. These techniques can also be used with other traffic simulation models that are based on a philosophy similar to TRAF-NETSIM, for example, FRESIM (*4*). Furthermore, the two variance reduction techniques described in this paper are applicable in situations where the analyst is interested in comparing alternative system designs or policies. One of these techniques can also be used in estimating the performance characteristics of an individual system.

A. K. Rathi, Oak Ridge National Laboratory, P.O. Box 2008, MS 6366, Oak Ridge, Tenn. 37831. M. M. Venigalla, University of Tennessee, 10521 Research Drive, Suite 200, Knoxville, Tenn. 37932.

PREVIOUS RESEARCH

Computer simulation models of traffic operations analysis have now been in existence for more than four decades. During this period, a number of sophisticated computer programs capable of simulating a wide array of traffic operations, network configurations, and control policies have been developed (*5*). However, only a handful of these models have been used beyond their development environment since most of these models are limited in scope. The need to synthesize the state of the art and to model virtually all traffic situations within a single software system led to the idea of the integrated traffic simulation systems in the late 1970s (*6*). By the mid-1980s, a version of FHWA's integrated traffic simulation system, named TRAF, was ready for beta testing (*1*). Following FHWA's lead, other prominent model families are now evolving as systems capable of simulating traffic operations on surface streets as well as freeways in an integrated fashion (*7,8*). Whereas the model developers have focused on integrated systems in the 1980s, the application of computer simulation models has increased considerably during this period as the potential of the simulation is realized by practitioners and researchers alike.

Despite the advances in the framework and methodology of traffic simulation models and their increasing usage, the area of statistical analysis of simulation output data and various other aspects of simulation experimental process (e.g., ranking of alternatives) has received very little attention. The model outputs generally do not include statistics other than sample means. On the other hand, the user generally has very limited control over input parameters that will allow for an efficient undertaking of simulation experiments. As a result, the application of simple concepts such as multiple replications is uncommon in traffic simulation studies, and the variance or confidence interval of parameter estimates is seldom computed and reported. There are no guidelines for conducting simulation experiments with the models that are used extensively. The review of literature reveals a handful of references dealing with the subject matter. On the other hand, the methods for analyzing output data from simulations, especially discrete event simulations, are widely used in the practice of simulation in manufacturing, queuing systems, distribution systems, and various other applications (*9*). The application of variance reduction techniques goes back as early as the 1960s, when Tocher (*10*) suggested the use of these techniques in the simulation of complex industrial systems.

The first demonstrated analysis of output data from a traffic simulation model was presented by Rathi and Nemeth (*11*). This study illustrated the effectiveness of variance reduction techniques for a freeway simulation model. It showed that significant variance reduction can be obtained by exploiting the random number seed in a microscopic simulation model. The implementation of variance reduction techniques for the TRAF-NETSIM simulation model described in this paper is essentially an extrapolation of that work. In the early 1980s, a research project was sponsored by FHWA to provide statistical guidelines for simulation experiments with emphasis on the NETSIM model. The work performed as part of this project is described in several reports (*12–14*). These reports contain an excellent discussion of the nature of stochastic simulation experiments and statistical issues relating to the analysis of simulation data. In terms of parameter estimation, the bulk of this work focused on the efficacy of the "ratio of means" estimation procedure for the NETSIM output. However, it was never made clear how the users might employ the suggested procedure to their advantage, and the effectiveness of the procedure was not demonstrated in simulation experiments with NETSIM. That is, the study addressed the efficiency of simulation experiments vis-à-vis precision of parameter estimation but did not provide any implementable methodology to assist the users. As a result, much of this work remains academic in nature.

More recently, Chang and Kanaan (*15*) provided an assessment of the variability of TRAF-NETSIM output and tested the efficiency of the "batch means" method in parameter estimation. The results showed that multiple replications provided better coverage of sample means than the batch means approach. However, the coverage after 50 replications (simulation run length not known) was compared with batch means runs of only 90 min. Therefore, the results could have been different had fewer replications been performed. This concept of batch means–based variance reduction is appealing, but determining the batch size in a given situation is not a straightforward process. Chang and Kanaan (*15*) concluded that considerable work is necessary to establish guidelines for using the batch means procedure of parameter estimation. Therefore, this work has little utility to the end users at this time.

Finally, TRAF-NETSIM simulation model was recently modified to generate traffic streams exhibiting the same routing patterns, driver-vehicle characteristics, and certain other operational characteristics through a series of simulation runs (*16*). Thus, the users can make a series of simulation runs "under the same circumstances" by retaining the traffic stream of an initial run while performing simulations under different traffic controls, volumes, or any other operational conditions during the subsequent runs. This enhancement, referred to as the "identical traffic stream," also makes the model amenable to variance reduction concepts that are based on random number seeds controlled experiments. The effectiveness of variance reduction based on one of these concepts is described in a forthcoming technical note (*17*). In this paper, we have considered both the antithetic variates and common random number concepts and compared their relative effectiveness using the identical traffic stream feature of TRAF-NETSIM.

The variance-reduction techniques described in this paper have been well established in the literature. However, they are being used for the TRAF-NETSIM model for the first time, made possible by the identical traffic stream feature of the model. Also, for the first time, the users can actually employ these techniques in their simulation experiments, unlike past efforts, which are purely academic in nature. Therefore, the contents of this paper are useful to both the experienced simulationists and the novice users of the TRAF-NETSIM model.

VARIANCE PROBLEM IN TRAF-NETSIM

TRAF-NETSIM offers many conveniences that make it an attractive tool for evaluating a wide spectrum of traffic management strategies for urban street networks. The model pro-

vides the highest level of detail and accuracy of any existing empirical technique or simulation model of traffic operations. The availability of this model has provided the opportunity for the development and testing of new and innovative traffic management concepts and designs.

However, TRAF-NETSIM simulation output contains much variability from one simulation run to the next. The variability reflects the stochastic nature of the simulated environment and is generally more pronounced due to the microscopicity (simulation of individual vehicles) of the model. Of course, the amount of variability experienced in a given situation depends on several factors including the initialization time, simulation time, traffic volume, network geometry, the presence of incidents or disruptions in the simulated network, and other operational features of the simulated environment. An illustrative example of the TRAF-NETSIM output variability is provided by Chang and Kanaan (*15*). The average delay on a link of a sample data set is found to vary from 128.3 to 409.4 sec in 10 independent runs.

The model has often been used to compare system alternatives (i.e., system designs or operating policies). In this context, the presence of high variance of the response variable (i.e., measure of effectiveness) is the worst enemy of the experimenter. Since variance is inversely proportional to the sample size, the model must be run longer or be replicated many times to achieve a desired precision level, both of which are costly undertakings. The presence of high variance means that seemingly large differences in the system's performances may not be statistically significant. High variability in model output can also lead to concern about the model's reliability.

To illustrate the importance of the variance problem, an analysis of the data from Chang and Kanaan (*15*) is performed. Table 1 shows the simulation results obtained in that study. The average of 10 replications after 60 minutes of simulation can be used to obtain an estimate of the delay per vehicle, $\overline{D}_1$. The estimates of variance and standard deviation, $V(D_{1r})$ and $S(D_{1r})$, for the response variable (delay per vehicle in this case) can also be computed using standard statistical formulas. The results are

$$\overline{D}_1 = 276.28 \quad V(D_{1r}) = 6461.35 \quad S(D_{1r}) = 80.38 \tag{1}$$

TABLE 1 Average Delay Generated by 10 Independent Runs (*15*)

Replication	Delay (sec) on Link 23-03			
	15 min	30 min	45 min	60 min
1	112.1	180.4	230.3	261.5
2	99.0	128.0	118.6	128.3
3	145.2	221.8	293.0	347.4
4	170.3	238.1	264.2	279.3
5	142.8	251.5	367.9	409.4
6	138.0	154.9	164.7	195.5
7	144.9	145.3	192.2	242.8
8	252.8	299.3	345.5	333.8
9	206.0	242.9	280.2	315.7
10	113.1	180.7	232.6	249.1

where D_{1r} is the delay per vehicle estimated in the rth replication.

From these estimates, one can construct a 95 percent confidence interval for the response variable, an interval within which we are 95 percent confident that the delay per vehicle value exists. For n replications, approximate values for the upper and lower confidence limits are

$$\text{Lower confidence limit} = \overline{D}_1 - \frac{2S(D_{1r})}{\sqrt{n}} = 225.44 \tag{2}$$

$$\text{Upper confidence limit} = \overline{D}_1 + \frac{2S(D_{1r})}{\sqrt{n}} = 327.11 \tag{3}$$

Considering an average delay per vehicle of 276.28 sec, a confidence interval of 101.67 (327.11 −225.44) is quite wide. The situation gets worse when one realizes that this may be the base case or alternative scenario of an experiment where the objective is to compare the performance of the two designs or policies. The standard experimental approach is to obtain independent results for the two alternatives and to compare the performance of the two systems by computing the mean, standard deviation, and variance for each case. The variance of the difference between the estimated values in a base case and alternative scenario is obtained by

$$V(\overline{D}_1 - \overline{D}_2) = V(\overline{D}_1) + V(\overline{D}_2) \tag{4}$$

$$= \frac{V(D_{1r}) + V(D_{2r})}{n} \tag{5}$$

where $(\overline{D}_2)$ is the estimate of delay per vehicle in the other scenario. All other notations are the same as for $(\overline{D}_1)$.

Assume for simplicity that the variance of the second case study is the same as the first study. Using Equation 5, the variance and standard deviation of the difference in expected value of delay per vehicle will be 1,292.27 [(6,461.35 + 6,461.35)/10] and 35.94, respectively. The 95 percent confidence interval for difference in expected value of delay will approximately be 142 sec. Therefore, even a change (plus or minus) in 71 sec of delay per vehicle would not be statistically significant. This confidence interval would be even wider if fewer replications are made. Therefore, many simulation runs may be required before valid conclusions are drawn.

This example is given to illustrate the variability in the TRAF-NETSIM simulation output and to suggest that statistical analysis must be used to properly interpret the simulation output data. Regardless of the type of simulation software used to perform the analysis, the use of simulation output data for decision making should be approached with care. The stochastic variations as well as fundamental system behavioral characteristics must be carefully analyzed before drawing conclusions from a simulation-based study. For instance, looking at the data in Table 1, one realizes that the simulated system has not reached a steady-state condition even after 60 min of simulation because delay is gradually increasing. This implies that there are either certain characteristics of the simulated network that serve to gradually increase the delay per vehicle over time or the system simply has not reached a steady-state condition. In a situation such as this, one might even want to investigate the validity of the

output data by examining the way in which such data are tallied during the simulation.

GENERATION OF RANDOM BEHAVIOR IN THE TRAF-NETSIM MODEL

TRAF-NETSIM is a discrete event simulation model of the dynamics of traffic operation in a network of urban streets. The vehicles are represented individually and their operational performance is determined every second. Furthermore, each vehicle is identified by a category (automobile, carpool, bus, or truck), a type (up to 16 different vehicle types with different operating and performance characteristics) within each category, and by a driver behavioral characteristic (passive, normal, or aggressive).

The physical environment in TRAF-NETSIM is represented as a network comprised of links and nodes. Generally, the nodes of the network represent intersections, and links represent one-way urban streets. The vehicles enter the network through links designated as entry links and are moved each second according to the car-following logic while responding to traffic control devices, pedestrians, neighboring vehicles, and other conditions that influence driver behavior.

TRAF-NETSIM simulates traffic flow on urban street networks by representing the movement of individual driver-vehicle combinations. The model uses a number of stochastic processes (or random sampling from discrete distributions) to represent the real-world behavior. For each vehicle entering the network, the vehicle and driver characteristics are generated randomly. As vehicles move from one link to another, their turning movements on the new link are randomly assigned while satisfying the link-specific turn movement percentages. Many other behavioral and operational decisions (e.g., free-flow speed and gap acceptance) are represented as random processes.

Like most simulation models using Monte Carlo procedures, the generation of random behavior in the TRAF-NETSIM model consists of a three-step process. The first step consists of generating a random number. Then a uniform deviate is generated from this random number. This is followed by the generation of a random observation from the probability distribution of interest. The model uses the initial random number seed as the basis for all stochastic decisions (random sampling) in the simulation process as described below.

