

TRANSPORTATION RESEARCH RECORD 905

Traffic Flow, Capacity, and Measurements

TRANSPORTATION RESEARCH BOARD

*NATIONAL RESEARCH COUNCIL
NATIONAL ACADEMY OF SCIENCES*

WASHINGTON, D.C. 1983

Transportation Research Record 905
Price \$21.80
Edited for TRB by Naomi Kassabian

mode
1 highway transportation

subject area
55 traffic flow, capacity, and measurements

Library of Congress Cataloging in Publication Data
National Research Council. Transportation Research Board.
Traffic flow, capacity, and measurements.

(Transportation research record; 905)

Reports prepared for the 62nd annual meeting of the Transportation Research Board.

Includes bibliographical references.

1. Traffic flow—Congresses. 2. Highway capacity—Congresses. 3. Traffic surveys—Congresses. I. National Research Council (U.S.). Transportation Research Board. II. Series.
TE7.H5 no. 905 [HE336.T7] 380.5s 83-17412
ISBN 0-309-03519-8 [388.3'14] ISSN 0361-1981

Sponsorship of the Papers in This Transportation Research Record

GROUP 3—OPERATION AND MAINTENANCE OF TRANSPORTATION FACILITIES

Patricia F. Waller, University of North Carolina, chairman

Committee on Traffic Control Devices

Ken F. Kobetsky, West Virginia Department of Highways, chairman
H. Milton Heywood, Federal Highway Administration, secretary
Robert L. Carstens, Robert Dewar, Paul Fowler, Robert M. Garrett, C. Arthur Geurts, Robert David Henry, Kay K. Kerkorian, C. John MacGowan, Hugh W. McGee, Zoltan A. Nemeth, Peter S. Parsonson, Harold C. Rhudy, David G. Snider, Ronald E. Stemmler, Richard H. Sullivan, James I. Taylor, James A. Thompson, Jonathan Upchurch, Clinton A. Venable, W. Scott Wainwright, Earl C. Williams, Jr.

Committee on Highway Capacity and Quality of Service

James H. Kell, JHK & Associates, chairman
William R. McShane, Polytechnic Institute of New York, secretary
George W. Black, Jr., Robert C. Blumenthal, James B. Borden, Arthur A. Carter, Jr., Joseph W. Hess, V.F. Hurdle, Frank J. Koepke, Jerry Kraft, Walter H. Kraft, Joel P. Leisch, Edward Lieberman, Louis E. Lipp, Carroll J. Messer, Stephen Edwin Rowe, Alexander Werner

Committee on Traffic Flow Theory and Characteristics

John J. Haynes, University of Texas at Arlington, chairman
Edmund A. Hodgkins, Federal Highway Administration, secretary
Patrick J. Athol, E. Ryerson Case, Said M. Easa, John W. Erdman, Nathan H. Gartner, Richard L. Hollinger, Matthew J. Huber, Joseph K. Lam, Tenny N. Lam, Edward Lieberman, C. John MacGowan, William R. McShane, Carroll J. Messer, Panos Michalopoulos, Robert H. Paine, Harold J. Payne, Thomas W. Rioux, Paul Ross, Richard Rothery, Steven R. Shapiro, Yosef Sheffi

Committee on Methodology for Evaluating Highway Improvements

William D. Glauz, Midwest Research Institute, chairman
William T. Baker, Federal Highway Administration, secretary
William D. Berg, Eugene D. Brenning, Charles Philip Brinkman, Forrest M. Council, Kenneth G. Courage, John F. DiRenzo, John P. Eicher, Robert L. Gordon, William Grenke, David L. Guell, Alfred-Shalom Hakkert, Bradley T. Hargroves, Carla Heaton, David L. Helman, Robert David Henry, Ernst Meyer, David Perkins, Jerry G. Pigman, Frederick S. Scholz, Roy W. Taylor, Harold T. Thompson, Leonard B. West, Jr., Paul H. Wright

David K. Witheford, Transportation Research Board staff.

Sponsorship is indicated by a footnote at the end of each report. The organizational units, officers, and members are as of December 31, 1982.

Contents

MINIMIZING COST OF MANUAL TRAFFIC COUNTS: CANADIAN EXAMPLE	
Satish C. Sharma	1
TANDEM TOLL BOOTHS FOR THE GOLDEN GATE BRIDGE	
Randolph W. Hall and Carlos F. Daganzo	7
TANDEM TOLL COLLECTION SYSTEMS (Abridgment)	
Louis D. Rubenstein	14
RELIABILITY OF CLASSIFIED TRAFFIC COUNT DATA	
Peter Davies and David R. Salter	17
APPLICATION OF COUNTING DISTRIBUTION FOR HIGH-VARIANCE URBAN TRAFFIC COUNTS	
Stephen G. Ritchie	27
DELAY MODELS OF TRAFFIC-ACTUATED SIGNAL CONTROLS	
Feng-Bor Lin and Farrokh Mazdeyasna	33
Discussion	
Kenneth G. Courage	36
Authors' Closure	36
ANOTHER LOOK AT BANDWIDTH MAXIMIZATION	
Karsten G. Baass	38
CALIBRATION OF TRANSYT PLATOON DISPERSION MODEL FOR PASSENGER CARS UNDER LOW-FRICTION TRAFFIC FLOW CONDITIONS (Abridgment)	
Patrick T. McCoy, Elizabeth A. Balderson, Richard T. Hsueh, and Abbas K. Mohaddes.	48
EVALUATION OF DYNAMIC FREEWAY FLOW MODEL BY USING FIELD DATA	
N.A. Derzko, A.J. Ugge, and E.R. Case	52
Discussion	
Harold J. Payne	59
Authors' Closure	60
PASSENGER-CAR EQUIVALENTS FOR RURAL HIGHWAYS	
Wiley D. Cunagin and Carroll J. Messer	61
Discussion	
Roger P. Roess	67

TRAFFIC DATA-COLLECTION SYSTEMS: CURRENT PROBLEMS AND FUTURE PROMISE (Abridgment)	
Richard W. Lyles and John H. Wyman	69
ANALYSIS OF TRANSYT PLATOON-DISPERSION ALGORITHM	
Nagui M. Roupail	72
OPTIMIZATION MODEL FOR ISOLATED SIGNALIZED TRAFFIC INTERSECTIONS	
W.B. Cronjé	80
COMPARISON OF SOAP AND NETSIM: PRETIMED AND ACTUATED SIGNAL CONTRLS	
Zoltan A. Nemeth and James R. Mekemson	84
ANALYSIS OF EXISTING FORMULAS FOR DELAY, OVERFLOW, AND STOPS	
W.B. Cronjé	89
DERIVATION OF EQUATIONS FOR QUEUE LENGTH, STOPS, AND DELAY FOR FIXED-TIME TRAFFIC SIGNALS	
W.B. Cronjé	93
DEVELOPMENTAL STUDY OF IMPLEMENTATION GUIDELINES FOR LEFT-TURN TREATMENTS	
Han-Jei Lin and Randy B. Machemehl	96
DETERMINING CAPACITY AND SELECTING APPROPRIATE TYPE OF CONTROL AT ONE-LANE TWO-WAY CONSTRUCTION SITES	
Panos G. Michalopoulos and Roger Plum	105
SINGLE-LANE CAPACITY OF URBAN FREEWAY DURING RECONSTRUCTION (Abridgment)	
Robert E. Dudash and A.G.R. Bullen	115
COMPARATIVE ANALYSIS OF SIGNALIZED-INTERSECTION CAPACITY METHODS	
Adolf D. May, Ergun Gedizlioglu, and Lawrence Tai	118
SPEEDS AND FLOWS ON AN URBAN FREEWAY: SOME MEASUREMENTS AND A HYPOTHESIS	
V.F. Hurdle and P.K. Datta	127
EFFECTIVENESS EVALUATION BY USING NONACCIDENT MEASURES OF EFFECTIVENESS (Abridgment)	
David D. Perkins and Brian L. Bowman	138
SURROGATE MEASURES FOR ACCIDENT EXPERIENCE AT RURAL ISOLATED HORIZONTAL CURVES	
Harold T. Thompson and David D. Perkins	142
CANDIDATE ACCIDENT SURROGATES FOR HIGHWAY SAFETY ANALYSIS	
David D. Perkins and Harold T. Thompson	147
SKELETON PROCEDURE FOR EVALUATION OF HIGHWAY SAFETY IMPROVEMENTS ON A ROAD NETWORK	
D. Mahalel	153

EVALUATING NEED FOR ACCIDENT-REDUCTION EXPERIMENTS	
William D. Berg and Camil Fuchs	159
COMMON BIAS IN BEFORE-AND-AFTER ACCIDENT COMPARISONS AND ITS ELIMINATION	
Ezra Hauer and Bhagwant Persaud	164

Authors of the Papers in This Record

Baass, Karsten G., Transportation Division, Department of Civil Engineering, Ecole Polytechnique, Université de Montréal, Case postale 6079, succursale "A", Montreal, Québec H3C 3A7, Canada
Balderson, Elizabeth A., Department of Civil Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0346
Berg, William D., Department of Civil and Environmental Engineering, Engineering Building, 1415 Johnson Drive, University of Wisconsin-Madison, Madison, WI 53706
Bowman, Brian L., Goodell-Grivas, Inc., 17320 W. Eight Mile Road, Southfield, MI 48075
Bullen, A.G.R., Department of Engineering, University of Pittsburgh, Pittsburgh, PA 15213
Case, E.R., Transportation Technology and Energy Branch, Ontario Ministry of Transportation and Communications, Room 330, Central Building, 1201 Wilson Avenue, Downsview, Ontario M3M 1J8, Canada
Courage, Kenneth G., Department of Civil Engineering, University of Florida, Gainesville, FL 32611
Cronjé, W.B., Department of Civil Engineering, University of Stellenbosch, Stellenbosch 7600, South Africa
Cunagin, Wiley D., Texas Transportation Institute, Texas A&M University, College Station, TX 77843
Daganzo, Carlos F., Department of Civil Engineering, University of California, Berkeley, CA 94720
Datta, P.K., Department of Civil Engineering, University of Toronto, Toronto M5S 1A4, Canada
Davies, Peter, Department of Civil Engineering, University of Nottingham, University Park, Nottingham NG7 2RD, England
Derzko, N.A., Department of Mathematics, University of Toronto, Toronto 181, Ontario, Canada
Dudash, Robert E., Michael Baker, Jr., Inc., 4301 Dutch Ridge Road, Beaver, PA 15009
Fuchs, Camil, Tel-Aviv University, Ramat Aviv, Tel Aviv, Israel
Gedizlioglu, Ergun, Macka Insaat Fak Ulastirma ve Trafik K. Besiktas, Istanbul, Turkey
Hall, Randolph W., Transportation Research Department, General Motors Research Laboratories, Warren, MI 48090
Hauer, Ezra, Department of Civil Engineering, University of Toronto, Toronto M5S 1A1, Canada
Hsueh, Richard T., Department of Civil Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0346
Hurdle, V.F., Department of Civil Engineering, University of Toronto, Toronto M5S 1A4, Canada
Lin, Feng-Bor, Department of Civil and Environmental Engineering, Clarkson College of Technology, Potsdam, NY 13676
Lin, Han-Jei, Department of Civil Engineering, University of Texas at Austin, Austin, TX 78712
Lyles, Richard W., College of Engineering, Michigan State University, East Lansing, MI 48824-1212
Machemehl, Randy B., Department of Civil Engineering, University of Texas at Austin, Austin, TX 78712
Mahalel, David, Transportation Research Institute, Technion-Israel Institute of Technology, Technion City, 32000 Haifa, Israel
May, Adolf D., Institute of Transportation Studies, 109 McLaughlin Hall, University of California at Berkeley, Berkeley, CA 94720
Mazdeyasna, Farrokh, Department of Civil and Environmental Engineering, Clarkson College of Technology, Potsdam, NY 13676
McCoy, Patrick T., Department of Civil Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0346
Mekemson, James R., Department of Civil Engineering, Ohio State University, 2070 Neil Avenue, Columbus, OH 43210
Messer, Carroll J., Texas Transportation Institute, Texas A&M University, College Station, TX 77843
Michalopoulos, Panos G., Department of Civil and Mineral Engineering, University of Minnesota, Minneapolis, MN 55455-0220
Mohaddes, Abbas K., Department of Civil Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0346
Nemeth, Zoltan A., Department of Civil Engineering, Ohio State University, 2070 Neil Avenue, Columbus, OH 43210
Payne, Harold J., VERAC Incorporated, 10975 Torreyana Road, Suite 300, San Diego, CA 92121
Perkins, David D., Goodell-Grivas, Inc., 17320 W. Eight Mile Road, Southfield, MI 48075
Persaud, Bhagwant, Department of Civil Engineering, University of Toronto, Toronto M5S 1A1, Canada
Plum, Roger, Department of Civil and Mineral Engineering, University of Minnesota, Minneapolis, MN 55455-0220
Ritchie, Stephen G., Department of Environmental Engineering, Cornell University, Ithaca, NY 14853
Roess, Roger P., Polytechnic Institute of New York, 333 Jay Street, Brooklyn, NY 11201
Rouphail, Nagui M., Department of Civil Engineering, Mechanics and Metallurgy, University of Illinois at Chicago, Chicago, IL 60680