Random number generation in TRAF-NETSIM is based on a linear recursive procedure, usually attributed to Lehmer (*18*). Using a multiplicative congruence technique, a sequence of random numbers is generated by always calculating the next random number from the last one obtained, given an initial random number (called the "seed"). In particular, the $(n + 1)$th random number X_{n+1} is calculated from the nth random number X_n by using the recursive relation

$$X_{n+1} = \text{mod}[(a * X_n + \text{mod}(X_n,k) * k), m] \quad (6)$$

where a, k, and m are positive integers. In the current implementation of TRAF-NETSIM, a value of 3 is used for a, 10,000 for k, and 100,000,000 for m. This procedure generates a "random" number between $k + 1$ and $m - 1$ (i.e., a number between 10,001 and 99,999,999). From this random number, a uniform deviate U between (0, 99) is obtained by dividing X_{n+1} by another integer p (1,000,000). Thus, a random number between 10,000 and 100,000,000 is generated by using the random seed and the desired random deviate is obtained by dividing it by 1,000,000 (integer to integer division). Since the numbers between 10,000 and 100,000,000 are generated randomly and nonrepeatedly, the procedure is able to generate a sequence of purely random deviates. Note that the integer p is selected such that U is between 0 and 99 and that its possible values are in direct proportion to its respective probabilities (i.e., 1 in 100). The uniform deviate U is then used to generate a random observation from the tabular decile probability distributions used in TRAF-NETSIM. This is accomplished by simply accessing the values stored in global arrays corresponding to the uniform deviate (*1*). Since the random number generation is a recursive procedure, the initial random number (seed) forms the basis for all stochastic decisions in the model.

The initial value of the random number seed X_0 is input by the user. The user is asked to provide an odd number (except one ending with 5) of up to eight digits. Selection of an odd number (except 5) for a as well as X_0 guarantees a full period for the random numbers. That is, the procedure will "randomly" generate, without repeating, all integer, nonnegative numbers between 10,001 and 99,999,999.

VARIANCE REDUCTION TECHNIQUES

Variance reduction techniques, as used in Monte Carlo studies and simulations, are based on the premise of replacing the original sampling ("crude" or "unplanned" sampling) procedure by a new procedure that produces the same parameter estimate but with a smaller variance. Some of these techniques replace the sampling process completely (e.g., importance sampling), whereas others simply use a different estimator (e.g., ratio estimators, batch means) from the sample average produced by the crude sampling method (*9*). However, perhaps the most subtle modification of the sampling process is used in variance reduction techniques that operate by controlling the random number seeds. These two techniques, common random numbers and antithetic variates, have proven to be effective and can be implemented with hardly any extra effort on the part of the user (*9,11*). These techniques take advantage of the fact that the random behavior observed in simulation experiments is under the analyst's control and can be used to increase the information gained in such experiments. The theory behind these two variance reduction techniques is discussed below.

Antithetic Variates

The antithetic variates technique reduces the variance of estimated parameter values by creating negative correlation between observations in paired replications of a single system. If M_{1r} and M_{2r} are the estimated mean of the same parameter variable for Runs 1 and 2 in rth replication, respectively, the

mean value of the parameter can then be estimated by

$$\overline{M}_r = \frac{(M_{1r} + M_{2r})}{2} \tag{7}$$

Note that M_{1r} and M_{2r} are themselves estimates. The estimated variance of M_r is given by

$$V(M_{1r} + M_{2r}) = V(M_{1r}) + V(M_{2r}) + 2p\sqrt{V(M_{1r})} \cdot \sqrt{V(M_{2r})} \tag{8}$$

$$V(\overline{M}_r) = \frac{V(M_{1r} + M_{2r})}{4} \tag{9}$$

where p (ranging from -1 to $+1$) is the correlation coefficient for random variables M_{1r} and M_{2r} and V_{1r} and V_{2r} are the estimated variances of M_{1r}, and M_{2r} respectively. If M_{1r} and M_{2r} are independent, the correlation p would be zero. This implies that the variance of $\overline{M}_r$ will decrease if M_{1r} and M_{2r} are negatively correlated. A negative correlation implies that if the value of M_{1r} is above its average, then M_{2r} is more likely to be below the average. A negative correlation between the observations can be created by generating observations from the random number r in one run and $(1-r)$ in the next run. See Kleijnen (*9*, pp. 183–200) for proof that the use of $(1-r)$ creates a negative correlation.

Common Random Numbers

Common random numbers represent a very intuitive yet simple-to-implement variance reduction procedure in simulation experiments. The ability to reduce variance in this manner is highly effective and practically cost-free (*18*). The only requirement to apply the procedure is the matching of the random number seeds in simulation experiments.

The variance of the difference between M_{1r}, the estimated sample mean of the response (output) variable for base case scenario, and M_{2r}, the estimated sample mean of the same response variable for the alternative system design or operating policy in the rth replication, respectively, is given by

$$V(M_{2r} - M_{1r}) = V(M_{2r}) + V(M_{1r}) - 2p\sqrt{V(M_{2r})} \cdot \sqrt{V(M_{1r})} \tag{10}$$

where p (ranging from -1 to $+1$) is the correlation coefficient for random variables M_{1r} and M_{2r}. Equation 10 suggests that the variance of the estimated difference in system performance is decreased if the response variables are positively correlated. A positive correlation implies that if the value of M_{2r} is above its average, then M_{1r} is more likely to be above the average. If the simulated environment reacts to the stochastic input variables in the same direction for both the base case and an alternative scenario, such a positive correlation can be created by using common random numbers. That is, if random number seeds $(r_1, r_2, \ldots r_n)$ are used for the base case, a positive correlation can be created by using the same random number seeds for the alternative scenario.

TEST OF THE EFFICIENCY OF THE VARIANCE REDUCTION CONCEPTS

The effectiveness of the variance reduction techniques described in the previous section (i.e., antithetic variates and common random numbers) is evaluated through its application for the sample data set distributed with the TRAF-NETSIM program. Figure 1 shows the link-node representation of the sample test data set. Although a small network, this data set includes bus operations, incidents, parking, actuated and fixed-time signal controls, time-varying input, and a variety of other features of traffic operations on urban surface street networks. The purpose of this data set is to demonstrate the features and capabilities of the model.

A base case scenario is developed by utilizing the signal control scheme at Node 1 shown in Figure 2a. An alternative scenario is created by including a leading green for eastbound traffic, as shown in Figure 2b. The control scheme shown in Figure 2b is the one used in the sample data set that is distributed with the TRAF-NETSIM program.

Using 30 independent random number seeds provided in Tables A.1 through A.3 of the classic textbook on discrete event simulation by Fishman (*18*, p. 486), 90 simulation runs were made for the base case and the alternative scenario (45 runs for each case). First, 30 independent replications for traffic operations under the base case scenario of the sample network were made using these seeds. Then, for the first 15 simulation runs, the antithetic runs were made using random deviates $(100 - d)$ where d is the random deviate generated in the original run of the pair. Therefore, a total of 45 simulation runs were made for the base case. The same replications were then made for the alternative scenario. Including an initialization time of 300 sec, the length of each simulation run was 900 sec.

Let us assume that the purpose of these simulations is to compare the two control strategies for the test network for some selected measures of performance. That is, the objective of our experiment is to determine the effect of the change in control strategy. The TRAF-NETSIM model generates a considerable amount of output data including estimates of speeds, delay, stops, travel time, fuel consumption, and emission on each link of the network (by turn movement if specified by the user), a group of selected links, and the entire network over user-specified time intervals. For illustration purposes, four of these measures of effectiveness (MOEs) were selected for analysis: average speed (in mph) of the left-turning vehicles on Link (2,1), number of vehicle trips on Link (2,3), number of signal phase failures at Node 1, and average travel time (in seconds) in the network. In selecting these output variables, an attempt has been made to include some of the commonly used performance measures across different levels of network aggregation. Tables 2 and 3 show the simulation results, along with the random number seeds employed in the simulation runs.

Using the data shown in Tables 2 and 3, the three estimates of variance for MOEs are obtained. First, the variance based on common random numbers is obtained by using the 30 paired replications of the base case and alternative scenario. Let A_r and B_r be the estimated value of the MOE in the rth replication of alternative scenario and base case, respectively.

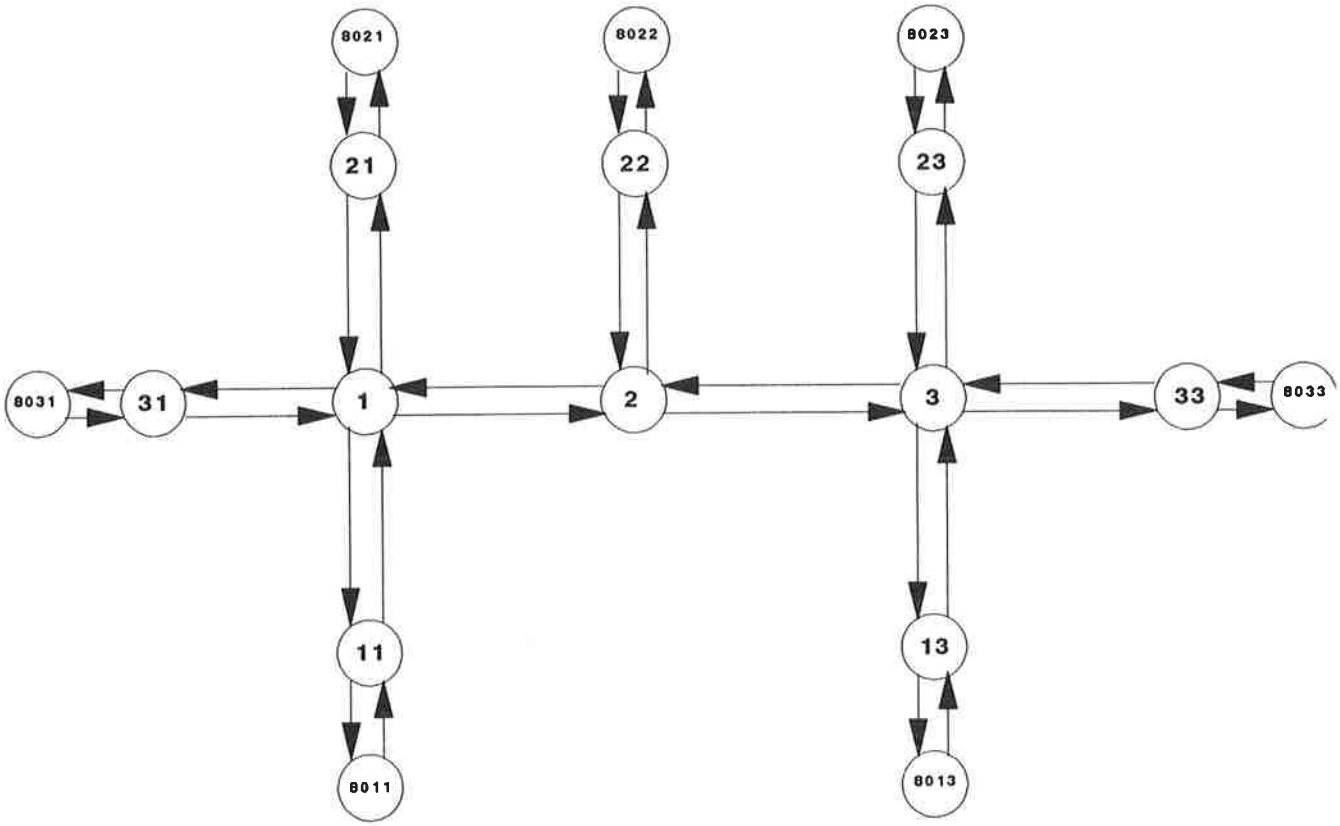


FIGURE 1 Link-node representation of TRAF-NETSIM sample data set.