Rubenstein, Louis D., Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109

Salter, David R., Department of Civil Engineering, University of Nottingham, University Park, Nottingham NG7 2RD, England

Sharma, Satish C., Faculty of Engineering, University of Regina, Regina, Saskatchewan S4S 0A2, Canada

Tai, Lawrence, Institute of Transportation Studies, 109 McLaughlin Hall, University of California at Berkeley, Berkeley, CA 94720

Thompson, Harold T., Department of Civil Engineering, University of Notre Dame, Box G, Notre Dame, IN 46556

Ugge, A.J., Transportation Technology and Energy Branch, Ontario Ministry of Transportation and Communications, Room 330, Central Building, 1201 Wilson Avenue, Downsview, Ontario M3M 1J8, Canada

Wyman, John H., Maine Facility Laboratory, Maine Department of Transportation, Augusta, ME 04330

Minimizing Cost of Manual Traffic Counts: Canadian Example

SATISH C. SHARMA

The accuracy and cost-effectiveness of short-period manual traffic counts are analyzed. Alberta's primary highway system is investigated in this study. The roads in the system under investigation are classified into four types: (a) commuter sites, (b) nonrecreational low-volume sites, (c) rural long-distance sites, and (d) recreational sites. The accuracy of short-period counts is expressed in terms of a deviation either side of the estimated volume, which defines limits of the interval in which the actual volume is most likely to be. In the tests of this study, a relative measure of deviation, namely, the coefficient of variation, is used in order to compare the variation in several sets of data for the counts of different durations and schedules. The analysis carried out in this paper illustrates clearly that the most important considerations for rationalization of short-period manual counts are (a) the type of road site being surveyed and (b) the hour-to-hour traffic variations within the same day. The month of the year, the day of the week, and the duration of counts are other significant factors that must be considered in order to devise the most efficient short-period counts.

Several types of traffic-counting programs are undertaken by roadway agencies to obtain values of average annual daily traffic (AADT) and other traffic data for their road networks. The most commonly used programs are (a) continuous counting by permanent traffic counters (PTCs); (b) seasonal counting by portable counters, where counts are taken a few times a year for periods from 48 h to 2 weeks in length; and (c) short-period counting, where manual traffic counts are undertaken for less than a day. The PTCs provide actual temporal distribution of traffic movement and the true values of AADT. The seasonal and short-period counts furnish only sample information and therefore need appropriate factoring to yield the estimates of AADT values.

In addition to the estimates of AADT, the short-period counting programs provide such important data as vehicle classification and turning movements, which are frequently required for planning and design of roads for both safety and economy purposes. The proposed study is concerned with the short-period manual counting programs.

All the provincial transportation agencies in Canada undertake short-period traffic-counting programs on an annual basis. Most of such traffic counting is carried out by students during the spring and summer seasons, which includes May, June, July, and August. The number of students hired for this purpose varies from province to province and is generally in the range of 10-20 students for the entire period.

Although it is true that the short-period manual counting is undertaken for a period of less than a day, there is a considerable difference in the actual durations and schedules adopted by the different provinces. For example, the Ministère des Transports in Quebec generally carries out 12-h (7:00 a.m. to 7:00 p.m.) and 8-h (7:00-11:00 a.m. and 3:00-7:00 p.m.) schedules, Alberta Transportation uses 12-h (7:00 a.m. to 7:00 p.m.) and 9-h (8:00 a.m. to 5:00 p.m.) schedules, and Ontario's Ministry of Transportation and Communications (MTC) employs 8-h (7:00-11:00 a.m. and 2:00-6:00 p.m.) schedules.

There are two important aspects that should be considered in relation to the short-period manual counting programs. One is that the quality of data used for planning and design purposes is a crucial factor that affects the reliability of the results.

The other is that, because collecting data is expensive, especially when considerable overtime wages are involved in 9-h or 12-h counting, the method of collection should be as cost-effective as possible.

During these times of budgetary constraints, some authorities feel that the improvement in accuracy obtained by extending a traffic count at a spot location beyond 6 h may be small. But very little scientific work has been done toward any systematic comparison between the 12-h traffic surveys and the shorter surveys in the context of the provincial or rural roads in Canada. The main objectives of this study are (a) to analyze the accuracy and cost-effectiveness of the existing programs, such as 12-h, 9-h, and 8-h counts, as compared with shorter manual traffic counts; (b) to study the influence of road type and traffic volume on the accuracy of different short-period surveys; and (c) to specify appropriate schedules of the shorter counts if they are reasonable in terms of the accuracy of the results.

BACKGROUND

Estimation of AADT from Short-Period Counts

The usual method for estimating AADT from sample counts is that advocated by the U.S. Bureau of Public Roads (BPR) in its Guide for Traffic Volume Counting Manual (1). In general, the BPR method involves (a) grouping together the PTC sites into similar patterns of monthly traffic variation, (b) determining average expansion factors for each group, (c) assigning road sections that do not have PTCs to one of these groups, and (d) applying the appropriate average expansion factor to sample counts to produce an estimate of AADT.

A commonly used form of mathematical relationship for estimating AADT from sample counts of shorter duration than 24 h is that in which the count is expanded first to 24-h volume by using an hourly expansion factor (H), second to average daily volume by using a daily expansion factor (D), and then to the annual flow by using a seasonal expansion factor (S). This formula may be expressed as follows:

$$\text{Estimated AADT} = \text{short-period volume count} \times H \times D \times S \quad (1)$$

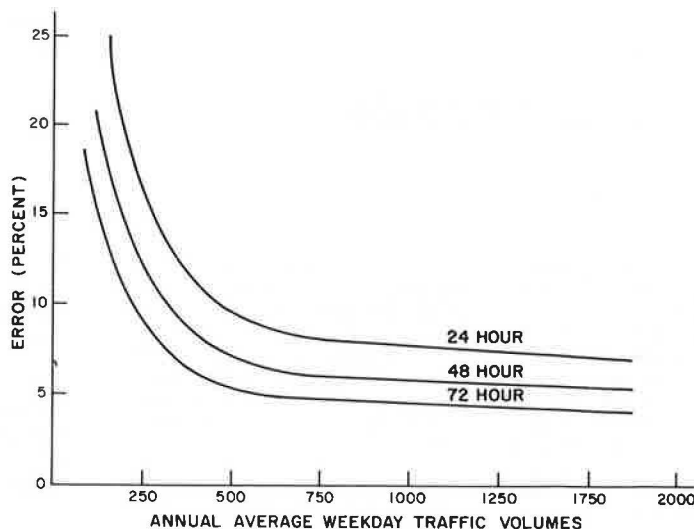
As indicated earlier, the values of average expansion factors for different groups of roads are computed from the PTC data. These factors are defined as follows:

$$\text{Hourly factor (H)} = (\text{avg volume for 24-h period}) / (\text{avg volume for particular duration of count}) \quad (2)$$

$$\text{Daily factor (D)} = [(\text{avg total volume for week}) / 7] / (\text{avg volume for particular day}) \quad (3)$$

$$\text{Seasonal factor (S)} = [(\text{total yearly volume}) / 12] / (\text{total volume for particular month}) \quad (4)$$

Figure 1. Estimation errors as function of weekday traffic volume and duration of counting.



Errors of AADT Estimates

There are four sources of error in estimating AADT at a point by using Equation 1:

1. The hourly factor at a counting site generally will not be exactly equal to the group mean;
2. The daily factor at the counting site will differ from the average daily factor for the group;
3. The seasonal factor at the site will not be exactly the same as the mean group seasonal factor; and
4. The road section on which a count is taken may have been assigned to a wrong PTC or road group; this error is assumed to be negligible (2).

The magnitude of error due to the hourly factor is generally expected to be a function of the duration and schedule of a particular short survey. However, any variation in the duration and schedule of a short-period count will not affect the errors due to the daily factor and the seasonal factor.

In the past, there have been studies to determine the effect of the duration of sample counts on the accuracy of resulting AADT estimates. The results of one such study concerning the so-called "coverage counts" (or seasonal traffic counts) were published by Petroff and Blensly (3). Commenting on Figure 1, which is adapted from the U.S. study (3), the authors write (3, p. 364):

The observation of the data presented in Figure 1 which is of utmost practical significance is that traffic counts of 24-hour duration on weekdays have a coefficient of variation of 10 percent or less when compared with the mean volume for a weekday in a given month at stations having the mean volume of about 500 vehicles per day or more. This applies usually to all months except the winter months in some of the states.... Counts of 48 hours duration improve the accuracy by 20 to 25 percent, thus raising the confidence limit from 68 percent to about 75 percent for one standard deviation of 10 percent, also extending the range of volumes down to about 300 vpd.

This translated into everyday language means that two thirds to three fourths, depending on the length of the count, of all coverage or blanket counts may be expected to have an error of about 10 percent or less when compared with the true mean weekday volume of the month during

which they were taken when volumes are 300 to 500 vehicles per day or more.

The above observations are for sample counts of one day's duration or longer in rural areas. A very limited amount of work has been reported in the literature concerning the errors of shorter-duration counts.

The study conducted by the Local Government Operational Research Unit (4) for the Department of Transport in Britain showed that a reasonably accurate estimate of AADT can usually be based on a single 16-h count. Another British study (5) on short-period counting reported:

A six-hour count in the afternoon (13:00h-19:00h or 14:00h-20:00h) on a weekday in late spring or early autumn will provide a reasonable estimate of the annual flow....

If a more accurate estimate of annual flow is required it would be better to repeat the six-hour count later in the same month rather than increase the length of counting to 16 hours. The expected accuracy from two six-hour counts is similar to that from a single 16-hour count, even though four hours less counting is undertaken.

The principal focus of this paper is to analyze the errors associated with the existing manual counting programs as used in Alberta, Ontario, and Quebec and to investigate the potential increases in the errors if shorter counts (e.g., 6-h and 4-h programs) are adopted by the provincial authorities. The study does not include the errors associated with daily and monthly expansion factors.

STUDY APPROACH

Statistical Accuracy of Short-Period Counts

The accuracy of short-period counts can be expressed in terms of deviation either side of the estimated 24-h volume, which defines limits of the interval in which the actual 24-h volume is most likely to lie. In the tests of this study, it was necessary to use a relative measure of deviation in order to compare the variation in several sets of data for counts of different durations. The coefficient of variation (CV) was used for this purpose. For a particular short-period count at a given roadway site, CV was defined as follows:

$$CV = \left\{ \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X})^2 \right\}^{1/2} / \bar{X} = SD/\bar{X} \quad (5)$$

where

- X_i = i th volume count at site,
 n = total number of volume counts taken at site,
 and
 \bar{X} = average value of n volume counts.

CV, as defined in Equation 5, is the SD as a fraction (or percentage) of the mean value \bar{X} . A low value of CV, which is associated with less dispersion of individual volume data about their mean, reflects a high accuracy in estimating the actual traffic volumes.

Study Data

Since the PTCs when grouped are considered to represent the population of the group, their statistical measures of variation (such as SD and CV) are also the measures of the errors at the short-period counting sites; the latter are also samples taken out of the same population.

For the purpose of this study, Alberta's PTC data were analyzed. The PTC information for the years 1978, 1979, and 1980 were included in the study. After the reliability of the available Alberta PTC data for those years was considered, a total of 41 PTCs was selected to be used in this investigation.

Because the short-period manual traffic counting in Alberta and other provinces is carried out mainly in the spring and summer seasons, only the months of May, June, July, and August were chosen for the analysis. Computations for the CV were carried out by each day of the week for different counting schedules. The holidays, such as Victoria Day and Canada Day, were eliminated from the analysis. A number of certain other days that displayed a very unusual volume and pattern of traffic (such as that due to reconstruction of a facility) were also excluded from the study. Otherwise, it was assumed that the hourly traffic variations at study sites did not change significantly from 1978 to 1980.

Classification of Road Sites

One of the most important causes of variability in the traffic flows and the errors of AADT estimates is the nature of the road sites surveyed. Actually, it is believed that the difference between sites can be so large as to overwhelm other causes of variability in predictions. On the basis of two recent studies (6,7) on Alberta highways, the study sites were classified into four broad types according to their temporal variations in traffic flows and such other characteristics as trip purpose and trip-length distribution. These types are as follows:

1. Commuter sites, e.g., the PTC site C9 located on Highway 3 east of Lethbridge;
2. Nonrecreational low-volume (rural) sites, e.g., the PTC site C147 located on Highway 35 north of Grimshaw;
3. Rural long-distance sites, e.g., the PTC site C18 located on the Trans-Canada Highway west of Medicine Hat; and
4. Recreational sites, e.g., the PTC site C114 located on Yellowhead Highway east of Jasper National Park.

Trip purpose information for the typical examples of these classes, i.e., sites C9, C147, C18, and C114, has been provided elsewhere (6).

Selection of Study Schedules

Other important considerations in manual traffic

surveys are the duration and the schedule of counting. For a particular day, the best short period will be that which (a) has the most stable relationship with the annual flow and (b) contains the most representative and important levels (e.g., the evening peak) of traffic that occur in the course of the day.

Sample schedules in this study were selected mainly from the considerations of (a) the current practice in Canada and (b) the variability in the volumes of different duration counts in the day. According to current practice, three schedules were chosen. These are 12 h (7:00 a.m. to 7:00 p.m.), 9 h (8:00 a.m. to 5:00 p.m.), and 8 h (7:00-11:00 a.m. and 2:00-6:00 p.m.).

The CVs for continuous counts of 8 h and shorter duration were investigated for all classes of roads, as shown in Figure 2 for the rural long-distance road site C18. It became apparent that for the best results shorter manual counts would have to be carried out with their midpoints at 3:00 or 4:00 in the afternoon. However, for the purpose of detailed discussion and comparisons, some other schedules were also selected in this study. On the assumption that all manual counts would be conducted between 7:00 a.m. and 7:00 p.m., the following schedules were chosen in addition to the existing three schedules:

1. 6 h (a.m.) (7:00 a.m. to 1:00 p.m.),
2. 6 h (p.m.) (1:00-7:00 p.m.),
3. 4 h (a.m.) (7:00-11:00 a.m.),
4. 4 h (p.m.) (2:00-6:00 p.m.), and
5. 2 h (p.m.) (4:00-6:00 p.m.).

RESULTS AND DISCUSSION

Tables 1-4 contain the computed values of CVs of traffic volume recorded at the typical road sites for the selected schedules of short-period counting. The tabulations are made for each day of the week for the months of May and July. Figures 3-6 are drawn by using the data from Tables 1-4, respectively, but for the sake of simplicity of presentation and discussion, the values of CV are averaged

Figure 2. Stability of weekday volume counts: May 1978, 1979, and 1980 at rural long-distance site C18.

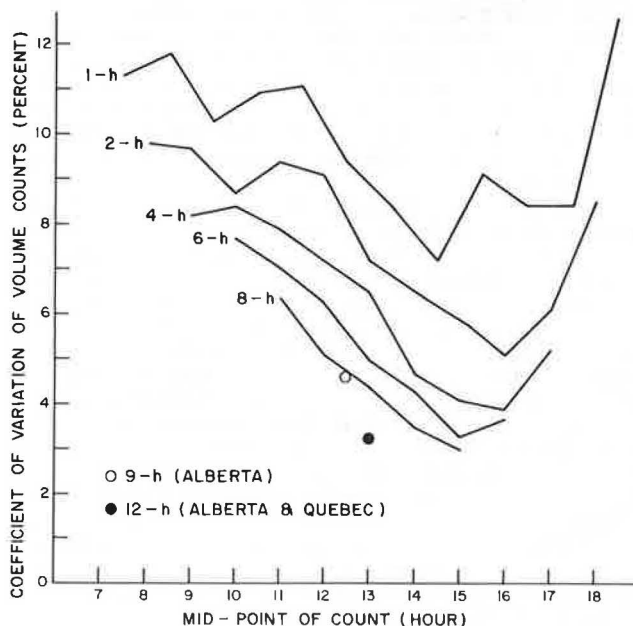


Table 1. CV of traffic volume recorded at PTC site C9 (commuter) for selected schedules of short-period counting.

Month	Counting Schedule (h)	CV of Recorded Traffic Volume (%)						
		Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
May	12	2.7	1.4	1.3	1.3	1.4	2.2	2.0
	9	3.0	2.1	2.5	2.0	2.4	3.1	2.2
	8	2.8	1.7	2.0	1.8	1.7	2.1	3.6
	6 (a.m.)	7.2	2.6	2.9	2.6	2.0	3.7	4.8
	6 (p.m.)	2.7	1.7	2.3	2.0	2.1	2.1	2.9
	4 (a.m.)	9.6	3.3	4.3	4.4	2.8	3.5	8.3
	4 (p.m.)	2.8	2.1	2.4	2.1	2.8	2.6	3.0
	2 (p.m.)	3.6	2.5	4.3	3.4	2.5	2.3	3.8
	2 (p.m.)	3.6	2.5	4.3	3.4	2.5	2.3	3.8
July	12	2.9	1.1	1.2	2.1	2.1	2.3	4.1
	9	2.7	1.4	1.3	3.1	2.5	2.5	5.1
	8	3.2	3.5	2.1	3.6	2.9	3.3	5.2
	6 (a.m.)	6.0	3.0	3.6	3.5	3.3	2.6	9.1
	6 (p.m.)	2.4	1.4	1.8	2.2	1.9	2.8	3.3
	4 (a.m.)	7.9	6.3	5.1	4.8	5.0	3.7	13.4
	4 (p.m.)	3.4	2.3	2.1	3.8	2.6	3.6	3.9
	2 (p.m.)	4.4	2.6	3.2	2.9	3.8	4.4	4.0
	2 (p.m.)	4.4	2.6	3.2	2.9	3.8	4.4	4.0

Table 2. CV of traffic volume recorded at PTC site C147 (nonrecreational low volume) for selected schedules of short-period counting.

Month	Counting Schedule (h)	CV of Recorded Traffic Volume (%)						
		Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
May	12	2.5	2.5	2.2	1.8	3.6	3.4	3.0
	9	6.5	5.6	3.2	3.4	5.9	5.2	4.3
	8	7.3	1.8	2.8	1.5	4.5	4.1	4.9
	6 (a.m.)	14.8	9.5	4.6	4.9	6.8	4.0	7.7
	6 (p.m.)	5.2	4.0	4.8	4.1	3.3	5.1	4.2
	4 (a.m.)	23.8	8.3	6.5	6.3	7.2	5.1	14.1
	4 (p.m.)	10.7	4.4	5.5	5.7	6.6	6.7	7.1
	2 (p.m.)	11.5	9.2	6.6	8.9	8.4	6.6	10.8
	2 (p.m.)	11.5	9.2	6.6	8.9	8.4	6.6	10.8
July	12	4.3	2.7	3.2	2.5	2.4	2.0	3.2
	9	6.7	4.4	4.5	3.6	3.6	2.4	3.7
	8	6.3	5.4	4.1	2.9	2.7	3.0	4.6
	6 (a.m.)	11.6	6.6	5.2	3.8	3.3	4.1	5.4
	6 (p.m.)	4.4	3.3	4.3	3.3	3.7	5.0	4.4
	4 (a.m.)	14.9	10.8	6.6	4.9	4.6	7.3	10.1
	4 (p.m.)	4.6	5.6	5.7	3.7	3.6	6.0	6.0
	2 (p.m.)	5.1	8.8	6.0	6.2	8.6	3.8	8.7
	2 (p.m.)	5.1	8.8	6.0	6.2	8.6	3.8	8.7

for the weekdays and the results for May and July are plotted in these figures. It may be mentioned here that the results for the months of June and August are omitted because of their close similarity to the results of May and July, respectively.

The following presentation of results and discussion is made by taking the examples of the typical road sites C9, C147, C18, and C114. But it should be noted that the computations of CV were also made for the remaining 37 sites and the results obtained at those sites were similar to their respective typical sites.

Examination of Figures 3-6 reveals several important facts that are common to each type of road site. One observation is that the CV for a 12-h schedule is smaller than those for other schedules. Therefore, the 12-h schedule can be expected to provide the most accurate estimates of traffic statistics. Another striking observation from these figures is that Alberta's 9-h short-period schedule produces higher values of CV as compared with the 6-h (p.m.) schedule in a great majority of cases. Actually, the results, such as those shown in these figures, indicate that many times even the 4-h (p.m.) schedule can provide as good an estimate of volume as does the 9-h schedule. It may also be noted from the values of CV that the 8-h schedule yields better results than the 9-h schedule. The results for the 8-h schedule are generally similar to those of the 6-h (p.m.) schedule.

Table 3. CV of traffic volume recorded at PTC site C18 (rural long distance) for selected schedules of short-period counting.

Month	Counting Schedule (h)	CV of Recorded Traffic Volume (%)						
		Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
May	12	2.7	2.8	2.2	2.5	3.0	3.4	2.7
	9	4.1	2.3	3.6	3.5	4.0	4.0	3.7
	8	3.1	2.9	2.5	3.0	3.5	4.2	3.2
	6 (a.m.)	11.7	3.3	3.3	2.8	5.1	8.4	5.3
	6 (p.m.)	6.8	3.0	2.7	3.5	3.0	3.2	5.0
	4 (a.m.)	15.4	4.2	3.8	2.8	5.2	9.7	8.3
	4 (p.m.)	7.0	3.2	3.1	3.9	4.1	3.0	6.6
	2 (p.m.)	8.5	7.0	5.3	7.3	6.7	4.3	8.7
	2 (p.m.)	8.5	7.0	5.3	7.3	6.7	4.3	8.7
July	12	2.2	2.0	2.0	1.9	2.5	1.8	2.2
	9	3.2	3.0	3.3	2.7	3.6	2.3	2.6
	8	2.0	2.6	2.2	2.3	2.6	3.1	2.4
	6 (a.m.)	2.5	3.1	3.3	2.5	3.9	3.5	3.6
	6 (p.m.)	3.6	1.7	3.5	2.9	3.1	2.6	3.9
	4 (a.m.)	5.5	3.2	4.8	3.5	4.6	5.2	5.5
	4 (p.m.)	4.1	3.1	4.2	3.2	3.2	3.6	4.0
	2 (p.m.)	5.6	3.5	4.5	3.2	3.9	3.9	3.4
	2 (p.m.)	5.6	3.5	4.5	3.2	3.9	3.9	3.4

Table 4. CV of traffic volume recorded at PTC site C114 (recreational) for selected schedules of short-period counting.