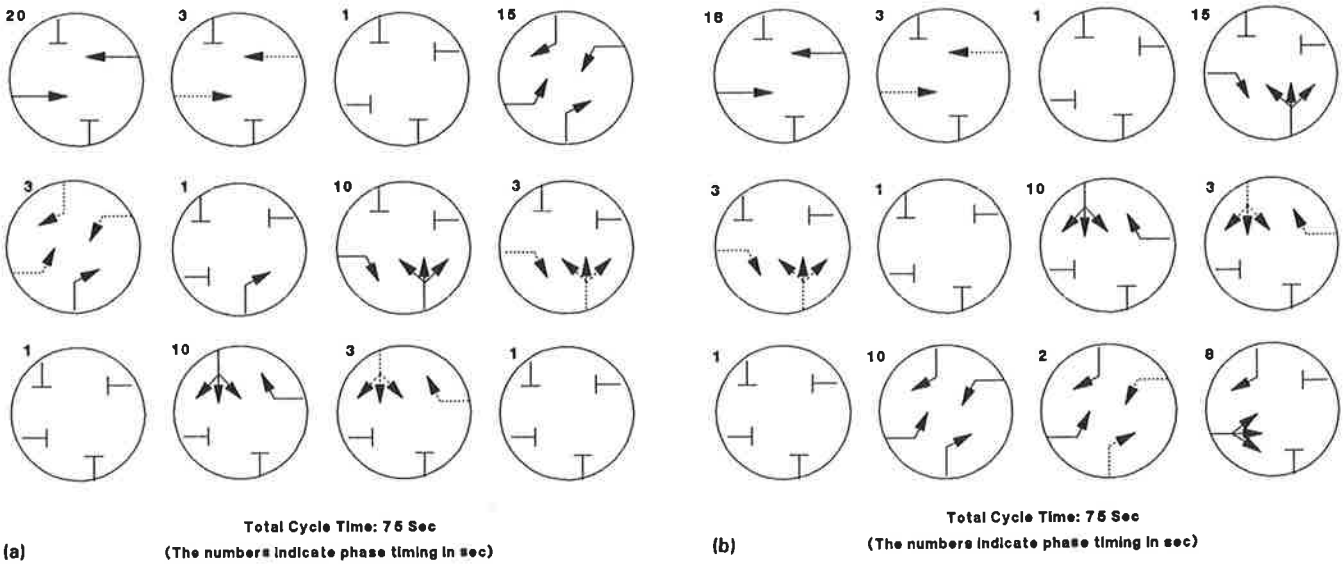


FIGURE 2 Signal timing at Node 1: (*a*) base case, (*b*) alternative case.

TABLE 2 Simulation Results from 45 Replications of Base Case

Run	Random Number Seed	Speed (mph) of Left Turning Vehicles on Link (2,1)		Number of Vehicle Trips on Link (2,3)	
		Run		Run	
		Base	Antithetic	Base	Antithetic
1	7781	8.5	7.4	212	247
2	75253171	7.0	7.3	232	239
3	76271663	7.1	7.3	237	229
4	68911991	8.2	8.0	230	237
5	67784357	8.0	6.6	256	245
6	9072619	5.5	4.2	245	233
7	99791377	6.6	8.7	230	231
8	81120351	7.2	7.2	225	236
9	90563473	8.8	9.1	234	227
10	24918189	7.9	8.6	230	220
11	20464843	8.8	6.5	232	230
12	75809031	8.4	6.8	225	232
13	51703891	8.0	8.7	229	258
14	26784697	7.1	6.6	256	233
15	62954703	8.7	8.0	238	221
16	41758441	5.3		244	
17	29977537	8.6		222	
18	9409871	6.8		224	
19	38090947	7.6		237	
20	59133357	6.8		242	
21	55161629	6.9		246	
22	37391779	7.8		216	
23	97264441	5.7		226	
24	73907653	7.7		237	
25	55730273	7.3		230	
26	74792037	8.8		237	
27	97859453	7.3		232	
28	53479	7.9		242	
29	12217	8.0		239	
30	40577	5.9		230	

Run	Random Number Seed	Average Travel Time (in seconds/vehicle trip)		Number of Phase failures at Node 1	
		Run		Run	
		Base	Antithetic	Base	Antithetic
1	7781	82.2	84.0	21	14
2	75253171	89.4	84.0	21	22
3	76271663	87.0	85.8	18	20
4	68911991	82.8	86.4	18	22
5	67784357	86.4	84.0	20	21
6	9072619	92.4	92.4	27	22
7	99791377	82.8	83.4	17	18
8	81120351	83.4	87.0	20	25
9	90563473	90.6	81.6	18	23
10	24918189	86.4	79.2	21	19
11	20464843	83.4	87.0	21	23
12	75809031	91.2	88.8	23	24
13	51703891	84.0	85.8	16	19
14	26784697	82.2	87.0	12	25
15	62954703	79.8	87.0	11	17
16	41758441	89.4		26	
17	29977537	85.2		22	
18	9409871	87.6		18	
19	38090947	88.8		22	
20	59133357	85.8		18	
21	55161629	85.2		16	
22	37391779	88.8		20	
23	97264441	83.4		17	
24	73907653	87.6		23	
25	55730273	81.6		19	
26	74792037	86.4		20	
27	97859453	84.6		21	
28	53479	87.6		23	
29	12217	85.8		16	
30	40577	88.2		25	

TABLE 3 Simulation Results from 45 Replications of the Alternative Scenario

Run	Random Number Seed	Speed (mph) of Left Turning Vehicles on Link (2,1)		Number of Vehicle Trips on Link (2,3)	
		Run		Run	
		Base	Antithetic	Base	Antithetic
1	7781	6.5	6.0	211	242
2	75253171	5.4	5.6	233	237
3	76271663	5.7	4.1	232	231
4	68911991	7.1	5.6	234	232
5	67784357	4.8	5.1	254	244
6	9072619	2.4	2.2	239	231
7	99791377	5.1	5.9	231	224
8	81120351	5.6	5.7	227	239
9	90563473	6.6	5.3	238	229
10	24918189	5.6	6.0	234	222
11	20464843	7.5	2.6	224	231
12	75809031	7.5	5.8	214	238
13	51703891	5.6	5.7	225	255
14	26784697	5.3	5.0	256	227
15	62954703	6.7	4.9	235	220
16	41758441	4.3		239	
17	29977537	7.8		226	
18	9409871	5.2		221	
19	38090947	6.2		239	
20	59133357	5.6		240	
21	55161629	5.3		245	
22	37391779	7.0		214	
23	97264441	3.8		227	
24	73907653	6.8		234	
25	55730273	4.8		227	
26	74792037	8.0		240	
27	97859453	4.8		226	
28	53479	4.7		244	
29	12217	6.5		242	
30	40577	4.4		231	

If D_r equals the difference in MOE for replication r (i.e., $D_r = A_r - B_r$), then the estimated variance of a pair of replications is given by

$$V(D_r) = \frac{1}{(n-1)} \sum_{r=1}^{n} (D_r - \overline{D})^2 \tag{11}$$

where

$$\overline{D} = \frac{1}{n} \sum_{r=1}^{n} D_i \tag{12}$$

and n is the number of replications.

Then, the variance of base case and alternative scenario are estimated individually by using the 15 paired replications in each case. If M_{1r} and M_{2r} are the estimated values of an output variable in the base run and its antithetic pair in replication r, respectively, then the mean value of the parameter and its variance can also be estimated by the equations

$$M_r = \frac{(M_{1r} + M_{2r})}{2} \tag{13}$$

$$\overline{M} = \frac{1}{n} \sum_{r=1}^{n} M_r \tag{14}$$

and

$$V(M_r) = \frac{1}{(n-1)} \sum_{r=1}^{n} (M_r - \overline{M})^2 \tag{15}$$

where M_r is the estimated value of the output variable from pair r. $\overline{M}$ is the estimated mean value of the output variable, n is the number of replications, and $V(M_r)$ is the variance of M_r. Again, note that M_r is itself an estimate.

Finally, the variance without the use of variance reduction techniques (i.e., paired replications) is estimated for the base case and alternative scenario individually on the basis of 30 independent replications. Using standard statistical formulas,

$$\overline{M} = \frac{1}{n} \sum_{r=1}^{n} M_r \tag{16}$$

and

$$V(M_r) = \frac{1}{(n-1)} \sum_{r=1}^{n} (M_r - \overline{M})^2 \tag{17}$$

Table 4 gives the computed variances for each of the four output variables and the reduction in variance derived by using antithetic variates and common random numbers after 10, 20, and 30 replications. An examination of simulation results in Table 4 leads to the following observations:

- The two variance reduction techniques indeed significantly reduce the variance of the MOEs. With the exception of three situations, at least 30 percent variance reduction is obtained in all cases. The average reduction in variance of the parameter estimate is nearly 65 percent for the 24 comparisons that are made. For individual MOEs, variance reduction in the range of 40 to 85 percent is obtained. Variance

TABLE 4 Computed Variances of the Performance Measures for the Test Data Set

Sampling Procedure	Average Travel Time	Number of Phase Failures at Node 1	Speed of Left Turning Vehicles on Link (2,1)	Number of Vehicle Trips on Link (2,3)
Independent Replications				
Variance: After 10 Replications	35.20	39.03	2.67	250.11
After 20 Replications	30.80	36.63	2.64	244.04
After 30 Replications	24.61	30.49	2.61	213.59
Antithetic Paired Replications				
Variance: After 10 Replications (5 Pairs)	14.53	26.90	0.43	45.38
After 20 Replications (10 Pairs)	21.98	24.11	1.31	37.34
After 30 Replications (15 Pairs)	20.65	28.83	1.11	55.62
Common Random Number Replications				
Variance: After 10 Replications (5 Pairs)	0.18	16.00	0.66	11.33
After 20 Replications (10 Pairs)	11.12	17.56	0.50	13.29
After 30 Replications (15 Pairs)	8.57	14.40	0.44	20.82
Variance Reduction (%)				
Antithetic Pairs Vs. Independent Replications Variance				
After 10 Replications	59	31	84	82
After 20 Replications	29	34	50	85
After 30 Replications	16	5	57	74
Common Random Numbers vs. Variance Assuming Independence:				
After 10 Replications	99	59	75	95
After 20 Replications	64	52	81	95
After 30 Replications	65	52	83	90

reduction of over 75 percent is realized in 10 of the 24 comparisons.

- With the exception of one comparison, the common random numbers–based variance reduction is more effective than the antithetic variates procedure in this case study. The result is not surprising considering the identical traffic stream feature of the TRAF-NETSIM model. In general, the common random number variance after 10 replications is less than the variance after 20 or 30 independent replications or the antithetic paired replications.
- The common random numbers strategy reduced variance by at least 50 percent in all comparisons. One out of two times, the reduction in variance was greater than 80 percent.
- The use of common random numbers produces consistent differences in the MOEs between the base case and alternative scenario (See Tables 2 and 3).
- In all cases, the variance after 10 antithetic replications is considerably less than the variance after 20 or 30 independent replications. Therefore, after making as few as one-third as many simulation runs, better statistical precision is obtained.
- In general, the effectiveness of variance reduction is decreased as the number of replications is increased. The only exception is the output statistics for the speed of left-turning vehicles on Link (2,1). This parameter is the least sampled among all performance measures used in this study. In this case, for example, fewer than 100 observations are made for the speed of left-turning vehicles compared with more than 2,000 for the average travel time in the network. As the number of observations is increased (through longer simulations) for this parameter, the results could be different. Nonetheless, this anomaly reinforces the requirement for large sample sizes in parameter estimation. As far as antithetic variates are concerned, variance is reduced regardless.
- The variance reduction techniques are as effective for the output measures that are cumulative statistics (number of phase failures) as they are for the measures that are themselves estimates (i.e., average values such as average travel time).

Finally, a note about the TRAF-NETSIM output itself. There is very little variability in the average travel time values; the mean to standard deviation ratio, known also as the *t*-statistic, is nearly 20. On the other hand, the same ratio for the number of phase failures is approximately 2. In general, though, the variability exhibited by the MOEs between the runs is small because of the congestion in simulated environment and because of larger sample sizes resulting from a 15-min simulation. In all cases, however, the variance is reduced by the use of the common random numbers and antithetic variates procedure.

SUMMARY

This study illustrates the effectiveness of two variance reduction techniques, antithetic variates and common random numbers, in simulation experiments with the TRAF-NETSIM simulation model. The primary motivation for the analyst to use these techniques is to reduce the risk of drawing incorrect conclusions and to increase the precision of the estimates without significant additional effort. The results clearly show that both techniques reduce the variance of the output performance measures and illustrate their utility in simulation

experiments. These variance reduction techniques can easily be applied by using the identical traffic stream feature of the model.

ACKNOWLEDGMENT

The authors are grateful to Henry Lieu and Alberto Santiago, both of the IVHS Division of FHWA, and Amar Kanaan of AEPCO, Inc., for providing the TRAF-NETSIM source code. Many helpful comments from David Metzger of the University of Tennessee are appreciated.

REFERENCES

1. M. Y. Yedlin, E. B. Lieberman, B. Andrews, A. K. Rathi, and J. F. Torres. *TRAF User Guide*. Federal Highway Administration, U.S. Department of Transportation, 1988.
2. A. K. Rathi and A. J. Santiago. Identical Traffic Streams in the TRAF-NETSIM Simulation Program. Presented at the Transportation Research Board's 69th Annual Meeting, Washington, D.C., 1990.
3. Shui-Ying Wong. TRAF-NETSIM: How it Works and What it Does. *ITE Journal*, April 1990, pp. 22–27.
4. A. Halati, J. F. Torres, and S. L. Cohen. FRESIM—Freeway Simulation Model. Presented at Transportation Research Board's 70th Annual Meeting, Washington, D.C., 1991.
5. A. S. Byrne, A. B. deLaski, K. G. Courage, and C. E. Wallace. *Handbook of Computer Models for Traffic Operations Analysis*. FHWA-TS-82-213. Federal Highway Administration, U.S. Department of Transportation, 1982.
6. E. Lieberman. Traffic Simulation: Past, Present, and Future. *Proc., International Symposium on Traffic Control Systems*. UGB-ITS-P-79-2. Institute of Transportation Studies, Berkeley, Calif., 1979.
7. A. May. Freeway Simulation Models Revisited. In *Transportation Research Record 1132*, TRB, National Research Council, Washington, D.C., 1989, pp. 94–99.
8. P. Michalopoulos, E. Kwon, and J.-G. Kang. Enhancement and Field Testing of a Dynamic Freeway Simulation Program. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991.
9. J. P. C. Kleijnen. *Statistical Techniques in Simulation: Part I*. Marcel Dekker, 1974.
10. K. D. Tocher. *The Art of Simulation*. Van Nostrand, 1963.
11. A. K. Rathi and Z. A. Nemeth. An Application of Variance Reduction Techniques in Freeway Simulation. *Transportation Research*, Vol. 19B, No. 3, 1985, pp. 209–215.
12. J. F. Torres, A. Halati, and A. Gafarian. *Statistical Guidelines for Simulation Experiments: Vols. 1, 2 and 3*. Contract DTFH-61-80-C-00124. Federal Highway Administration, U.S. Department of Transportation, 1983.

TABLE 5 TRAF-NETSIM Simulated Stopped Delays[a]

SIMULATION TIME (SECOND)	SIMULATED STOPPED DELAY (SEC/VEH) IDENTICAL TRAFFIC STREAM RANDOM NUMBER SEED 7781	45451891	2135011	97116143	95051027	53493673	95612789	70257223	67121881	66687679	74071517	MEAN	STANDARD DEVIATION	REQUIRED SAMPLE SIZE[b] 95% CONFIDENCE LEVEL WITH TOLERABLE ERROR 10% OF MEAN	15% OF MEAN	90% CONFIDENCE LEVEL WITH TOLERABLE ERROR 10% OF MEAN	15% OF MEAN
LINK 1-2																	
600	16.1	17.7	16.0	16.8	17.2	20.1	13.4	21.8	19.4	20.7	17.8	17.9	2.429	9.1	4.1	6.0	2.7
1200	13.9	21.8	14.4	19.3	17.2	16.2	16.5	25.9	19.4	16.8	22.3	18.5	3.652	19.3	8.6	12.8	5.7
1800	14.7	19.6	14.9	17.2	17.9	16.1	15.4	24.4	19.2	17.8	18.8	17.8	2.771	12.0	5.3	7.9	3.5
2400	14.8	21.7	17.2	17.0	17.1	18.1	16.2	21.6	17.1	19.9	19.6	18.2	2.216	7.3	3.3	4.9	2.2
3000	15.3	20.7	18.4	16.7	17.0	18.4	17.7	23.7	17.3	19.8	21.1	18.7	2.399	8.1	3.6	5.4	2.4
3600	15.6	20.5	19.6	16.1	17.8	17.5	17.9	22.6	16.5	18.6	20.0	18.4	2.103	6.5	2.9	4.3	1.9
4200	16.4	21.1	19.1	16.0	18.1	18.0	17.6	21.7	16.7	19.0	21.2	18.6	1.996	5.7	2.5	3.8	1.7
4800	17.3	21.1	18.6	15.9	17.8	17.4	17.7	20.4	16.7	18.6	21.1	18.4	1.757	4.5	2.0	3.0	1.3
5400	17.3	21.1	19.2	16.3	17.2	17.4	18.4	19.5	16.9	18.8	20.8	18.4	1.592	3.7	1.6	2.4	1.1
6000	17.4	20.6	19.6	17.2	16.7	17.2	18.6	19.4	16.5	19.3	20.6	18.5	1.527	3.4	1.5	2.2	1.0
6600	16.9	20.0	19.1	17.4	17.1	17.4	17.8	19.2	16.2	19.5	19.7	18.2	1.316	2.6	1.2	1.7	0.8
7200	17.0	19.6	19.0	17.8	17.1	17.7	18.0	18.8	17.0	19.0	19.4	18.2	0.979	1.4	0.6	0.9	0.4
7800	17.3	19.8	19.1	18.1	17.2	17.6	17.9	18.7	16.9	19.3	19.3	18.3	0.995	1.5	0.7	1.0	0.4
8400	17.2	19.6	18.8	17.8	17.0	17.8	17.8	18.4	16.7	19.1	19.1	18.1	0.953	1.4	0.6	0.9	0.4
9000	17.1	19.9	19.0	17.8	16.8	18.6	17.6	18.2	16.7	19.3	18.6	18.1	1.047	1.7	0.7	1.1	0.5
9600	17.5	19.7	18.9	18.2	16.7	18.8	17.5	18.0	16.8	19.0	18.4	18.1	0.947	1.4	0.6	0.9	0.4
9950	17.6	19.9	18.9	18.4	16.8	19.1	17.4	18.0	17.0	19.0	18.1	18.2	0.965	1.4	0.6	0.9	0.4
LINK 4-3																	
600	37.4	21.8	20.4	-	26.0	20.6	24.3	30.6	14.0	21.0	51.5	26.8	10.226	74.7	33.2	49.1	21.8
1200	39.6	17.4	26.9	22.1	20.2	20.0	20.9	22.7	12.0	19.9	48.8	24.6	10.548	91.3	40.6	60.4	26.9
1800	34.0	17.9	25.5	24.3	23.4	22.1	21.3	21.3	13.5	20.4	40.4	24.0	7.418	47.4	21.1	31.3	13.9
2400	34.7	16.9	24.9	23.3	23.2	21.2	21.5	21.3	14.8	19.9	34.3	23.3	6.245	35.7	15.9	23.6	10.5
3000	33.8	17.6	23.3	23.6	22.7	23.0	23.0	20.2	15.9	18.8	30.3	22.9	5.248	26.0	11.6	17.2	7.6
3600	31.9	17.7	23.4	23.3	22.6	22.1	24.0	20.0	17.9	19.9	35.2	23.5	5.484	27.1	12.1	17.9	8.0
4200	30.3	19.0	23.9	24.4	22.3	21.4	23.1	21.5	18.8	19.6	34.4	23.5	4.833	21.0	9.3	13.9	6.2
4800	29.4	18.5	23.3	23.2	21.9	20.1	22.8	21.5	20.4	19.5	32.1	23.0	4.184	16.5	7.3	10.9	4.8
5400	29.3	19.4	23.8	23.2	22.4	21.7	23.4	21.7	20.7	19.7	30.7	23.3	3.632	12.1	5.4	8.0	3.6
6000	28.4	19.7	24.5	27.5	22.6	22.7	22.9	22.5	21.1	19.7	30.9	23.9	3.630	11.5	5.1	7.6	3.4
6600	27.6	19.7	23.8	26.7	22.2	22.4	22.9	23.6	21.4	19.1	31.4	23.7	3.613	11.5	5.1	7.6	3.4
7200	26.8	19.5	23.7	26.6	22.3	22.6	22.6	23.2	21.0	19.2	30.5	23.5	3.368	10.2	4.5	6.8	3.0
7800	26.4	22.2	23.4	26.1	22.1	23.6	22.3	24.2	20.9	19.3	29.9	23.7	2.929	7.6	3.4	5.0	2.2
8400	26.4	23.2	22.8	26.9	21.6	23.1	22.1	24.1	20.9	19.4	29.0	23.6	2.835	7.2	3.2	4.7	2.1
9000	25.7	23.3	22.5	28.7	21.7	22.6	24.1	23.7	21.3	19.7	28.4	23.8	2.824	7.0	3.1	4.6	2.1
9600	25.0	23.5	22.4	28.0	22.3	22.5	24.0	23.4	21.0	19.9	27.8	23.6	2.529	5.7	2.5	3.8	1.7
9950	24.7	23.3	22.2	27.6	22.9	22.7	23.7	23.0	20.8	19.9	27.3	23.5	2.365	5.0	2.2	3.3	1.5

[a]Adapted from (1, p. 481).

[b]From equation 1.

13. A. Halati. *Estimation of the Ratio of the Steady-State Means of a Bivariate Stochastic Process*. Ph.D. thesis. University of Southern California, Los Angeles, 1985.
14. A. V. Gafarian and A. Halati. Statistical Analysis of Output Ratios in Traffic Simulation. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1989, pp. 29–36.
15. G. L. Chang and A. Kanaan. Variability Assessment for TRAF-NETSIM. *ASCE Journal of Transportation Engineering*, Vol. 116, No. 5, 1990, pp. 636–657.
16. A. K. Rathi and A. J. Santiago. Urban Network Traffic Simulation: TRAF-NETSIM Program. *ASCE Journal of Transportation Engineering*, Vol. 116, No. 6, 1990, pp. 734–743.
17. A. K. Rathi. The Use of Common Random Numbers To Reduce the Variance in Network Simulation of Traffic. Forthcoming Technical Note, *Transportation Research*, 1991.
18. G. S. Fishman. *Principles of Discrete Event Simulation*. Wiley Interscience, New York, 1978.

DISCUSSION

SHUI-YING WONG
Federal Highway Administration, Office of Traffic Operations and IVHS, HTV-32, 400 7th St., S.W., Washington, D.C. 20590

The author provided a good discussion of the variation of the outputs from TRAF-NETSIM. In illustrating the variance reduction techniques, the author only considered TRAF-NETSIM runs with a simulation time of 900 sec. In a case study applying TRAF-NETSIM to estimate capacity and level of service (*1*), we found that the TRAF-NETSIM output is sensitive to simulation time. In the study, we applied 11 random number seeds to the same TRAF-NETSIM run. We found four interesting phenomena: (*a*) The output may have great variation when the simulation time is short (short being 1,800 sec or less); the variation becomes less as simulation time becomes longer. (*b*) If we trace the output value from one random number seed versus simulation time, the output value may fluctuate when the simulation time is short, and the output value stabilizes as simulation time becomes longer. (*c*) Different performance measures have different degrees of variation. (*d*) The same performance measure may have different degrees of variation under different situations.

Tables 5 and 6 show these phenomena. Figures 3 and 4 are based on the data from Tables 5 and 6, respectively. For example, the variation of stopped delay on Link 4-3 (Figure 3b) ranged from 12 (random number seed 7025223) to 49 (random number seed 66687679) sec per vehicle when the simulation time was 1,200 sec. The variation ranged from 20 (random number seed 74071517) to 28 sec per vehicle (random number seed 66687679) when the simulation time

TABLE 6 TRAF-NETSIM Simulated Capacities[a]