Month	Counting Schedule (h)	CV of Recorded Traffic Volume (%)						
		Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
May	12	7.6	6.1	6.0	5.6	3.9	6.2	7.1
	9	13.0	8.8	8.9	8.0	5.7	8.6	11.2
	8	9.3	7.9	8.3	7.5	6.9	6.6	9.7
	6 (a.m.)	23.6	11.2	9.1	7.4	7.4	13.0	12.1
	6 (p.m.)	9.3	7.8	5.7	6.8	5.7	8.7	8.8
	4 (a.m.)	37.5	21.0	16.5	11.7	14.0	17.9	21.0
	4 (p.m.)	10.8	8.0	6.5	8.6	6.7	9.5	10.9
	2 (p.m.)	14.2	8.9	9.5	10.8	6.8	8.8	11.7
	2 (p.m.)	14.2	8.9	9.5	10.8	6.8	8.8	11.7
July	12	4.1	2.6	4.0	3.2	2.7	5.9	2.7
	9	5.7	4.9	5.8	6.1	4.1	8.6	4.1
	8	7.0	5.6	5.9	5.5	5.0	8.6	5.9
	6 (a.m.)	12.8	15.2	12.3	12.3	8.5	18.4	10.9
	6 (p.m.)	4.5	6.1	3.7	4.4	4.4	3.6	4.3
	4 (a.m.)	25.9	29.8	19.0	19.7	18.0	27.7	19.8
	4 (p.m.)	5.3	5.7	4.6	4.0	4.5	4.2	3.9
	2 (p.m.)	7.8	9.5	7.4	7.9	6.3	9.5	7.1
	2 (p.m.)	7.8	9.5	7.4	7.9	6.3	9.5	7.1

Figures 3-6 also indicate clearly that the morning hours are not very appropriate for short-period counting. For example, the values of CV for the 6-h (a.m.) and 4-h (a.m.) schedules are considerably higher than the values for the 6-h (p.m.) and the 4-h (p.m.) schedules, respectively. In fact, it should be noted that a short-period count of 2-4 h in the late afternoon will provide the same accuracy of data as a 6-h count in the morning.

Prior to examining the effect of road sites on the accuracy of short-period counts, let us consider the influence of traffic volume (or AADT) on the expected values of CV. Figure 7 presents the average values of CV as a function of AADT. Smooth hand-fitted curves in the figure are drawn from the scatter of CV statistics for all 41 study sites for the weekdays of May to August. The X-axis in the figure is the average AADT for 1978, 1979, and 1980--the years for which PTC data are analyzed in this investigation.

These curves clearly indicate that the errors in estimated volumes are expected to be less for the road that carries large volumes of traffic. However, it can also be noted that beyond a certain critical range of AADT, the errors are not likely to decrease significantly with further increase in traffic volume. For example, this critical range of AADT from Figure 7 is (a) 2000 for the 12-h schedule, (b) 3000 for the 6-h (p.m.) schedule, and (c) 4000 for the 2-h (p.m.) schedule.

There is one important factor to be considered in relation to the effect of road sites on the accuracy of counts. It is that, in general, different classes of road sites carry different amounts of traffic volume. For example, commuter sites located near large population centers are likely to carry much heavier traffic volumes as compared with other road sites, which by their nature serve primarily such specific purposes as long-distance farm-to-market or highly recreational trips. The average

AADT values for the four typical road sites of this analysis are 8800 for C9, the commuter site; 1180 for C147, the nonrecreational low-volume site; 3500 for C18, the rural long-distance site; and 2055 for C114, the recreational site.

The effect of road types on estimated errors of counting can be deduced by referring back to Figures 3-6. It is evident that the values of CV or the errors of estimation are functions of the location of short-period counts. As generally expected, the

Figure 3. CVs of traffic volumes: weekdays of May and July in 1978, 1979, and 1980 at commuter site C9.

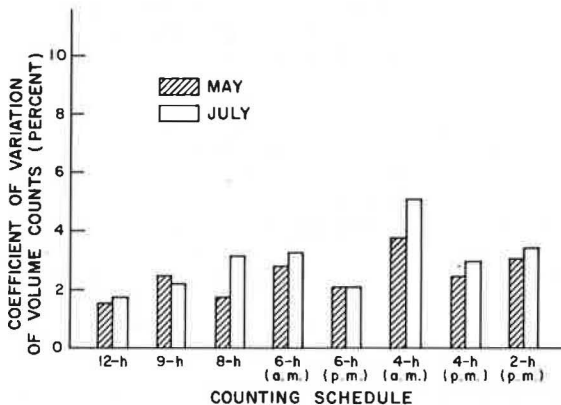


Figure 4. CVs of traffic volumes: weekdays of May and July in 1978, 1979, and 1980 at nonrecreational low-volume site C147.

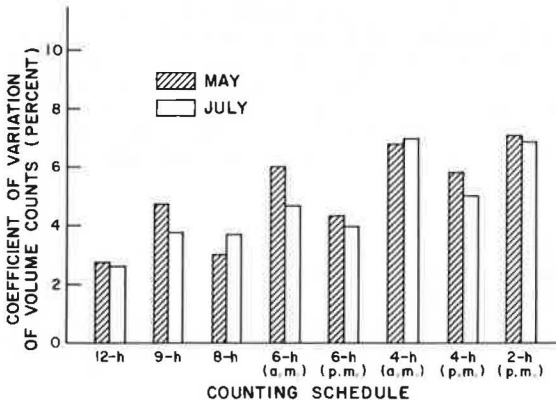


Figure 7. CVs of traffic volumes: weekdays of May, June, July, and August in 1978, 1979, and 1980; averages of all study sites.

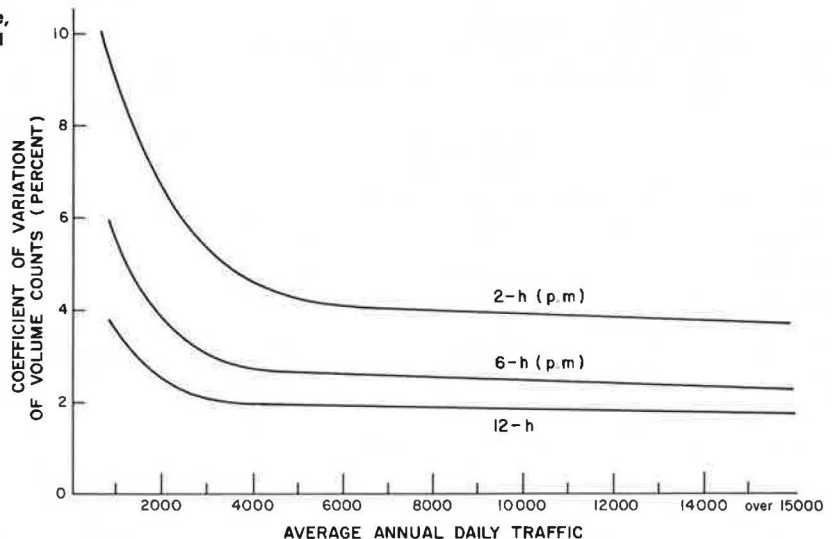


Figure 5. CVs of traffic volumes: weekdays of May and July in 1978, 1979, and 1980 at rural long-distance site C18.

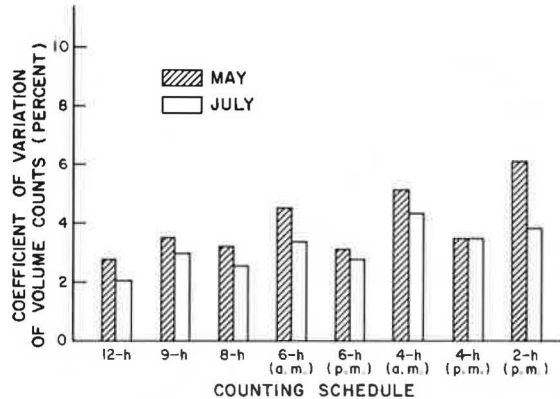
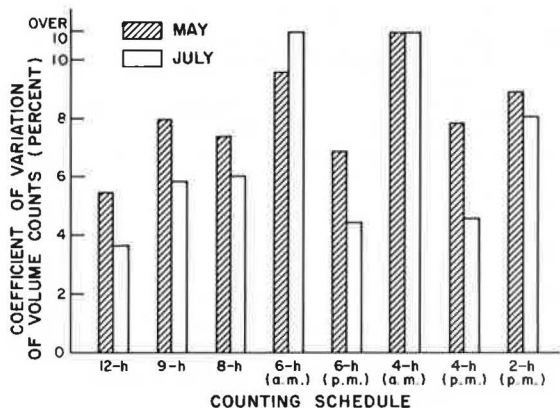


Figure 6. CVs of traffic volumes: weekdays of May and July in 1978, 1979, and 1980 at recreational site C114.



lowest errors are observed in the case of commuter site C9 and the highest in the case of recreational site C114. The values of CV in those figures also suggest that the 2-h (p.m.) schedule at a commuter site, or the 4-h (p.m.) schedule at a rural long-distance site, or the 6-h (p.m.) schedule at a non-recreation low-volume site are all likely to produce the same accuracy as would the 12-h schedule at a recreational site.

Even though the volume of traffic, particularly when AADT is less than 2000, may have some effect on the accuracy of counts (Figure 7), it is believed that the difference in the accuracy of counts at various sites is due primarily to trip-purpose characteristics at the site being surveyed. The work-business trips are considered to be less variable from day to day as compared with the social-recreational trips. A good example in this respect is to compare the results for the C147 low-volume rural site and those for the C114 recreational site. Even though site C147 has a lower value of AADT, its results are more accurate as compared with those of site C114. This difference can be attributed to the trip-purpose characteristics. The social-recreational components of trip purpose for C147 and C114 during the summer weekdays are 22 and 75 percent, respectively (6).

A formal analysis of variance pertaining to the effects of days, months, and their possible interactions with different road types was not carried out in this study. However, the study results such as those included in Tables 1-4 and Figures 3-6 seem to indicate that (a) there is no effect of weekdays on the accuracy of counts and (b) the months have some interaction with the nature of the sites surveyed. It is apparent that site C18 (rural long distance) and site C114 (recreational) are expected to produce the most accurate estimates during the months of July (and August) when the tourist-recreational travel is at peak levels in Alberta. In contrast, the other two sites seem to provide better results during May (and June) when the work-business trips are still at their normal levels.

CONCLUSIONS

The analysis carried out in this paper illustrates clearly that the most important considerations for rationalization of short-period manual counts are (a) the type of road site being surveyed and (b) the hour-to-hour traffic variations within the same day. The month of the year, the day of the week, and the duration of the counts are other significant factors that must be considered in order to design the most efficient schedules of short-period counts. The following specific conclusions are drawn from this investigation of 41 PTC sites in the province of Alberta:

1. For the counts of 8 h or less on weekdays, a period with a midpoint at 3:00 or 4:00 p.m. is expected to provide the most accurate volume estimates for each class of road. An additional advantage of including this period is that peak-hour turning movements and vehicle classification can still be observed because such a counting period generally includes the evening peak of traffic volume.

2. The accuracy of short-period counts is a function of the nature of road sites surveyed. The greatest accuracy is expected for commuter sites and the least in the case of highly recreational sites. In fact, the values of CV computed for the study sites of this investigation indicate that the 2-h (p.m.) schedule at a commuter site, the 4-h (p.m.) schedule at a rural long-distance site, or the 6-h (p.m.) schedule at a nonrecreational low-volume site

are all likely to produce the same accuracy as would the 12-h schedule at a recreational site.

3. There seems to be some interaction between the type of road surveyed and the month of counting. For example, recreational roads produce more reliable traffic estimates in the months of July and August when tourist-recreational travel is at peak levels. In contrast, commuter sites provide better information during May and June.

4. The 9-h schedule as currently used in Alberta produces less accurate volume estimates compared with the 6-h (p.m.) schedule in a great majority of cases. Actually, the results indicate that many times even the 4-h (p.m.) schedule can provide as good estimates of volume as does the 9-h schedule. The accuracy of the 8-h schedule as used in Ontario is generally similar to that of the 6-h (p.m.) schedule.

5. The traditional 12-h surveys yield the most accurate estimates of 24-h volume. The differences in the accuracy of the 12-h schedule and the carefully selected 6-h (p.m.) schedule are (a) 0-1 percent at commuter and long-distance rural sites and (b) 1-2 percent for nonrecreational low-volume sites and recreational sites. However, since the overtime wage rules in Canada increase the personnel cost of the 12-h count to nearly three times that of a shorter 6-h count, the traditional 12-h surveys are less cost-effective than the 6-h (p.m.) surveys. Actually, in many cases, even the 4-h (p.m.) schedules could be considered reasonably accurate and cost-effective.

6. Another conclusion of this study is the effect of AADT on the accuracy of short-period counts. In general, the errors in volume are expected to be less for the roads that carry large volumes of traffic. But beyond a certain critical range of AADT, the errors are not likely to decrease significantly with further increase in traffic volume. For example, this critical range of AADT appears to be 2000 for the 12-h survey, 3000 for the 6-h (p.m.) survey, and 4000 for the 2-h (p.m.) survey.