	SIMULATED CAPACITY (VEHICLES/HOUR)													REQUIRED SAMPLE SIZE[b]			
	IDENTICAL TRAFFIC STREAM RANDOM NUMBER SEED													95% CONFIDENCE LEVEL WITH TOLERABLE ERROR		90% CONFIDENCE LEVEL WITH TOLERABLE ERROR	
SIMULATION TIME (SECOND)	7781	2135011	45451891	95051027	97116143	95612789	53493673	67121881	70257223	74071517	66687679		MEAN	STANDARD DEVIATION	10% OF MEAN	15% OF MEAN	10% OF MEAN	15% OF MEAN
LINK 1-2																	
600	906	720	690	702	876	714	870	738	798	900	750		787.6	84.899	5.8	2.6	3.8	1.7
1200	882	807	771	747	801	795	822	786	822	789	723		795.0	41.699	1.4	0.6	0.9	0.4
1800	842	814	772	790	842	788	856	754	858	716	744		797.8	48.485	1.8	0.8	1.2	0.5
2400	811	753	787	810	834	810	852	765	846	705	763		794.2	44.634	1.6	0.7	1.0	0.5
3000	816	782	795	792	856	814	846	740	823	728	771		796.6	40.183	1.3	0.6	0.8	0.4
3600	802	802	813	801	853	780	837	766	804	715	758		793.7	38.058	1.1	0.5	0.8	0.3
4200	798	791	793	798	828	767	824	766	804	720	775		787.6	30.170	0.7	0.3	0.5	0.2
4800	780	786	779	786	795	770	823	776	809	731	778		783.0	23.259	0.4	0.2	0.3	0.1
5400	786	780	782	790	802	763	812	788	817	719	778		783.4	26.402	0.6	0.3	0.4	0.2
6000	784	777	783	790	795	770	810	769	823	716	784		781.9	27.205	0.6	0.3	0.4	0.2
6600	784	775	792	790	797	766	806	771	817	715	788		781.9	26.730	0.6	0.3	0.4	0.2
7200	779	780	793	785	798	767	799	773	813	729	783		781.7	21.827	0.4	0.2	0.3	0.1
7800	768	770	786	782	795	768	798	779	809	727	771		777.5	21.547	0.4	0.2	0.3	0.1
8400	771	774	778	779	798	765	798	787	803	736	768		777.9	18.987	0.3	0.1	0.2	0.1
9000	761	769	785	774	796	760	802	793	808	744	778		779.1	19.816	0.3	0.1	0.2	0.1
9600	753	778	794	763	793	759	805	795	808	741	780		779.0	22.298	0.4	0.2	0.3	0.1
9950	749	777	791	762	792	761	804	795	808	739	777		777.7	22.668	0.4	0.2	0.3	0.1
LINK 4-3																	
600	450	420	522	516	510	528	450	444	516	534	546		494.2	43.950	3.9	1.7	2.6	1.2
1200	471	483	498	525	495	516	483	468	498	438	549		493.1	30.128	1.9	0.8	1.2	0.5
1800	476	506	478	506	474	490	486	482	488	480	526		490.2	16.086	0.5	0.2	0.4	0.2
2400	483	508	481	505	459	480	484	484	472	495	486		485.2	13.862	0.4	0.2	0.3	0.1
3000	494	510	482	489	476	488	458	484	465	492	494		484.7	14.423	0.4	0.2	0.3	0.1
3600	486	510	483	497	487	493	456	494	478	488	488		487.3	13.320	0.4	0.2	0.2	0.1
4200	491	520	478	498	474	503	467	498	468	471	487		486.8	16.940	0.6	0.3	0.4	0.2
4800	493	510	478	492	474	501	462	483	471	462	480		482.4	15.397	0.5	0.2	0.3	0.1
5400	490	510	478	500	470	504	466	474	467	460	490		482.6	17.043	0.6	0.3	0.4	0.2
6000	491	505	482	511	462	501	478	476	478	466	495		485.9	15.928	0.5	0.2	0.4	0.2
6600	485	484	476	507	459	498	466	483	477	470	484		480.8	13.688	0.4	0.2	0.3	0.1
7200	480	484	473	506	465	504	467	484	473	479	486		481.9	13.315	0.4	0.2	0.3	0.1
7800	477	481	472	501	456	500	468	486	460	477	477		477.7	14.255	0.4	0.2	0.3	0.1
8400	478	480	475	500	456	495	464	486	465	477	470		476.9	13.232	0.4	0.2	0.3	0.1
9000	484	480	479	496	462	496	462	487	466	480	474		478.7	12.001	0.3	0.1	0.2	0.1
9600	484	483	480	492	457	493	460	479	469	475	476		477.1	11.562	0.3	0.1	0.2	0.1
9950	485	482	481	494	458	496	461	478	471	478	476		478.2	11.814	0.3	0.1	0.2	0.1

[a]Adapted from (1, p. 480).

[b]From equation 1.

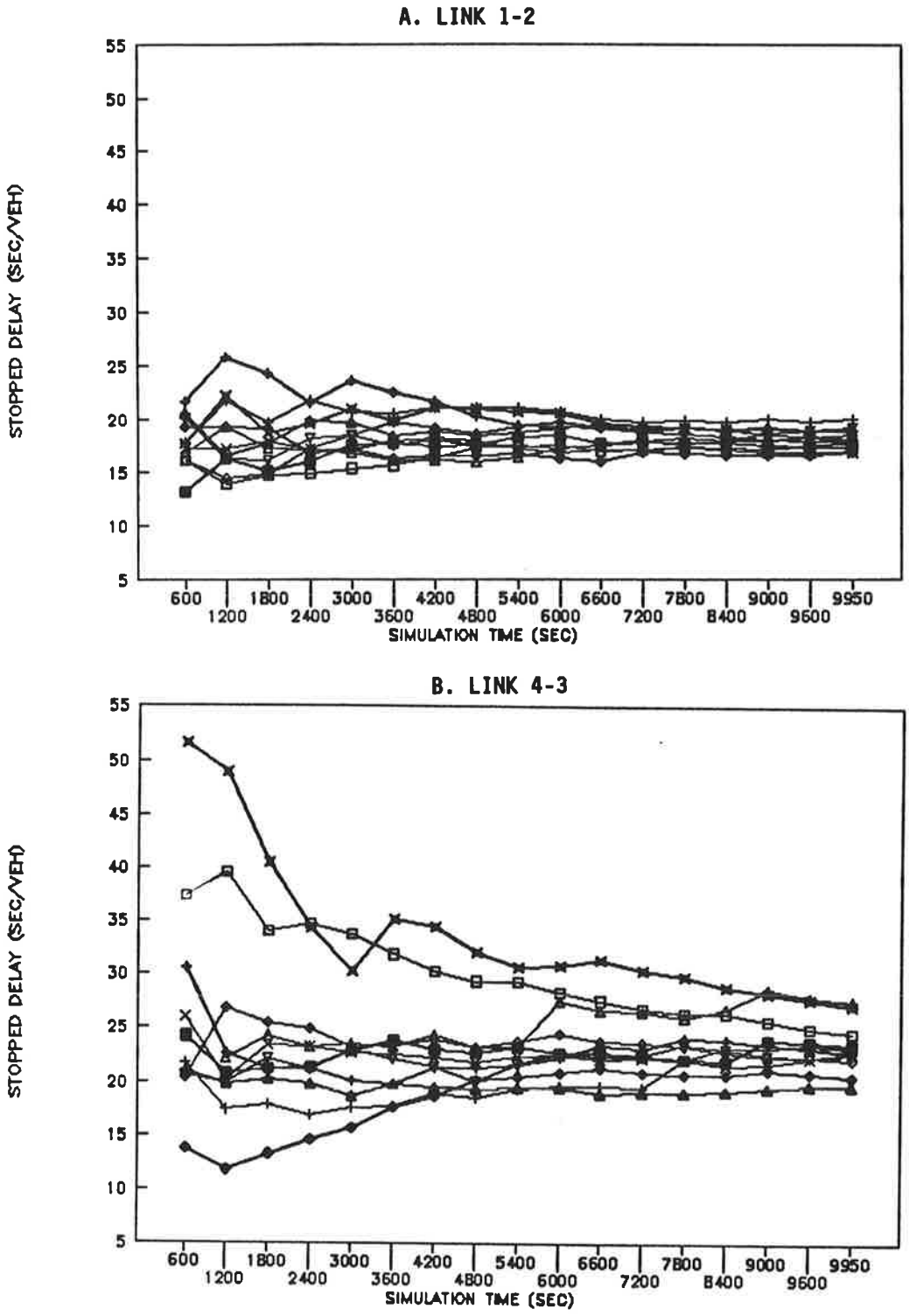


FIGURE 3 TRAF-NETSIM simulated stopped delays (based on Table 5).

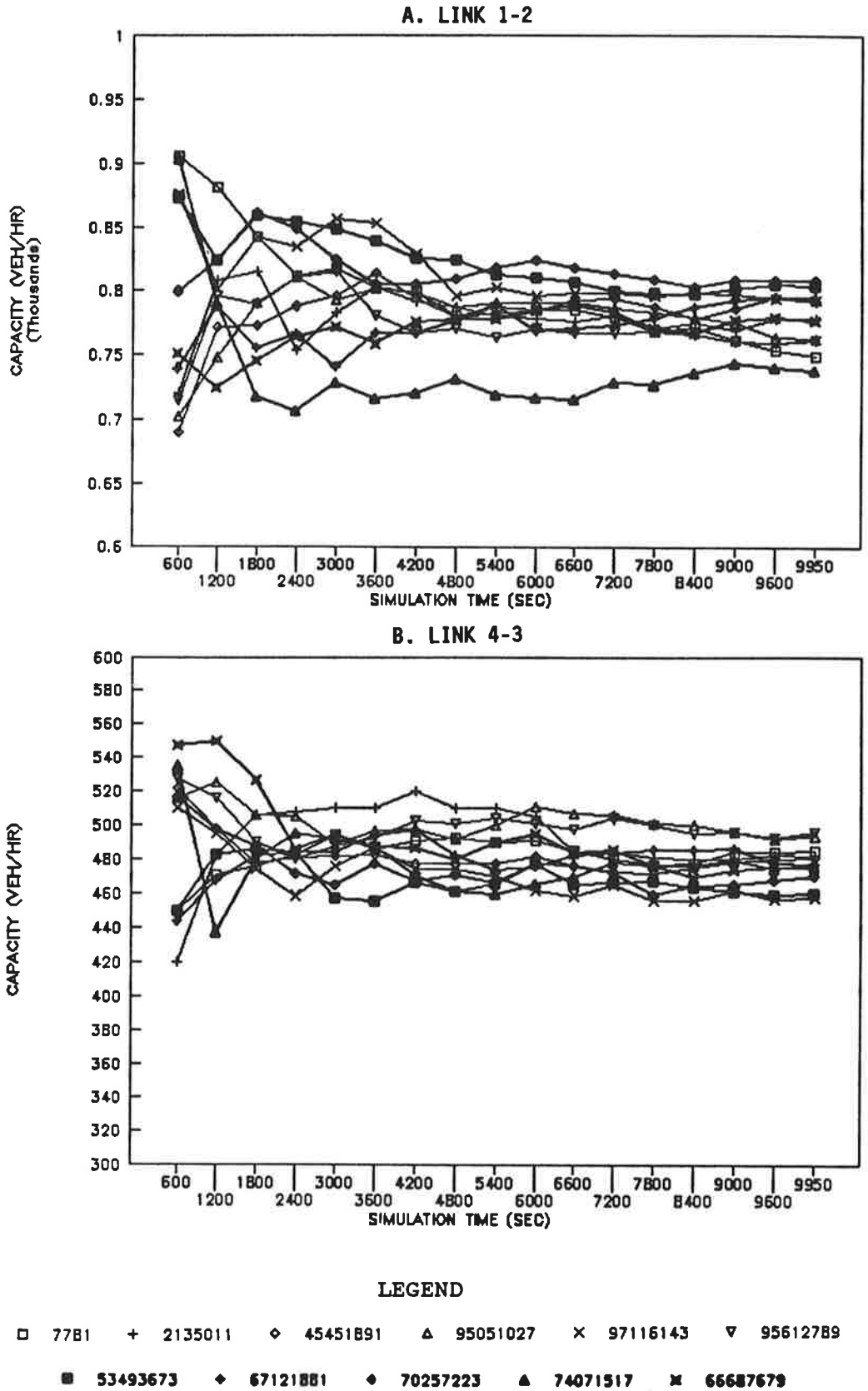


FIGURE 4 TRAF-NETSIM simulated capacities (based on Table 6).

was increased to 9,600 sec. If we trace the value of stopped delay for random number seed 66687679 (the top curve in Figure 3b), the value fluctuated from 52 to 30 sec per vehicle during the first 3,600 sec of simulation. The value stabilized around 28 to 30 sec per vehicle as simulation time became longer. This pattern was true for other random number seeds. The variation of capacity (Table 6 and Figure 4) was less than the variation of stopped delay, as evidenced from comparing the required sample size (number of runs) on the last four columns of Tables 5 and 6. The variation of stopped delay on Link 4-3 was greater than that on Link 1-2, because the situation was different. Left turns were permitted on Link 4-3, whereas they were not permitted on Link 1-2.

In the case study, stopped delay was from the "Stop Time" of the Cumulative NETSIM Cumulative Statistics report. Capacity was from the "Volume" of the Cumulative NETSIM Statistics report, with the link "flooded" or saturated with vehicles (*1*, p. 468). The network had 23 nodes and 33 links. A 9,600-sec simulation run took about 45 min to execute in an 80386, 16 megahertz microcomputer. All random number seeds, except the default value of 7781, were randomly generated. The required sample size was computed from

$$X = [(T_{1-a/2,\,n-1})\,(S)/E]^2 \quad (1)$$

where

X = required number of runs;
T = critical value for Student's t-distribution, with $(1-a)$ 100% level of confidence and $(n-1)$ degrees of freedom;
a = coefficient of confidence;
n = number of samples from which S is computed;
S = sample standard deviation; and
E = tolerable error.

REFERENCE

1. S. Y. Wong. Capacity and Level of Service by Simulation—A Case Study of TRAF-NETSIM. *Proc., International Symposium on Highway Capacity*, A. A. Balkema Publishers, Rotterdam, the Netherlands, 1991, pp. 467–483.

AUTHORS' CLOSURE

We thank Mr. Wong for taking the time to prepare a discussion of our paper. Mr. Wong's discussion essentially reinforces the concern about the variability of TRAF-NETSIM output and its implications for the valid usage of the model as a decision-support tool.

The first paragraph of Mr. Wong's discussion and the data from his study are essentially a reiteration of the theme of our paper. That is, TRAF-NETSIM simulation output contains much variability from one simulation run to the next, and the variability experienced in a given situation depends on several factors including the initialization time, simulation time, traffic volume, network geometry, the presence of incidents or disruptions in the simulated network, operational features of the simulated environment, and the performance measure itself. Mr. Wong's data suggest that after about 9,600 sec of simulation, the values of measures of effectiveness are stabilized.

Mr. Wong suggests that we have considered TRAF-NETSIM runs with only 900 sec of simulation. We are not sure if it is a concern or criticism, but our study is an attempt to address the problems associated with simulation runs of short durations. If one makes simulation runs of 9,600 sec or more, there would be no need (in most cases) to use the variance reduction techniques discussed in our paper. Since variance is inversely proportional to the sample size, the model can be either run longer or replicated many times (with short simulation durations) to obtain tighter confidence intervals on simulation output data. The variance reduction techniques discussed in this paper are presented as an alternative to simulation runs of long durations. Mr. Wong may find it interesting to use the antithetic variates techniques for his experiments and come up with the conclusions that we have reached on the basis of experiments discussed in this paper.

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

Network Programming To Derive Turning Movements from Link Flows

PETER T. MARTIN AND MARGARET C. BELL

It is generally accepted that there is a need to develop traffic models to manage and control congestion in real time. This control needs to be integrated with route guidance systems in order to achieve rerouting. A basic requirement of such models is to automatically establish turning movements from link flows. Conventional models such as entropy maximizing and information minimizing have been developed for use off line for transportation planning and are not suitable for application on line. A novel approach is proposed that uses linear programming to forecast traffic congestion in an urban network and define junction turning flows. The algorithms, originally developed for optimizing flows of water and electricity, use detector flow measurements, weighted links, and constrained upper and lower flow bounds. The principles underlying this approach are explained. The development, calibration, validation, and implementation of the model in a real network in the city of Leicester, England, are described. The results show that turning movements can be predicted with sufficient accuracy to justify further work, in particular to carry out a demonstration of the application of the algorithm on street.

Current coordinated traffic signal systems, both fixed time and demand responsive, are designed to control networks operating at 90 and 95 percent saturation. Some demand-responsive systems have facilities to gate, meter, and favor traffic to ease congestion. Currently, congestion management relies heavily on the engineer's judgment of the traffic situation viewed through closed-circuit television systems and can only be achieved with operator intervention. Often small adjustments to timings help to alleviate congestion locally. However, comprehensive management and control of congestion need strategic control because traffic has to be redistributed along those routes with sufficient spare capacity, otherwise the overload would simply be transferred to another part of the network. Such a control system would be based on a philosophy that optimizes space in time. It would need to monitor traffic movements to forecast the onset of congestion and select appropriate control strategies. The system would monitor performance on line, build up a knowledge base, and eventually learn from experience, in other words, an expert system.

Traffic detection is advanced and reliable (*1*,*2*), and the current research effort is investigating methods for data-base management (*3*) of large volumes of information on line. Therefore, the infrastructure to support the implementation of the expert system techniques will shortly be available.

P. T. Martin, Department of Civil Engineering, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom. Current affiliation: ARDFA, California Polytechnic State University, San Luis Obispo, Calif. 93407. M. C. Bell, Department of Civil Engineering, University of Nottingham, University Park, Nottingham NG7 2RD, United Kingdom.

Traffic routes vary for many different reasons: longer-term shifts follow changes in land use and implementation of traffic management schemes, and in the short term changes occur in response to the build-up of recurrent congestion and incidents. Also, drivers' choice of route and demand for travel vary daily and seasonally. A fourth generation of signal control will therefore have to continuously monitor the shifts in traffic routes that take place and the corresponding transient changes in traffic conditions. The linear programming method proposed here, once calibrated, could be used to predict these changes in traffic patterns as they happen, thus allowing remedial congestion control strategies to be defined on line.

On a large scale, traffic routes are described by origin and destination matrices. On a small scale, traffic routes are defined by the flow along links and the turning movement at the junctions. The corresponding signal control strategy, defined off line in a fixed-time system and on line in a demand-responsive system, matches or governs the demand for travel on each link. The vehicle detectors supply a measure of link flows and these, with the constraints of the signal control timings, enable the turning movements to be inferred. The linear programming technique is therefore system driven, which is quite different from conventional transportation modeling, which is behavioral.

Conventional models that derive origin-destinations from link flows have been developed for transportation planning rather than for signal control. Since they are applied off line, computer run time and storage are not critical. Maximum entropy methods (*4*) select the most likely origin and destination matrices to be consistent with the given set of link flows with the least bias. Information minimization models (*5*) are founded on the principle that for a junction, a set of turning movements exists that is most probable. However, a solution fails to produce converged solutions when the data are noisy, which is usually the case for traffic. Therefore, dynamic methods (*6*) have been developed to interpret the rhythmic nature of the traffic data, but success has been limited.

Predicted flows from these large-scale transportation models rarely agree with those measured on the street, but are adequate for planning. These models are both slow and exhaustive on computer time and in their present form are unlikely ever to have application for traffic monitoring and control on line.

Previous research (*7*) had demonstrated that linear programming had potential. The particular advantage of the method was in the speed with which solutions to fairly large and complex problems could be achieved. This early study was based on traffic data generated by Monte Carlo simulation

modeling of flows on the network. The main conclusion of this study was that before substantial advances could be made in solving the problem of deriving turning movements from detector flows, there was a need to research the algorithms applied to traffic data from a real network. This could only be done by having direct access to signal timing and detector flow data from a network and carrying out surveys to enable simultaneous measurement of junction turning movements. It was for this reason that funding from the Science and Engineering Research Council (SERC) was secured to set up surveys to provide unique data sets. The third annual survey was made in May 1991.

THE STUDY AREA

Leicester, in the East Midlands of England, has a population of 289,000 and is a city with a series of radials and ring roads. In 1988 the fixed-time signal control system was replaced by a demand-responsive traffic signal control system, Split Cycle Offset Optimization Technique, or SCOOT (*8*). A subarea of SCOOT, Region R, to the south of the city was chosen for this research.

Figure 1 shows the street network of six signal-controlled junctions. This particular SCOOT subarea was chosen because it has a spatial geometry that offers alternative routes to traffic both into and out of the city center. During the morning peak, congestion builds up along London Road and drivers are known to seek alternative routes along Regent Road. During the evening peak, traffic can leave the city either along London Road or Regent Road. Another reason for choosing Region R was the knowledge that substantial route changes were expected following traffic management alterations. A new link to the ring road was built in 1989, and the city center was pedestrianized on October 14, 1990. These network changes are substantial and likely to have created noticeable shifts in route patterns in Region R.

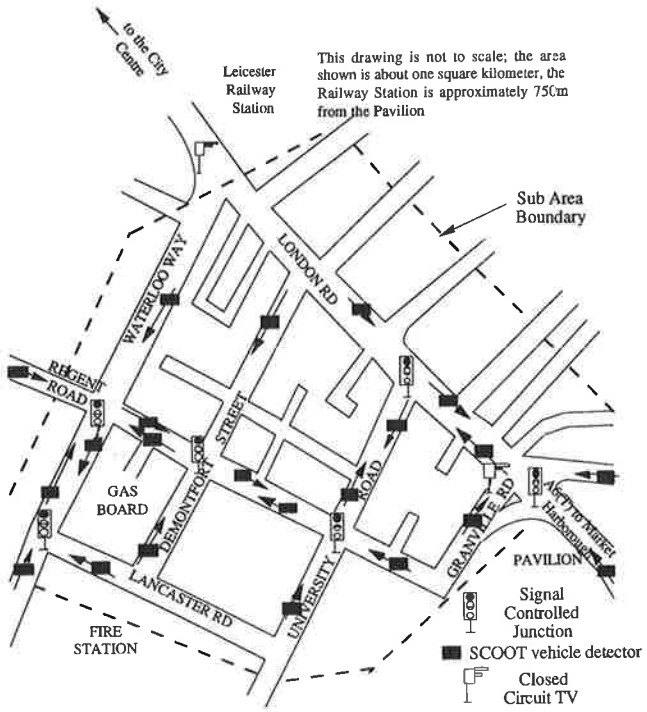


FIGURE 1 Street layout, SCOOT Region R, central Leicester.

The Nottingham University Transport Research Group (NUTRG) has a computer work station with a dedicated communication link with the Leicestershire County Council Traffic Management Computer. This link allows traffic flows, congestion indicators, and signal timing data to be retrieved at 5-min intervals continuously from detectors throughout Leicestershire. Detector flow data were monitored via SCOOT for 21 of the 170 links in the study area. Seventy on-street observers provided a comprehensive set of turning movements measured simultaneously with the gathering of the link traffic flows from the detectors. Three surveys, on Wednesday May 10, 1989, Wednesday May 9, 1990, and Wednesday May 8, 1991, between 1500 and 1800 hours provided comprehensive data. The linear programming method was used to infer the 54 unknown turning movements at 7 signal-controlled junctions from the 21 measured flows from SCOOT.

THE ALGORITHMS

A linear program model has three components. There must be a system, a problem, and a solution. The system is the network represented by a series of nodes connected by arcs. The geography of the town or city defines the structure of the network. Each arc accommodates a traffic demand or flow. The term "arc" is the network representation of a "link," which is used to describe the road. The signal timings quantify the control environment, that is, the capacities of both junctions and links. The problem is to predict a comprehensive set of turning flows from a sparse set of link flows measured by detectors. The solution is the feasible set of turning movements predicted with acceptable accuracy that satisfies the constraints.

When the system is constrained too severely, no feasible solution can be found. When the constraints are too lax, feasibility is too readily accomplished and the prediction is poor. The challenge of this research was to identify the constraints for the linear programming algorithms in such a way that all relevant information available from the control system is used so that the particular feasible solution, the correct one, is identified. The resulting solution then represents the turning flows reliably. It is important that the constraints conform to some network or traffic characteristics that are quantifiable. This enables the algorithm to be both repeatable and transferable. In this way the model can be applied to any control network for varying traffic demands at different times of day.

The algorithm relies on the Simplex method (*9*), which defines and solves a series of inequalities. If a set of solutions to the series of equations exists, such solutions are said to be feasible. The one solution that results from the minimization of a cost or a weighted function of attributes is known as the optimum solution. An algebraic procedure moves from one basic feasible solution to a better adjacent feasible solution by choosing a basic entering and leaving variable to solve a system of linear equations using Gaussian elimination. An iterative procedure is applied until the solution cannot be

further improved, which is deemed the optimal solution, and the computation of the algorithm stops. The program NETFLOW (*10*), written in FORTRAN, is the Simplex method in matrix form, which serves to streamline the original method considerably and makes it computationally faster with potential on-line application. At each iteration, sparse matrices hold the minimum of information in a more compact form, and therefore the procedure requires less computer storage. As a result much larger network problems can be solved. The modeling process is summarized in Figure 2.