The findings of this research provide a better understanding of the factors that affect the accuracy of estimating traffic volumes. It is hoped that with this better understanding, agencies will be able to design and schedule more cost-effective short-period traffic-counting programs without any loss in accuracy.

ACKNOWLEDGMENT

I am grateful to Alberta Transportation for providing the necessary data. Particular thanks are due Al Werner, manager of planning services at Alberta Transportation, for his time and interest in the project. Michel De Caen assisted in the computations for this project.

The financial assistance of the Natural Sciences and Engineering Research Council of Canada is also acknowledged.

REFERENCES

1. Guide for Traffic Volume Counting Manual. Bureau of Public Roads, U.S. Department of Commerce, 1965.
2. R.R. Bodle. Evaluation of Rural Coverage Count Duration for Estimating Annual Average Daily Traffic. Bureau of Public Roads, U.S. Department of Commerce, Highway Planning Technical Rept. 5, 1966.
3. B.B. Petroff, and R.C. Blensly. Improving Traffic-Count Procedures by Application of Statistical

- cal Method. HRB Proc., Vol. 33, 1954, pp. 362-375.
4. Annual Traffic Multipliers for Use in COBA. Local Government Operational Research Unit, Department of Transport, London, England, 1978.
5. G. Phillips. When to Mount a Traffic Count. Traffic Engineering and Control, Vol. 21, No. 1, 1980, pp. 4-6.
6. S.C. Sharma and A. Werner. Improved Method of Grouping Provincewide Permanent Traffic Counters. TRB, Transportation Research Record 815, 1981, pp. 12-18.
7. S.C. Sharma. Toward a Standard Functional Classification of Provincial Roads in Canada. Presented at 1981 Roads and Transportation Association of Canada Annual Conference, Winnipeg, Canada, 1982.

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

Tandem Toll Booths for the Golden Gate Bridge

RANDOLPH W. HALL AND CARLOS F. DAGANZO

Many toll plazas are constrained in width by buildings or other physical barriers. These barriers may make the cost of adding toll lanes prohibitive. One method for increasing the capacity of a toll facility without increasing its width is to use tandem toll booths. A tandem toll booth consists of two toll takers in a single toll lane both serving alternating sets of vehicles simultaneously. The capacity of tandem toll booths is calculated with time-space diagrams and the cumulative headway distributions of vehicles at a conventional toll booth. The capacity depends on the maximum of two random variables, which correspond to the service times at the two booths, and is found by taking the product of their cumulative headway distributions. Adjacent tandem toll booths were found to increase the capacity of the Golden Gate Bridge toll plaza by about 15 percent, and batch tandem toll booths increase capacity by 25 percent or more. Thus, tandem toll booths would eliminate the queueing that now exists during the morning commute period without the cost of expanding the toll plaza's width.

The Golden Gate Bridge is the primary transportation link between Marin County, California, and the City of San Francisco. Each day 100 000 vehicles traverse the six-lane 2-mile span, 20 000 of which travel in each commute period. The bridge is operated by the Golden Gate Bridge, Highway, and Transportation District and is financed by user fees. Tolls are collected in the southbound direction at the south (San Francisco) end of the bridge.

In February 1981 the bridge board of directors approved an increase in the automobile toll from \$1.00 to \$1.25. Soon afterward, it became evident that the toll plaza could no longer accommodate peak-period traffic. Queues extended as far back as 4 miles; occasionally delays were up to 30 min. Predictably, many motorists were upset and the new toll became the subject of media scrutiny.

The ability of a toll facility to accommodate large traffic flow depends on two factors: the number of servers (toll takers) and the service time per vehicle. The Golden Gate solution to the queuing problem was to reduce the service time per vehicle. Delay had increased with the \$1.25 toll because more motorists needed change and because some motorists folded their dollar bill around their quarter (which caused the toll taker to spend extra time sorting money). The added delay caused the \$1.25 toll to be rescinded in July 1981. The toll was eventually replaced by a split toll of \$1.00 from Sunday through Thursday and \$2.00 on Friday and Saturday. Since that time the number of complaints has dropped considerably, but the equity of the split toll has been questioned. Clearly, the split toll was not motivated by traditional pricing con-

siderations but simply by the need to reduce service time while maintaining revenue.

The alternative approach of increasing the number of servers was perceived to be infeasible in the short run. The peculiar geography of the facility meant that adding toll lanes would require relocation of the entire toll plaza at a cost of \$16 million (1). Other ideas, such as building separate toll facilities for the different bridge lanes, would also be capital intensive.

The purpose of this paper is to demonstrate the consequences of adding toll booths to existing lanes in series rather than adding toll lanes in parallel. The unit of study will be a pair of toll takers serving a single traffic lane. Such a unit, which we call a tandem toll booth, can potentially increase the per-lane throughput, or capacity, and may obviate the need for increasing the width of the toll facility. When the width of a toll facility is physically constrained (by buildings or other barriers), as is the case with the Golden Gate Bridge, tandem toll booths may be cost efficient. Because the length of the toll plaza is also restricted, this paper focuses on tandem-toll-booth strategies that are not greatly affected by the distance between the upstream and downstream toll takers. Although other strategies (such as staggered toll plazas for different lanes or the alternating-toll-booth strategy, described by Rubenstein in another paper in this Record) might increase capacity more, queuing can be nearly eliminated at the Golden Gate Bridge without resorting to these capital-intensive alternatives.

We next describe how to calculate the capacity of tandem toll booths and then report on a case study at the Golden Gate Bridge.

PER-LANE CAPACITY

Single Toll Booths

Before the operation of a tandem toll booth is explained, a single toll booth will be considered. Let the service position be the location of a vehicle when it pays its toll and let the waiting position be the location of the following vehicle in line (Figure 1). Furthermore, let vehicles be numbered 0, 1, 2, ... beginning from the vehicle in the service position.

At the time vehicle 1 leaves the service position, the headway between it and vehicle 0 (H) equals the sum of (a) the headway at the time vehicle 0 left service (H_0) and (b) the service time for vehicle 1 (S), which is the time needed to pay the toll. For the vehicle trajectories on the time-space diagram of Figure 2, H_0 is the sum of a reaction time (R) and a move-up time (M). The reaction time equals the elapsed time between the moment vehicle 0 leaves the service position and the moment vehicle 1 begins to move into the service position. The move-up time equals the time needed to drive from the waiting position to the service position. However, some drivers do not actually stop before entering the service position (Figure 3). Rather, many coast into service at a slow speed, coming to a halt only if vehicle 0 takes particularly long to pay its toll. For these drivers, the reaction time

R is virtually zero. (In general, reaction times are very small because drivers anticipate the departure of vehicles from the booth by watching the toll takers.)

If the headway varies randomly from vehicle to vehicle, its expectation is given by

$$E(H) = E(H_0) + E(S) \quad E(H) = E(R) + E(M) + E(S) \quad (1a)$$

One vehicle is processed per headway, so the capacity per toll lane (C) is the inverse of $E(H)$:

$$C = 1/E(H) \quad (1b)$$

Tandem Toll Booth: Adjacent Servers

Adjacent tandem toll booths have two service positions (position 1 in the front and position 2 in the rear) and no waiting room between (Figure 4). The waiting positions for the following two vehicles are identically numbered. Vehicles that wait at position 1 are served at position 1, and vehicles that wait at position 2 are served at position 2. Let vehicles now be numbered 0, 1, 2, ... beginning from the vehicle initially at service position 2.

In Figure 5, the elapsed time from the moment vehicle 0 leaves service until the moment vehicle 1 completes and leaves service (T_1) is the sum of a reaction, move-up, and service time:

$$T_1 = R_1' + M_1' + S_1' \quad (2)$$

where R_1' , M_1' , and S_1' are, respectively, the reaction time, move-up time, and service time of vehicle 1.

Vehicle 2 cannot enter its service position until vehicle 1 does the same. Thus, the elapsed time from the moment vehicle 0 leaves service until vehicle 2 completes service (T_2) contains an additional reaction time:

$$T_2 = R_1' + (R_2' + M_2' + S_2') \quad (3)$$

Figure 1. Conventional toll booth.

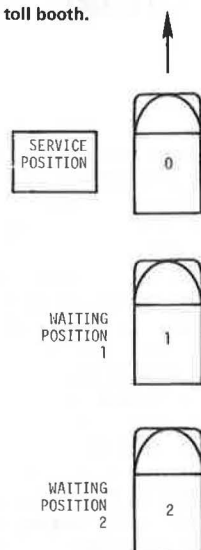


Figure 2. Vehicle trajectories at conventional booth.

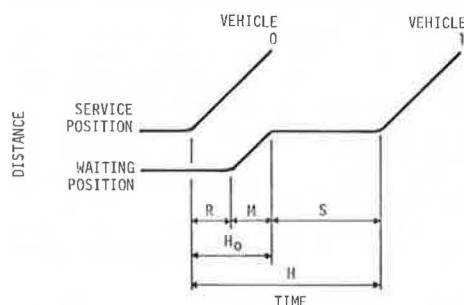


Figure 3. Vehicle coasting into service at conventional booth.

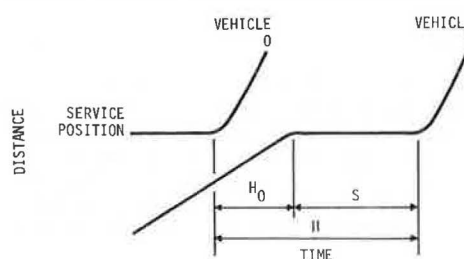


Figure 4. Tandem toll booth.

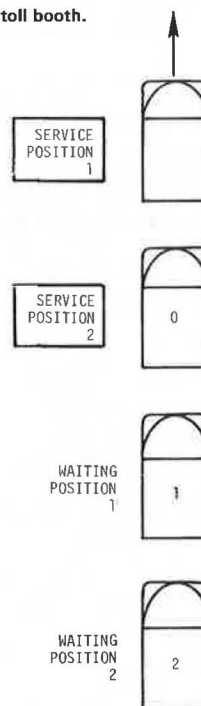
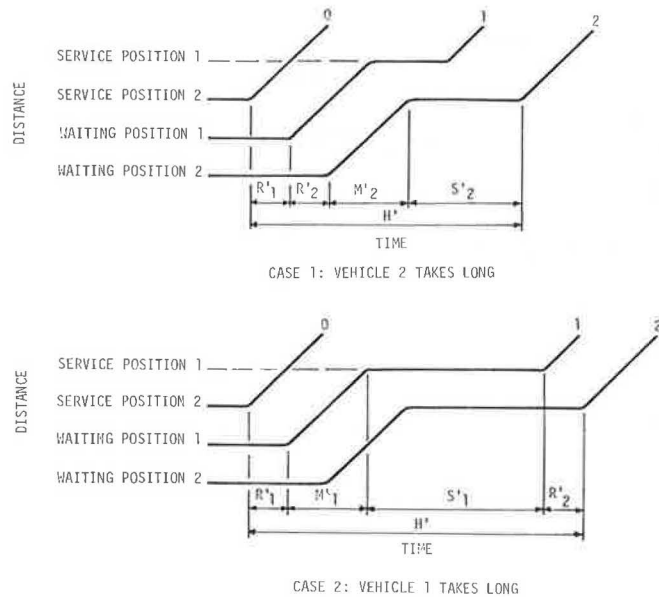


Figure 5. Vehicle trajectories at adjacent tandem toll booths.



However, since service position 1 is just one car space from service position 2, vehicle 2 cannot leave service until its driver perceives that vehicle 1 has left service. Thus, the elapsed cycle time from the moment vehicle 0 leaves service until vehicle 2 leaves service (H') must be greater than or equal to $T_1 + R_2$:

$$H' \geq T_1 + R_2 \quad (4)$$

In fact, vehicle 2 will depart after both T_2' and $(T_1 + R_2)$ have elapsed. Thus, the cycle time H' between vehicles 0 and 2 is

$$H' = \max(T_2, T_1 + R_2) \\ = \max[R_1 + (R_2 + M_2 + S_2), R_2 + (R_1 + M_1 + S_1)] \quad (5)$$

Suppose that vehicles 1 and 2 were observed as they passed through a conventional toll booth. R_1 and R_2 would then be their approximate reaction times at this single booth (R_1 and R_2), and S_1 and S_2 would be their service times (S_1 and S_2). However, move-up time for the single booth would be less than that for a tandem booth. Move-up time for vehicle 1 (M_1) would equal the M_1 observed at the single booth plus a value ΔM equal to the time needed to traverse the extra distance between the two service positions (Figures 2 and 5). The extra move-up time for vehicle 2 would equal the time needed to traverse an extra car position in the queue. Since the spacing between service positions is approximately the same as the distance between cars in the queue, the move-up time for vehicle 2 would also be $M_2 + \Delta M$. Thus, additional move-up time (ΔM) can be estimated by dividing the distance between vehicles queued at a conventional booth by their peak velocity as they drive into the service position.

If we substitute the service, move-up, and reaction times observed at a single booth for the corresponding variables in Equation 5 and assume (in agreement with observation) that reaction time and ΔM do not vary greatly among vehicles, the following simplified cycle-time equation is obtained:

$$H' \approx R + \Delta M + \max(R_2 + M_2 + S_2, R_1 + M_1 + S_1) \\ \approx R + \Delta M + \max(H_2, H_1) \quad (6)$$

where H_1 and H_2 are the headways observed at a conventional booth. As noted earlier, not all drivers stop before entering service at a single toll booth, and most have very short reaction times. Thus, R was not measured precisely. Instead, the sum $R + \Delta M$ was approximated by the headway between cars discharging from traffic signals. According to the well-known Greenshields paper (2), this headway is approximately 3.2 s. However, headways at tandem toll booths might be somewhat larger than those at traffic signals because drivers handle money while driving to the booth and because drivers might be delayed by vehicles at the front booth. But reaction times have decreased since Greenshields' paper [due to automatic transmissions (3)], so the net effect makes 3.2 s a reasonable estimate for $R + \Delta M$. Fortunately, $R + \Delta M$ is easily bounded. It must be between ΔM and H_0 ($R + M$). Thus, the next section gives capacity estimates for the following reaction times:

1. Low, $R + \Delta M = \Delta M$;
2. Medium, $R + \Delta M = 3.2$ s; and
3. High, $R + \Delta M = H_0$.

The expectation of H' also depends on the expectation of $\max(H_1, H_2)$. The probability that $\max(H_1, H_2)$ is less than any time t equals the probability both that H_1 is less than t and that H_2 is less than t . Thus, the cumulative distribution function for $\max(H_1, H_2)$ is the product of the distribution functions for H_1 and H_2 (Figure 6). The expectation equals the area above the cumulative distribution for $\max(H_1, H_2)$ between $t = 0$ and $t = \infty$ and is calculated by numerical integration:

$$E[\max(H_1, H_2)] = \int_0^{\infty} 1 - P(H_1 < t)P(H_2 < t)dt \quad (7)$$

The capacity of the tandem toll booth in vehicles per unit time (C') is twice the inverse of the cycle time:

$$C' = 2/E(H') \quad (8)$$

because two vehicles are served in each cycle.

The percentage of increase in capacity with adjacent tandem toll booths is derived from Equations 1 and 8:

$$\text{Percent capacity increase} = 100 \left\{ \frac{2/E(H')}{1/E(H)} - 1 \right\} \\ = 100 \left\{ \frac{2E(H)}{R + \Delta M + E[\max(H_1, H_2)]} - 1 \right\}$$

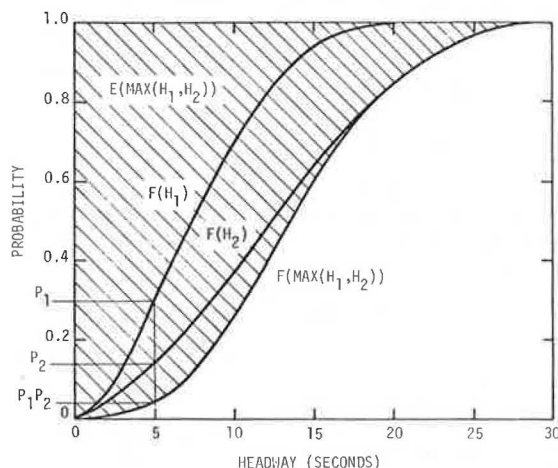
Figure 6. Expectation of $\max(H_1, H_2)$.

Figure 7. Vehicle trajectories: long service time for first n vehicles.

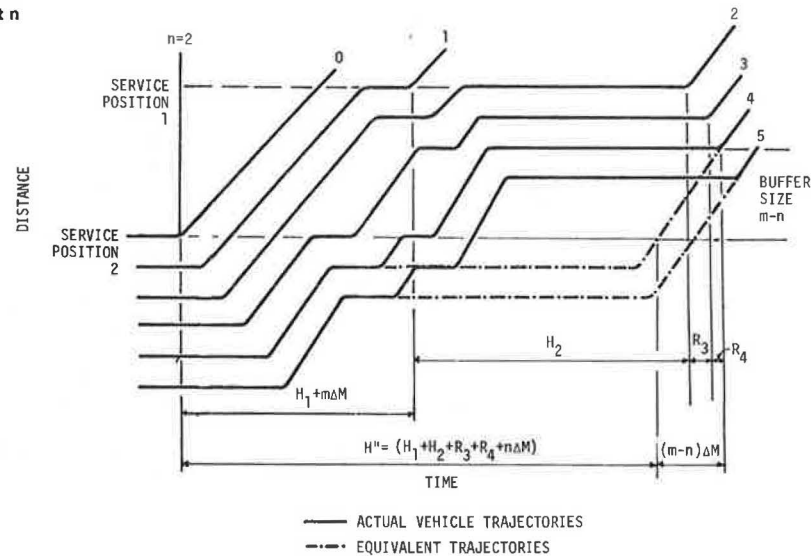
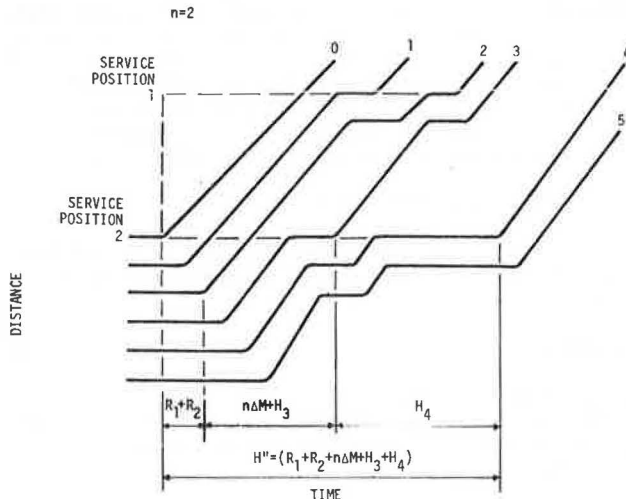


Figure 8. Vehicle trajectories: short service time for first n vehicles.



Tandem toll booths increase capacity if their cycle time is less than twice the headway of conventional booths. Because the reaction time R and move-up time M tend to be small compared with the service time, capacity increase should be substantial. Also note that capacity increases most substantially if headways do not vary greatly among vehicles. Otherwise, $E[\max(H_1, H_2)]$ will be much larger than $E(H)$, and the capacity ratio will be small.

Tandem Toll Booths: Batch Servers

The capacity of tandem toll booths is increased with batch processing. Rather than have the servers collect tolls one vehicle at a time, each server would serve a batch of vehicles in succession. The advantage of this strategy is that random variations in service times are moderated and idle time is reduced.

Suppose that each toll booth processes n vehicles at a time. Then, the last vehicle in a batch served at position 2 would be followed by $2n$ vehicles. The first n would stop at service position 1 and be processed in succession (each leaving as soon as it

pays its toll). The second n vehicles would pay their tolls at service position 2. However, they must queue up behind the vehicles at service position 1 until they all finish paying their tolls. The time-space diagrams in Figures 7 and 8 depict batch tandem booths of size $n = 2$ for two cases: one in which the service time for the first n vehicles is large and the other in which the service time for these vehicles is small. Again, vehicles are numbered 0, 1, 2, ... beginning from the vehicle initially at service position 2.

To ensure that service is not blocked at position 2 by vehicles waiting for service at position 1, the distance between servers should be somewhat greater than n vehicle position lengths. In fact, from standard results of point processes (renewal theorem), blocking would be rare if the number of positions (m) is at least as follows:

$$m \geq n + [2\sigma_H/E(H)](n-1)^{1/2}$$

where σ_H is the standard deviation of H . The ratio $2\sigma_H/E(H)$ is generally close to 1 (see the next section) and should not change much with different tolls. Therefore, the rounded-up value of $n + (n-1)^{1/2}$ gives the minimum number of positions between servers needed to prevent blocking:

$$n \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad \dots \quad 10 \quad \dots$$

$$m \quad 1 \quad 3 \quad 5 \quad 6 \quad 7 \quad 9 \quad \dots \quad 13 \quad \dots$$

Values of m smaller than these could result in occasional blockage, and larger values could result in unnecessary distance between servers.

The cycle time between the moment vehicle 0 leaves service and vehicle $2n$ leaves service (H'') equals the sum of n reaction and move-up times and n headways (Figures 7 and 8):

$$H'' = n(R + \Delta M) + \max \left(\sum_{i=1}^n H_i, \sum_{i=n+1}^{2n} H_i \right) \quad (9)$$

If $\bar{H}_1^{(n)}$ and $\bar{H}_2^{(n)}$ are the average of two sets of n independent headways (which represent the n vehicles at servers 1 and 2), Equation 9 reduces to the following:

$$E(H'') = nE \{ R + \Delta M + \max[\bar{H}_1^{(n)}, \bar{H}_2^{(n)}] \} \quad (10)$$

Note that Equations 9 and 10 do not depend on the spacing between servers. Although increasing spacing increases the travel time between servers, vehicles can leave service at booth 2 earlier (because they can queue up behind vehicles at booth 1). These two factors cancel each other. Capacity (C'') equals $2n$ (the number of vehicles per cycle) divided by the cycle time:

$$C'' = 2n/E(H'') \quad (11)$$

If toll takers are similar, $\bar{H}_1^{(n)}$ and $\bar{H}_2^{(n)}$ will have the same expectation $[E(H)]$ and standard deviation $[\sigma_H/(n)^{1/2}]$. If n is greater than 2 or 3, $\bar{H}_1^{(n)}$ and $\bar{H}_2^{(n)}$ will also be approximately normally distributed. Under these conditions, the expectation of the maximum is (4)

$$E\{\max[\bar{H}_1^{(n)}, \bar{H}_2^{(n)}]\} \approx E(H) + [0.4\sigma_H/(n)^{1/2}] \quad (12)$$

which yields the capacity

$$C'' \approx 2 \{[E(R) + E(\Delta M) + E(H)] + [0.4\sigma_H/(n)^{1/2}]\}^{-1} \quad (13)$$

Even though H is not exactly normally distributed, Equation 12 is fairly accurate, even for small n . Thus, Equation 13 with $n = 1$ approximates C' (adjacent booths) and eliminates the need for numerical integration.

The capacity improvement with batch processing exhibits decreasing marginal returns with n , as illustrated for the following hypothetical values:

$$E(R) + E(\Delta M) = 2.0 \text{ s}, E(H) = 5 \text{ s}, \text{ and } \sigma_H = 2.5 \text{ s}.$$

From Equation 1b we obtain

$$C = 0.200 \text{ vehicle/s (720 vehicles/h),}$$

and from Equation 13 we obtain

$$\begin{aligned} C' &\approx 0.250 \text{ vehicle/s (900 vehicles/h),} \\ C'' &\approx 0.267 \text{ vehicle/s (960 vehicles/h) if } n = 4, \\ C'' &\approx 0.276 \text{ vehicle/s (1000 vehicles/h) if } n = 16, \\ C'' &\approx 0.286 \text{ vehicle/s (1028 vehicles/h) if } n = \infty. \end{aligned}$$

TANDEM TOLL BOOTHS AT GOLDEN GATE BRIDGE

The Golden Gate Bridge toll plaza consists of 13 bidirectional traffic lanes and 12 conventional toll booths. Typically 11 or 12 booths serve southbound (toll-paying) vehicles during the morning commute, whereas just one lane serves northbound traffic. These lanes are reassigned throughout the day to match prevailing traffic patterns.