NETFLO ALGORITHM

The network is modeled as a series of nodes connected by arcs. Arcs represent homogeneous stretches of road between junctions. Nodes represent the intersection of arcs, which are unidirectional. The NETFLO model applies a heuristic procedure to obtain the initial basic feasible solution. The main idea of this procedure is to quickly find the low-weight paths through the network that will accommodate the demand to and from each node in the network. Mathematically, the linear programming problem reduces to simply minimizing an objective function f that takes the form

$$f = \mathbf{w}\,\bar{\mathbf{q}} \tag{1}$$

such that

$$\mathbf{A}\,\bar{\mathbf{q}} = \bar{\mathbf{r}} \tag{2}$$

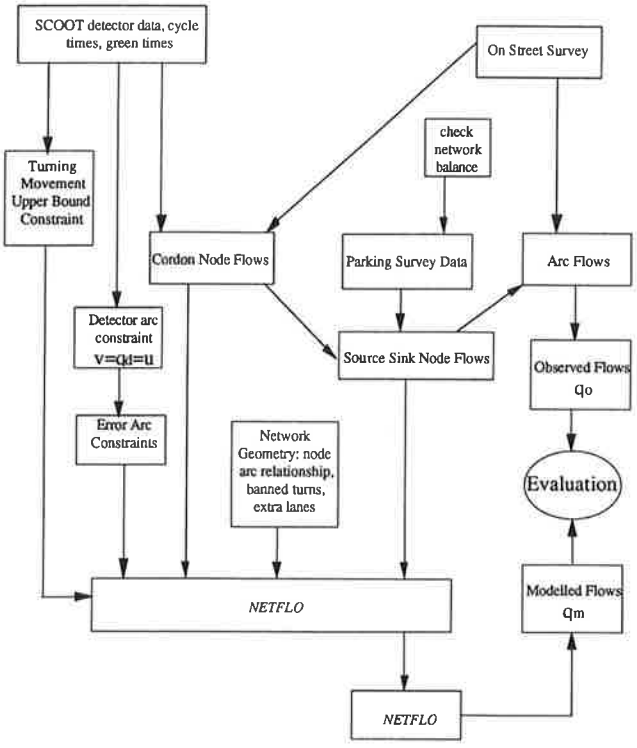


FIGURE 2 The modeling process.

and

$$\mathbf{0} \le \bar{\mathbf{v}} \le \bar{\mathbf{q}} \le \bar{\mathbf{u}} \tag{3}$$

where

- $\mathbf{A}$ = $i \times j$ node-arc incidence matrix, which defines the network structure;
- $\mathbf{w}$ = $1 \times j$ vector of unit weights associated with each arc;
- $\bar{\mathbf{r}}$ = $i \times 1$ vector of flows in and out of nodes, known as the requirement vector;
- $\bar{\mathbf{v}}$ = $j \times 1$ vector of arc lower bounds;
- $\bar{\mathbf{u}}$ = $j \times 1$ vector of arc upper bounds or arc capacities; and
- $\bar{\mathbf{q}}$ = $j \times 1$ vector of arc flow, the decision variable that is the unknown quantity.

The node-arc incidence matrix $\mathbf{A}$ is defined so that element A_{ij} of the array takes a value of $+1$ if Arc j is directed away from Node i and a value of -1 if Arc j is directed toward Node i. Should Arc j and Node i not meet, the value 0 is applied. If for Node i, $r > 0$, Node i is a supply node where traffic flows into the network; that is, the supply is equal to r. If for Node i, $r < 0$, Node i is a demand node and traffic flows out of the network. The value of entry or exit flow is r. Internal nodes or turning points at a junction have $r = 0$; these are known as transshipment points. Introducing the vectors $\mathbf{q}$, $\mathbf{r}$, and $\boldsymbol{\alpha}$ in order to transform the lower bound $\bar{\mathbf{v}}$, now let

$$\bar{\mathbf{q}} = \mathbf{q} + \bar{\mathbf{v}} \tag{4}$$

and

$$\mathbf{r} = \bar{\mathbf{r}} - \mathbf{A}\,\bar{\mathbf{v}} \tag{5}$$

$$\boldsymbol{\alpha} = \mathbf{w}\,\bar{\mathbf{v}} \tag{6}$$

$$\mathbf{u} = \bar{\mathbf{u}} - \bar{\mathbf{v}} \tag{7}$$

as the lower bounds are reduced to zero. The objective function then reduces to

$$f_{\min} = \mathbf{wq} + \boldsymbol{\alpha} \tag{8}$$

such that

$$\mathbf{Aq} = \mathbf{r} \tag{9}$$

and

$$\mathbf{0} \le \mathbf{q} \le \mathbf{u} \tag{10}$$

The NETFLO program applies a heuristic procedure by quickly finding paths through the network to satisfy node continuity. Spanning trees are supplemented by artificial arcs. A spanning tree is a subsidiary network that contains all the nodes but a number of arcs are reduced so that loops are eliminated. Part of a spanning tree is formed so as to satisfy the induced supply and induced demand. For each supply node, s, an artificial node, y, is set up, connected by an artificial arc (s,y) of weight ∞, capacity ∞, and flow t_{sy} satisfying node continuity. For each

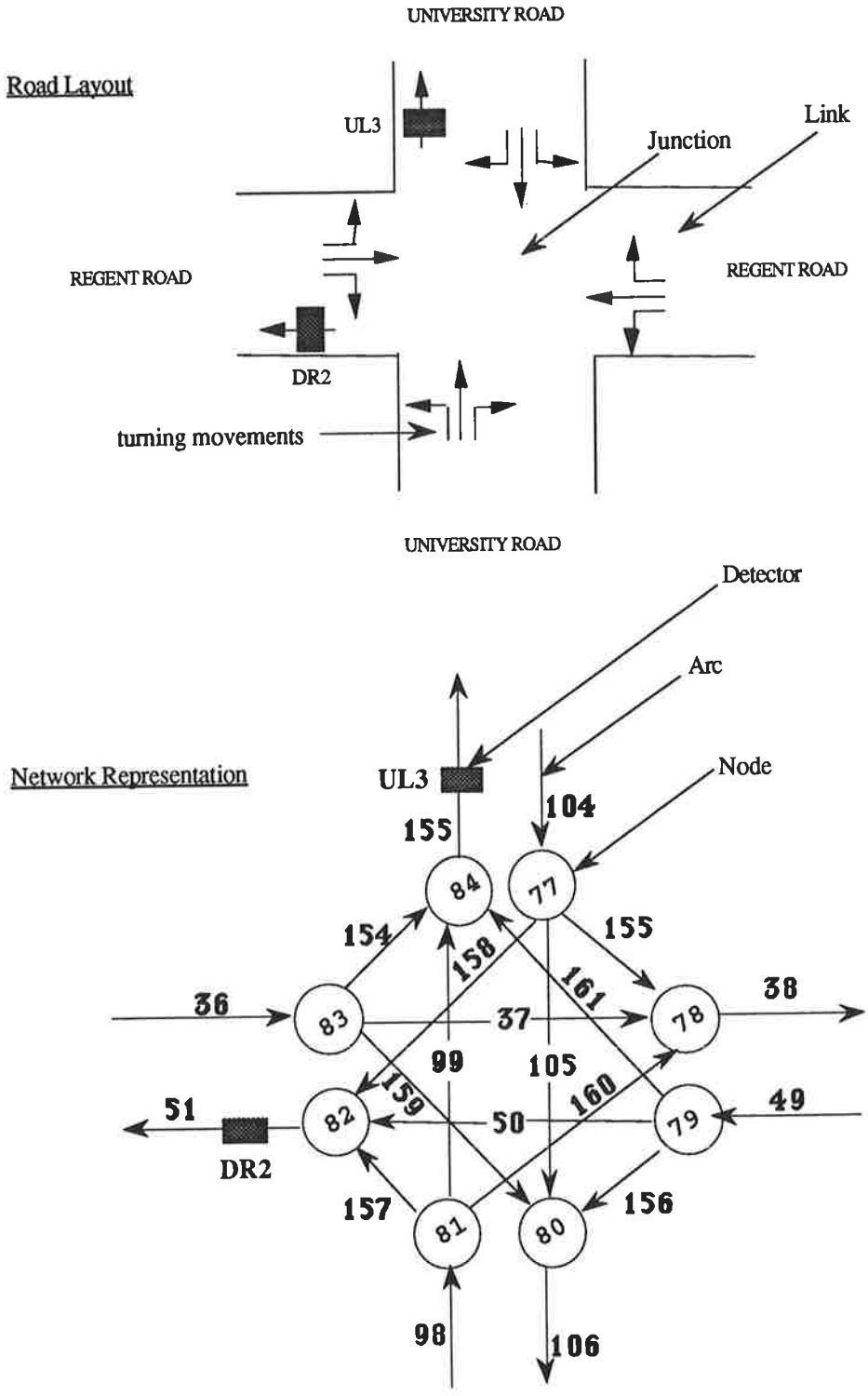


FIGURE 3 Node arc representation of a junction.

demand node (d), where flow leaves the network, an artificial node p is set up connected by an artificial arc (p,d) of weight ∞, capacity ∞, and flow t_{pd} satisfying node continuity.

Since the artificial arcs are given such high weights, the optimum solution is dominated by a set of flows whereby all artificial arcs are assigned zero flow. The approach is similar to the "Big-M" method (*11*), whereby the objective function is supplemented with an additional term M, which denotes a very large positive number and thus carries an overwhelming penalty.

The objective function now becomes

$$f_{\min} = \mathbf{wq} + \boldsymbol{\alpha} + M\mathbf{t} \quad (11)$$

where $\mathbf{t}$ is an $[m \times 1]$ vector of artificial arc flows, where m is the number of artificial arcs generated by the particular network. The spanning trees are completed with the addition of more artificial arcs, and basis exchanges are performed to achieve optimality.

NETWORK DEFINITION

A formal node-arc classification has been devised. The node-arc representation of a simple two-way junction at the intersection of Regent and University roads is shown in Figure 3. The notation is shown in Figure 4. Each traffic movement at the road intersection is expanded so that each turn is represented by its own unique arc. The arcs are represented by straight lines and the nodes by circles, triangles, or squares. Left-turn movements are uncrossed. All other turning movements, including straight-on maneuvers, are shown as crossed lines. Banned turning movements are accommodated by the absence of an arc in the network. Channelized lanes are modeled as separate arcs. The network representation of the street network of Region R is shown in Figure 5. Figures 2 and 5 have direct equivalence.

Five types of arc are defined:

1. An internal arc transmits flows from any node type to any other node type,
2. An external arc represents entry and exit flows; it is not afforded a label;
3. The external detector arc sits on the cordon perimeter and models the site of a SCOOT detector that supplies flow to a cordon node,
4. An internal detector arc models the site of a SCOOT detector within the network, and
5. A sorsink arc, like an external arc, is virtual, signifying the inflow or outflow of traffic from a source or sink within the cordon.

An external arc is a virtual arc that signifies the entry or exit flow into or out of the cordon. The model does not assign flow along either external arcs or external detector arcs because one end of the arc is free.

Three types of real node are defined:

1. A cordon node marks the point of entry to or exit from the network. Each is associated with an external arc and is connected to any number of simple arcs or detector arcs, or both.
2. Sorsink nodes mark the point of injection or extraction within the cordon. They always have one virtual sorsink arc, one entering arc (simple or detector), and one exit arc (simple or internal detector).
3. Transshipment nodes are not connected to external or sorsink arcs. They have any number of inflow arcs (simple or internal detector) and any number of outflow arcs (simple or internal detector).

CONSTRAINT REGIME

Node continuity is the principle of conservation of traffic flows at a node; it is analogous to Kirchoff's law for electric flow. For the linear program to respect node continuity and obey the law, net inflow must match net outflow. Cordon arcs model the external flows that enter and leave the network. The net imbalance of external flows is matched by a set of sorsink flows. Sorsink arcs model flows generated by internal sources and flows absorbed by internal sinks. The imbalance is distributed among the sources and sinks in proportion to the capacity of on-street parking and car parks. (Later enhancement of the model will include data showing level of

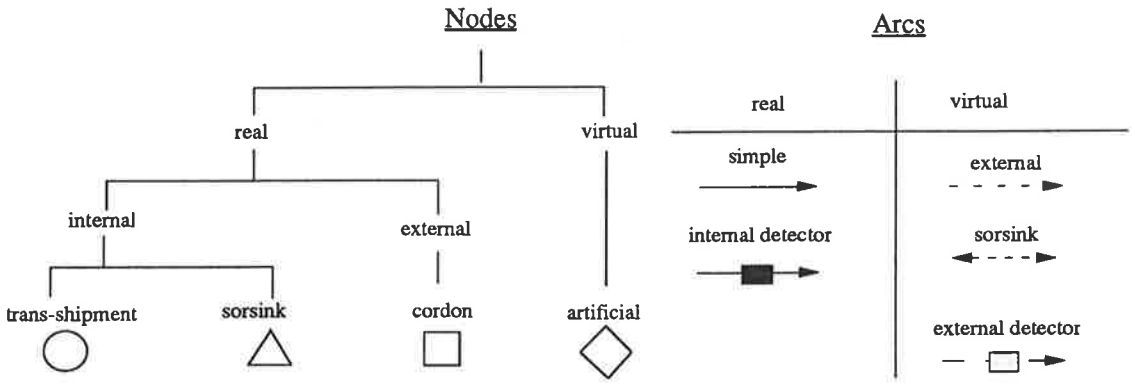


FIGURE 4 Network notation.