Following the procedures outlined earlier, data were collected to determine the impact of tandem toll booths on the plaza's vehicle capacity. The morning commute period was chosen for study because the flow of toll-paying vehicles (approximately 7000 vehicles/h) is largest during these hours. As mentioned previously, the bridge collects a split toll; \$2.00 is charged two days a week and \$1.00 five days a week. Data were collected in good weather for both tolls and then used to estimate the change in capacity and delay. Capacity with a \$1.25 toll was also estimated.

Capacity

Vehicles were individually timed as they passed through the toll plaza on \$1.00 days and \$2.00 days. Consecutive departure times (the times when vehicles began to accelerate from the service position) were recorded for approximately 100 vehicles

Table 1. Average and SD of headways.

Item	No. of Observations	\bar{H} (s)	S(H) (s)
\$1.00 Toll Day			
Lane			
A	95	5.47	2.23
B	86	5.26	2.42
C	78	5.81	1.93
D	96	5.56	2.31
Total	355		
Avg		5.52	2.24 ^a
Bridge avg ^b		5.56	
\$2.00 Toll Day			
Lane			
A	118	7.43	3.35
B	116	6.39	3.79
C	113	6.30	3.44
D	108	7.26	4.00
Total	455		
Avg		6.84	3.65 ^a
Bridge avg ^b		6.53	

^a Root mean square.

^b Derived from average traffic volume, 7:30-8:00 a.m.

in each of four toll lanes. Table 1 gives the average and standard deviation of H for the different lanes. The averages closely match the prevailing traffic volumes recorded by the bridge authority during peak periods.

Reaction and move-up times were later recorded by timing vehicles as they approached the toll plaza. The variable ΔM was estimated by timing vehicles as they traversed a measured distance of 25 ft. Only those vehicles that maintained a constant speed over the entire interval were recorded. The average value of ΔM was 2.7 s; the standard deviation was 0.41 s, which translates to a speed of 6.3 mph.

The expectation of H_0 was determined by recording the elapsed time from the moment a vehicle in the service position began to accelerate away from the toll plaza until the moment the outstretched arm of the following driver reached the toll taker. (Due to time constraints, these observations were recorded at the San Francisco--Oakland Bay Bridge. However, the observed value of H_0 should not differ greatly from that at the Golden Gate Bridge since H_0 is not influenced by the toll charged.) The average of H_0 was 4.22 s; the standard deviation was 0.67 s.

The cumulative headway distributions for the four surveyed lanes were paired into all possible combinations and by following the method described earlier, they were multiplied and integrated to estimate the maximum headways (Table 2). The capacity change was then determined by adding $E(\Delta M)$ (optimistic estimate), 3.2 s (medium estimate), or $E(H_0)$ (pessimistic estimate) to the expectation of the maximum and performing necessary calculations (Equations 7 and 8).

As shown in Table 2, the capacity change for adjacent tandem booths does not vary greatly among the lane combinations. The optimistic estimates for \$1.00 days range from a 16 percent increase to a 22.2 percent increase; the average is 18.5 percent. The capacity increase is greater on \$2.00 days because service time is longer relative to reaction and move-up times. The increase ranges from 15.2 to 29.0 percent; the average is 21.1 percent. However, the optimistic estimate differs considerably from the pessimistic estimate: 18.5 percent versus 1.8 percent on \$1.00 days, 21.5 percent versus 7.0 percent on \$2.00 days. Although it is impossible to

predict exactly how drivers would behave with tandem toll booths, the actual capacity would likely be close to the medium estimate. This assertion is based on the short reaction times of commuters at the bridge toll plaza.

The capacity estimates for batch tandem toll booths are given in Table 3. As discussed earlier, capacity increases with batch size, although the amount of increase decreases with batch size. As batch size becomes large, lane capacity increases by as much as 10-20 percent over adjacent booths. Even with a batch of size 5, the additional increase falls in the range of 7-13 percent. According to the pessimistic prediction, batch tandem toll booths of size $n = 5$ increase capacity by at least 8.8 percent on \$1.00 days and 17.0 percent on \$2.00 days over conventional booths. More realistically, the increase on \$1.00 days should be more than 20 percent and on \$2.00 days close to 30 percent.

Although the \$1.25 toll was discontinued prior to this study, the effect of tandem toll booths on capacity can be inferred. Traffic flow per toll lane was approximately 9.5 vehicles/min with the \$1.25 toll, which is nearly the same as the 9.2-vehicle/min rate with the \$2.00 toll. Thus, it is likely that a similar number of drivers needed change and that the headway distributions were comparable. Therefore, the percentage capacity increase with tandem booths would likely be similar to that of the \$2.00 toll; it would fall in the range of 7-21 percent for adjacent booths and most likely be more than 15 percent. The capacity increase with a batch of size 5 would fall in the range of 17-24 percent; most likely it would be close to 30 percent.

Delay

A precise estimate could not be obtained for the distribution of vehicle arrival times within the constraints of this study. Therefore, attention was given to the capability of tandem toll booths to handle the maximum traffic volume entering the plaza. During one morning commute period, vehicles were counted at the north end of the bridge 2.2 min from the plaza in uninterrupted flow conditions. The maximum flow rate was slightly less than 7800 vehicles/h, which is close to saturation for the Golden Gate Bridge roadway (four lanes with no shoulder), and was sustained for about 15 min. One can deduce that the arrival rate of vehicles at the plaza would never greatly exceed 7800 vehicles/h over any reasonably long time interval (1 min or greater). Thus, a toll plaza capacity in excess of

7800 vehicles/h would ensure that transient queues and delays would always be short (30 s or less).

Even during the busiest periods, the average arrival rate is considerably less than 7800 vehicles/h. Toll plaza queues on \$1.00 days tend to be small or nonexistent. Therefore, vehicle counts recorded at the plaza on these days are representative of both arrivals and departures. The average 15-min vehicle count on \$1.00 days between 7:30 and 8:00 a.m. on 97 workdays in 1981 was 1782 vehicles, or a rate of 7127 vehicles/h. If the capacity of the toll plaza were approximately 7200 vehicles/h, transient queues would be common but they would not become increasingly long. Even if the maximum flow rate were sustained for 30 min (which is unlikely), delay would reach just 2.5 min. Average delay would likely be a minute or less.

Table 2. Effect of adjacent tandem booths on lane capacity.

Item	E[$\max(\bar{H}_1, \bar{H}_2)$] (s)	C' [vehicles/(lane-min)]		
		High Estimate	Medium Estimate	Low Estimate
\$1.00 Toll Day				
Lane combination				
A, A	6.57	12.94	12.28	11.12
A, B	6.46	13.09	12.42	11.23
A, C	6.70	12.77	12.12	10.99
A, D	6.66	12.82	12.17	11.03
B, B	6.33	13.29	12.59	11.38
B, C	6.62	12.88	12.22	11.07
B, D	6.55	13.02	12.31	11.14
C, C	6.80	12.63	12.00	10.89
C, D	6.78	12.65	12.02	10.91
D, D	6.74	12.71	12.07	10.95
Avg	6.62	12.88	12.22	11.07
Single toll booth			10.87	
\$2.00 Toll Day				
Lane combination				
A, A	9.09	10.18	9.76	9.02
A, B	8.66	10.56	10.12	9.31
A, C	8.56	10.66	10.20	9.40
A, D	9.15	10.12	9.72	8.97
B, B	8.05	11.16	10.67	9.78
B, C	7.96	11.31	10.75	9.85
B, D	8.64	10.59	10.14	9.34
C, C	7.87	11.35	10.84	9.92
C, D	8.56	10.66	10.20	9.39
D, D	9.17	10.10	9.70	8.95
Avg	8.57	10.66	10.21	9.39
Single toll booth			8.77	

Table 3. Effect of batch tandem toll booths on lane capacity.

Toll (\$)	Conventional Booth [vehicles/(lane-min)]	n = 1		n = 2 ^a		n = 5 ^a		n = 10 ^a		n = ∞	
		Vehicles per Lane per Minute	Increase (%)	Vehicles per Lane per Minute	Increase (%)	Vehicles per Lane per Minute	Increase (%)	Vehicles per Lane per Minute	Increase (%)	Vehicles per Lane per Minute	Increase (%)
High Estimate ^b											
1.00	10.87	12.88	18.5	13.52	24.4	13.89	27.1	14.08	29.5	14.58	34.1
2.00	8.77	10.66	21.5	11.36	29.5	11.79	34.4	12.01	36.9	12.60	43.7
Medium Estimate ^b											
1.00	10.87	12.22	12.40	12.83	18.0	13.17	21.0	13.33	22.6	13.76	26.6
2.00	8.77	10.21	16.4	10.84	23.6	11.22	28.0	11.43	30.3	11.95	36.3
Low Estimate ^b											
1.00	10.87	11.07	1.8	11.57	6.4	11.83	8.8	11.97	10.1	12.32	13.3
2.00	8.77	9.39	7.0	9.93	13.2	10.26	17.0	10.42	18.8	10.86	23.8

^aClark approximation (4).

^bVehicles per lane per minute averaged over all lane combinations.

Table 4. Vehicle capacity of toll plaza.

Capacity (vehicles/h)		Adjacent Tandem Booths			Batch Tandem (n = 2)			Batch Tandem (n = 5)		
Toll (\$)	Current ^a	High	Medium	Low	High	Medium	Low	High	Medium	Low
1.00	7127	8116	7790	7223	8431	8089	7469	8576	8250	7597
2.00	6616	7683	7430	6963	8080	7787	7271	8323	8005	7460
1.25	6824	7924	7663	7187	8274	8032	7499	8584	8257	7694

^aCapacity estimate based on average traffic flow from 7:30 to 8:00 a.m. on 97 \$1.00 days, 86 \$1.25 days, and 14 \$2.00 days in 1981.

The layout of the toll plaza prevents installing tandem toll booths on all traffic lanes, particularly the far-right lane (which serves many trucks and buses) and the two far-left lanes (which already operate below capacity when other lanes have long queues). Tandem toll booths could possibly be installed on the other nine lanes.

Table 4 gives the current capacity of the toll plaza under the \$1.00, \$1.25, and \$2.00 tolls and capacity estimates for adjacent and batch tandem toll booths. These capacities are based on the average traffic flow between 7:30 and 8:00 a.m. on workdays, a period when the plaza is usually congested. This estimate is slightly conservative since the arrival rate sometimes falls below capacity (such as when there is an accident on the bridge). The tandem estimates account for nine tandem and three conventional booths and are based on the percentages given in Table 3. Furthermore, it was assumed that all lanes carry equal amounts of traffic (the three lanes without tandem booths actually carry slightly less than average, so this underestimates tandem capacity).

Tandem capacity on \$1.25 days is between 7187 and 7924 vehicles/h. Thus, even under the most pessimistic estimate, capacity would still exceed the average traffic volume and queues would never become excessively long. Under the optimistic estimate, capacity would exceed the maximum arrival rate at the toll plaza (7800 vehicles/h), and queues would never become more than a few vehicles long. The medium estimate yields a capacity of more than 7600 vehicles/h. Therefore, we conclude that nine tandem booths would accommodate all traffic during the morning commute without significant delay.

The capacity estimates for \$2.00 days are also encouraging. It is very likely that adjacent tandem booths would handle the average traffic volume entering the toll plaza (see medium estimate). However, the toll plaza would not handle the maximum arrival rate. Also, the pessimistic capacity estimate falls below the average traffic volume recorded on \$1.00 days. The capacity of batch tandem toll booths (size $n = 2$) does exceed the average traffic volume and would likely be very close to the maximum arrival rate. Because part of the delay incurred on \$2.00 days results from confusion regarding the split toll, capacity would surely be greater than these estimates if a \$2.00 toll were charged every day of the week and likely be close to that for \$1.25 days.

DISCUSSION

It was not feasible to collect arrival and service time data for all time periods, but it is still possible to make a good estimate of the impact of tandem toll booths throughout the day and week. Of particular interest are Sunday afternoons, when many vacationers return to San Francisco, and Friday evenings, when many Marin residents travel into the

city. Queues have been particularly long at these times because fewer lanes are allocated to the toll-paying direction (the northbound traffic is greater than in the morning) and because service times are considerably longer than during the morning commute (drivers are less familiar with the toll-taking system). This longer service time makes tandem booths more desirable. Thus, the capacity should increase by a greater percentage than the estimates provided above and would likely be 20 percent or greater on \$2.00 days for adjacent booths.