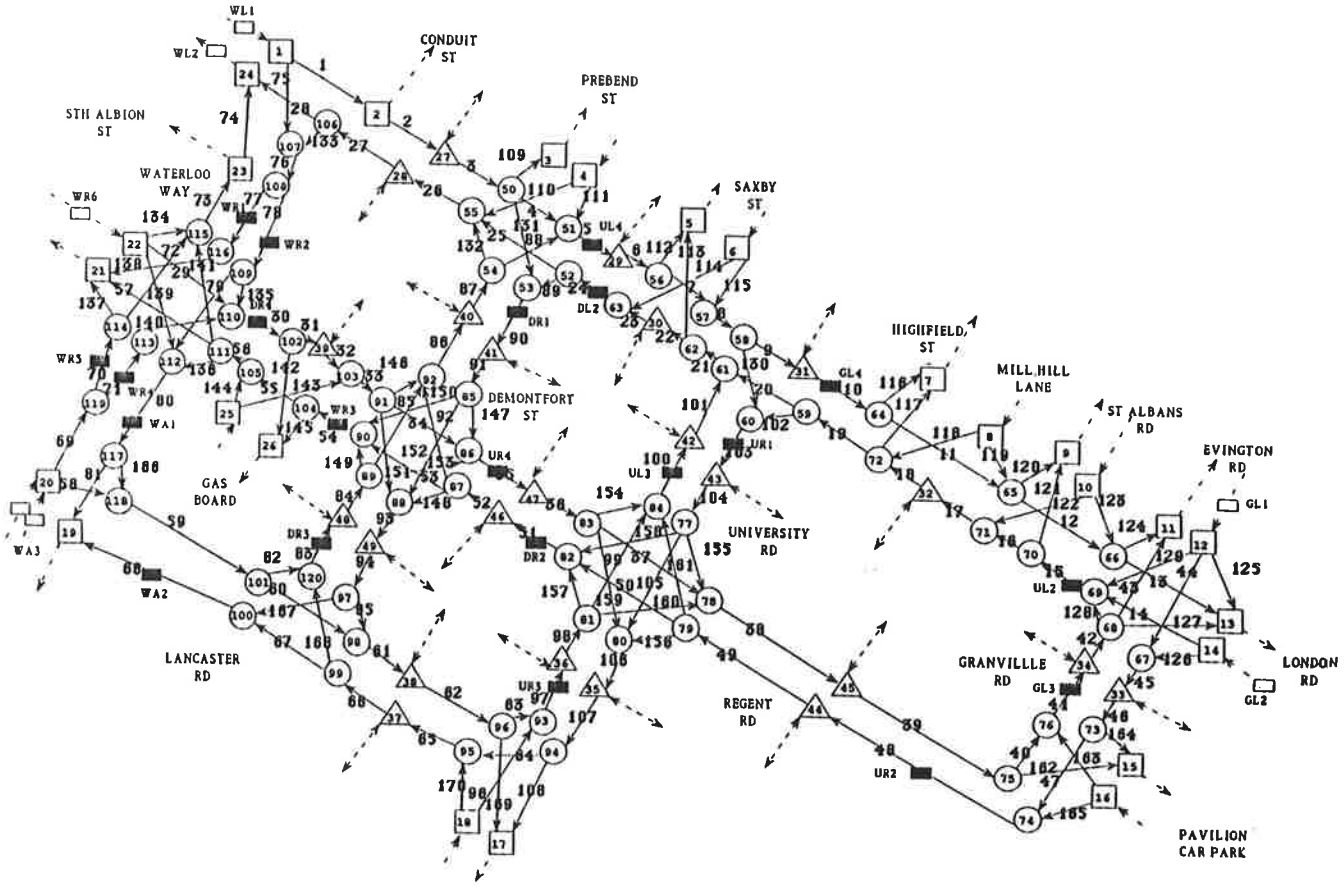


FIGURE 5 The network.

car park usage from automatic vehicle detection systems at the entrances and exits of carparks.)

Once the external and internal flows have been loaded onto the network, measured link flows from detectors are introduced. A degree of flexibility must be introduced to accommodate the dynamics or variability of the system and permit a feasible solution. A known flow, such as that measured on line from a detector, may be forced into the solution by specifying

$$u = v = q_d \tag{12}$$

where q_d is the detector flow.

However, fixing the upper and lower bounds in this way and tightly constraining the algorithm by forcing the detector flow along specified arcs leads to infeasibility. On the other hand, excessive relaxation of the upper and lower bounds merely serves to dilute the quality of the detector information. Therefore, a means of providing "room to breath" is needed. A series of constraint regimes was tested in order to simulate the effect of including detector data flows. Bell (7) proposed a method of introducing flexibility to the system by introducing an additional two error arcs $e+$ and $e-$ alongside the arc representing the measured flow, as shown in Figure 6.

This flexibility enabled the linear program to arrive at a feasible solution. In defining a solution for the network, the linear program assigns flows to arcs to minimize the sum of

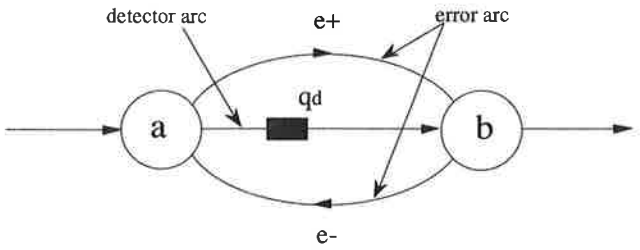


FIGURE 6 Error arc configuration.

the flows on the error arcs over the entire network. A solution is found such that

$\sum(e+ + e-)$ is a minimum for all links in the network

A series of tests investigated the sensitivity of the network to different error tolerances. These produced an optimum error arc configuration to produce a best and still feasible solution. The application of a range of weights to the error arcs failed to improve the reliability of reproducing the detector arc flows. A combination of varying upper and lower arc constraints along the detector arc itself with adjustment in the upper bounds of the error arc showed that an appropriate solution would be obtained with a maximum tolerance of 2.5 percent of the measured flow.

The SCOOT model supplies flow levels and signal timings every 5 min. The green splits and saturation flows provide capacities for each signal phase as follows:

$$u = \frac{gsx}{\tau} \tag{13}$$

where

g = green time for the link defined by the fixed-time plan,
s = saturation flow at the stop line,
x = degree of saturation, and
τ = cycle time, assumed fixed for the network over the 5-min period.

The effect of using measured flows on different arcs, equivalent to moving the locations of the detectors in the network, was simulated. Detectors were "moved" closer to junctions and straight-on maneuvers were simulated. The Demontfort Street–Regent Road junction was subject to complete definition by constraining each turning movement with the known flow. The existing SCOOT detector locations proved to give the best estimate.

Sensitivity tests applying lower-bound constraints demonstrated no useful purpose. On the contrary, they tended to reduce the reliability of flow prediction.

RESULTS

The simulated flows compared with the flows actually observed on site are shown in Figure 7. The cluster points on the x-axis, close to the origin, represent low flow turning movements, which have been assigned a zero by the algorithm. Straight-on maneuvers are reliably predicted. Half of the right-turn maneuvers are successfully predicted, but the model failed to predict the majority of left-turn movements. Overall, the model performs well, achieving a correlation coefficient of 92 percent.

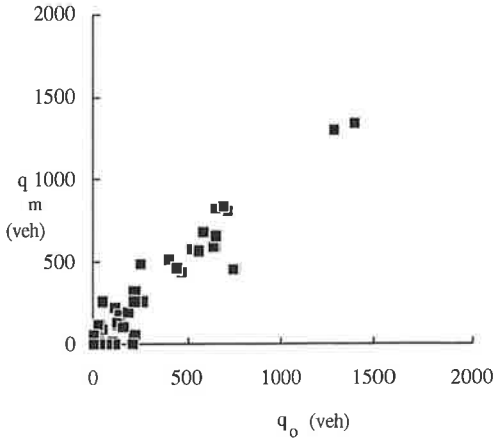


FIGURE 7 Turning movements: observed q_0 and modeled q_m.

CONCLUSIONS AND FURTHER WORK

The NETFLOW algorithm applied to traffic networks has been shown to be successful, yielding a correlation coefficient of 92 percent for predicted and actual traffic flows. However, it is the low flows, those of less importance in solving congestion problems, that are not reliably inferred.

The method proposed can deal with car parks as sinks and sources within the network. This element of the model is fairly complex and the modeling of car park storage is crude. More work is needed. Other limitations of the application of the model include the following:

1. The basic calculations are performed as integer arithmetic, which is especially important for the arc weights and for the calculation of sink and source flows from parking areas;
2. The minimum-flow parameter v cannot force the algorithm to assign a minimum flow to specified arcs; and
3. There is scope for an algorithm to be rendered dynamic by processing smoothed 5-min flow data in favor of the 1-hr aggregated flows.

The next step in this research program is to demonstrate a practical on-line application with traffic flows measured both into and out of the network. In the SCOOT environment, therefore, extra detectors would need to be installed. Further development of the algorithms will enable hypothetical incidents to be modeled by imposing heavy penalties on affected links. The flow predictions would provide a library of congestion control plans that could be implemented in conjunction with route guidance systems to improve traffic flow immediately after an incident. The algorithm could contribute to an expert system of traffic control.

The method stands apart from others in two ways. First, its route logic is node oriented. It moves through the network node by node, seeking to satisfy continuity by balancing inflows with outflows. The flows on consecutive links are effectively defined independently. Second, it has been structured to draw on detector data that are at the heart of a dynamically responsive unified traffic control model.

Transportation models that predict flows seek optimality by minimizing an objective function that usually has a behavioral basis. An objective function can take the form of the sum of a series of costed links. Costs reflect some form of material travel characteristic such as journey time, distance, or other perceived journey cost. An equilibrium model, for example, would suggest that drivers reroute when links approach capacity, minimizing journey length or travel time. Usually, there is a notion of cost implicit in the objective function minimized for optimality.

The approach described here has no behavioral basis because its purpose is to infer traffic movements from a limited set of detected flows. A weight function is introduced that serves as a controlling mechanism rather than a cost function. Some weights deter, whereas others encourage and no physical cost is minimized. The method does not seek to imitate drivers' route choice. In principle, it seeks to estimate the state of the traffic system (i.e., to derive turning movement flows and unknown flows) from geometric and traffic data. For this reason, the objective function plays a subordinate

role. It is a mechanism for identifying the one solution from the many possible solutions and has no tangible meaning. Since each flow prediction is associated with a set of constraints, the aim is to find a constraint regime whereby the minimized objective function is associated with a particular solution and unknown flows are reliably inferred.

Notwithstanding the subsidiary role of the objective function, however, there remains scope for its improvement. Furthermore, the link weight functions are linear. Since flow-speed curves are nonlinear, some degree of gradually increasing weights would offer more realistic traffic modeling.

ACKNOWLEDGMENTS

The authors acknowledge the Science and Engineering Research Council and the Department of Civil Engineering at the University of Nottingham for funding the research and the Leicestershire County Council for their close cooperation and support.

REFERENCES

1. M. McDonald, N. B. Hounsell, N. Sittampalam, and F. N. McLeod. *Traffic Incidents and Route Guidance in a SCOOT Network*. Science and Engineering Research Council, United Kingdom, Oct. 1987.
2. P. Davies, D. R. Salter, and M. Bettison. Loop Sensors for Vehicle Classification. *Traffic Engineering and Control*, Vol. 23, Feb. 1982, pp. 55–59.
3. M. C. Bell and P. T. Martin. The Use of Traffic Detector Data for Traffic Control Strategies. In *Institution of Electrical Engineers International Conference on Road Traffic Control*. Conference Publication 320. IEE, London, May 1990.
4. H. J. Van Zuylen and L. G. Willumsen. The Most Likely Trip Matrix Estimated from Traffic Counts. *Transport Research B*, Vol. 14B, 1980, pp. 281–293.
5. H. J. Van Zuylen. The Estimation of Turning Flows on a Junction. *Traffic Engineering and Control*, Vol. 20, Nov. 1979, pp. 539–541.
6. M. G. H. Bell, D. Inaudi, J. Lange, and M. Maher. Techniques for the Dynamic Estimation of O-D Matrices in Traffic Networks. In *Proceedings of the DRIVE Conference: Advanced Telematics in Road Transport*. 4–6 Feb. 1991, Brussels, pp. 1040–1056.
7. M. C. Bell. *The Use of Automatic Control Algorithms To Define Urban Traffic Routes*. Internal Report. Durham University, 1987.
8. P. B. Hunt, D. I. Robertson, R. D. Bretherton, and R. I. Winton. *SCOOT A Traffic Responsive Method of Coordinating Signals*. TRRL Report LR 1014. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1981.
9. G. B. Danzig, A. Orden, and P. Wolfe. The Generalised Simplex Method for Minimising a Linear Form under Linear Inequality Restraints. *Pacific Journal of Mathematics*, Vol. 5, 1955, pp. 183–195.
10. J. L. Kennington and R. V. Helgason. *Algorithms for Network Programming*. John Wiley & Sons, New York, 1980.
11. F. S. Hillier and G. J. Lieberman. *Introduction to Operations Research*. McGraw-Hill, 5th ed., 1990, 81 pp.

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