The estimates provided above apply specifically to days when the weather is fair. Tandem toll booths may not be as effective on rainy or very foggy days. Drivers would likely be more cautious moving into the service position. Thus, the move-up time and reaction time should be larger relative to the service time, and the capacity increase should be less than that predicted earlier. However, the capacity on the bridge itself may be sufficiently small on poor weather days that queues never appear at the toll plaza anyway.

The capacity of tandem booths is most accurately estimated with a simple experiment. An additional toll taker could temporarily collect tolls in tandem behind each of the existing toll booths. Car counts could then be compared with the current averages to obtain an accurate estimate of the capacity change. The experiment should be repeated over several time periods so that capacity throughout the day and week can be determined.

CONCLUSIONS

It is clear from Equation 12 that the capacity of tandem toll booths is greatest when service times have a small variance and the move-up and reaction times are small relative to the service time. Although the Golden Gate Bridge is an ideal location for tandem booths from the aspect of small service time variance, its small service time average keeps the capacity increase relatively low. Because traffic only slightly exceeds capacity, this increase would eliminate most of the existing delay at the toll plaza. A capacity increase of about 15 percent is expected for adjacent booths and about 25 percent for batch booths of size 5. In both cases the actual capacity is influenced by the size of the toll, familiarity of drivers with the toll-taking system, and weather conditions.

Although tandem toll booths use personnel less efficiently, they do increase land use efficiency; they offer increased capacity without increased plaza width. Since the capacity of the toll plaza at the Golden Gate Bridge is insufficient during only a few hours a day, the labor cost would be small compared with the cost of constructing new lanes. Tandem toll booths would cost \$1.5 million to install and \$150 000 per year to staff, which appears attractive compared with the \$16 million cost of installing a new toll plaza (1).

ACKNOWLEDGMENT

We would like to thank Ron Gordon for suggesting the idea of tandem toll booths and for funding data collection. The Golden Gate Bridge Authority is thanked for providing data and for their overall cooperation. The theoretical aspects of this research were supported by the National Science Foundation.

REFERENCES

1. R.B. Wong. Golden Gate Bridge Toll Plaza Master Plan Study. Ammann and Whitney, New York, 1982.

2. B.C. Greenshields, D. Shapiro, and E.L. Ericksen. Traffic Performance at Urban Intersections. Bureau of Highway Traffic, Yale Univ., New Haven, CT, Tech. Rept. 1, 1947.
3. W. Kunzman. Another Look at Signalized Intersection Capacity. ITE Journal, Vol. 48, No. 8, Aug. 1978, pp. 12-15.
4. C.E. Clark. The Greatest of a Finite Set of Variables. Operations Research, Vol. 9, 1961, pp. 145-162.

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

Abridgment

Tandem Toll Collection Systems

LOUIS D. RUBENSTEIN

By using two or more collection stations in the same traffic lane, tandem toll or parking-fee collection increases lane capacity and reduces the need for plaza widening. Data are presented relating processing rates to toll fee; e.g., the rate for a \$1.25 fee is 30 percent slower than that for a \$1.00 fee. Toll agencies that have implemented \$1.25 tolls have encountered extreme congestion, especially with the weekend recreational traveler. Several operational configurations of tandem tolls are described. A coordination device is described to automate the control of motorist traffic signal and payment signal to distinguish between axle registrations of successive vehicles, even under dense conditions. Slow collection devices such as paper-money acceptors or flexible-ticket readers that are impractical at a conventional active lane are feasible in tandem. The expected capacity increase depends on the conventional cycle time, its standard deviation, and the distance between the toll stations. When the distance is several vehicle lengths, the stations are buffered, which results in better performance and independence of capacity increase on cycle time variance. The slower the existing collection time, the greater the capacity increase, e.g., 6 s/vehicle yields a 34 percent increase, 20 s/vehicle, 1.75 percent increase, when buffered.

There is a growing need for measures such as tandem toll booths to rapidly increase the traffic capacity of existing toll plazas. The experience at the Golden Gate Bridge and the Triborough Bridge and Tunnel Authority with long queues when toll rates were raised to more than \$1 can be expected to be repeated at other tollway facilities. The high inflation rate of the last several years, one-way toll collection, and the use of toll surpluses to subsidize mass transit operations are pushing many toll fees to above the \$1 level.

Stop-watch surveys that I have conducted indicate the relative effect of the toll fee on the vehicle-processing rate; they are summarized in Table 1.

Many existing toll plazas were designed when traffic volumes were lower and vehicle-processing rates higher and are not equipped to accommodate fees of more than \$1. As toll fees rise, the problem will become more widespread.

This approaching problem will remain for years. Efforts by the U.S. Treasury Department to popularize the use of a \$1 coin have not been successful. Similarly, efforts by toll operators to promote use of high-value tokens have met public resistance and are not very effective with the weekend recreational traveler. Busy motorists are not willing to accept the inconvenience and advance payment requirements of token prepurchase without a substantial dis-

count. Even a 10 percent discount for tokens will reduce the revenue of many facilities a greater amount than the total cost of the existing toll-collection system. Token discounts also increase opportunities for employee fraud.

New technologies such as automatic vehicle identification had offered potential for speeding toll processing, but after years of development they have still not overcome the operational, cost, and privacy obstacles to their widespread implementation. Toll-collection computerization programs have been directed at improved auditing capabilities and not improved traffic flow.

The patronage of toll booths in the outer roadway lanes is lower than that in the central lanes, even under congested conditions. The approach to a toll plaza must be widened gradually over a long distance, which increases construction and maintenance costs, particularly on elevated plazas. If there are heavy weaving movements due to the location of particular entrance/exit ramps, even long tapers may not be effective. Tandem lanes would also lessen air-pollution levels in the toll plaza.

Table 1. Effect of toll fee on processing rate.

Passenger-Car Fee (\$)	Bridge Surveyed	Manual Lane-Processing Rate ^a (s/car)	
		Sample Avg	Best Avg
1.20	San Diego-Coronado, California	10.0	9.1
1.25	Throgs Neck, New York City	9.8	9.0
2.00	Golden Gate, San Francisco, California	8.8	6.5
0.75	San Francisco-Oakland Bay, San Francisco, California	6.9	6.6
1.00	Golden Gate	6.4	6.1
0.40	Carniquez, I-80, Vallejo, California	6.3	5.9
0.25	Vincent Thomas, Long Beach, California	5.9	5.5

^aObservations are based on 120 observations per bridge, under moderate traffic; survey was conducted in spring 1982 during hours when commuter carpool free-passage rates were not in effect. Plaza grades were zero to slight. Best averages exclude patrons with exceedingly long service times, apparently unrelated to the toll fee.

Table 2. Effect of processing rate on increase in capacity of tandem tolls.

Single-Station Toll Cycle Time (s/vehicle)	Application and Typical Fee	Increase in Capacity (%)	
		Two Toll Stations	Three Toll Stations
5.5	Single-coin toll, car	34	49
6	\$1 toll, car	36	55
9.5	\$1.25 toll, car; distance-related turnpike toll with ticket	54	89
14	Tractor-trailer	47	75
20	Time-related car parking fee with ticket	75	134

Estimates for an increased traffic capacity of more than 40 percent for tandem lanes are described in Table 2. These estimates are supported by actual experience. The New York State Thruway used tandem toll operations at major bottleneck toll plazas on several peak days per year for more than 10 years until plaza widenings were completed. In spite of the irregular use and makeshift, nonautomated coordination and accounting of the system, the Thruway Authority reports that tandem toll operations resulted in capacity increases of 25 percent. Los Angeles International Airport has used manually operated tandem parking-fee collection on peak days since October 1982; according to Bill Barnett, vice president of Parking Concepts, Inc., capacity increases of more than 50 percent have been reported. A demonstration of an automated toll station in tandem is planned on the Bronx Whitestone Bridge in the summer of 1983.

A recent study (1, p. 17) developed a model that describes the capacity increase available by tandem toll collection. The model indicates the importance of designing and operating the stations in a tandem toll lane so that delays in one booth are not communicated to the second booth.

A 40, 25, or even 15 percent increase in toll-lane capacity can have a tremendous impact on toll plaza queues. Application of standard queueing-analysis formulas (2, p. 364) indicates that in the situation where the vehicle arrival rate is 95 percent of the vehicle service rate, a 15 percent increase in the service rate will reduce the average queue length from 18 vehicles to 4 vehicles. At toll plazas with high traffic stream divergence angles, queue lengths as short as 9 vehicles can lead to blockage of access to the outer lanes and an unstable flow condition.

Early field tests with tandem tolls experienced difficulty with coordination of the consecutive toll booth activity and revenue accounting. Designs have been developed (U.S. Patent 3 686 627, August 1972) that electronically coordinate the traffic flow and revenue accounting and automatically adjust to patrons paying at the wrong booth. Auditing computers installed in modern toll plazas can implement these designs at minimal expense.

Tandem toll booths are most effective when used with automatic processing equipment or during peak hours or on peak days to provide supplemental capacity. By proper scheduling of operating shifts or use of manual and automated lanes in tandem, increases in operator costs can be minimized.

OPERATIONS GUIDELINES

Sequencing and Layout

To derive the full potential of tandem tolls, they

must be applied in appropriate locations and in an effective manner. Several operations guidelines that can be identified at this time are described.

A typical two-direction toll plaza is shown in Figure 1. It is characterized by a tapered approach roadway. Such conventional toll plazas have a single line of toll booths crossing the roadway. Tandem tolls permit the central lanes to carry a higher volume, and less traffic needs to be processed in the outer lanes. Figure 1 shows a set of tandem toll booths installed in the central section of the plaza. The lengthened toll island and the second treadle indicate a tandem lane in the diagram. The islands provide a barrier that restrains each vehicle in its lane and also screens visual disturbance that can slow operations.

Any equipment used in a conventional toll lane could be used in a tandem lane. Tandem toll lanes make feasible the use of some equipment that might otherwise be too slow to install in an active lane, e.g., a \$1 bill changer, stored-ride magnetic card reader, or at a parking lot exit a time-related ticket reader and fee collector. Several additional units of equipment are required in a tandem lane. Typical signs to instruct the motorist at which booth to pay are shown in Figure 2. In addition, a device to coordinate the activities of the two toll booths is desirable. Although less effective, manually operated and coordinated systems could implement tandem tolls on a temporary basis.

The coordination device must identify each vehicle as it travels through the toll lane and know whether it has or has not paid its toll. This is accomplished by keeping a tally of the differences in axle counts between treadles A1 and A2 in Figure 2. The sequence of axle counts and payment registrations detected at each station can be the basis of a system to automatically count and check the number of axles of each vehicle.

The device could be set to collect a toll from every other vehicle at each station or could use a different collection configuration. If a patron mistakenly pays at the upstream booth, the device memory will cause a proceed signal to occur when that vehicle reaches the downstream booth.

Alternative operational configurations for a tandem toll lane such as batch, alternate vehicles, synchronized stations, automatic-manual stations, three collection stations, and automatic truck classification are discussed elsewhere (3).

Plaza Width and Taper

Disadvantages of wide toll plazas that can be reduced by tandem tolls include increased right-of-way costs, relative lower activity of outer toll lanes, and unstable flow conditions caused when access to the outer toll lanes is blocked by queues.

Early toll plaza designers recognized the importance of minimizing toll plaza widths and employed techniques such as driver- and passenger-side toll-collection booths to increase the number of toll lanes in a given plaza width (4). With the increased popularity of bucket seats and as a result of operations research studies that noted driver reluctance to use the passenger-side motorist collection booths and their slower processing times, their use in toll plazas was discontinued (5).

The only quantitative evaluation of this driver reluctance available in the open literature pertains to the San Francisco-Oakland Bay Bridge (6, p. 137). An approximate 2.5 percent drop in traffic volume is reported to occur for each 1 percent increase in divergence.

To overcome this effect and to reduce accidents, several references recommend long toll plaza

Figure 1. Two-direction toll plaza with tandem tolls in central lane.

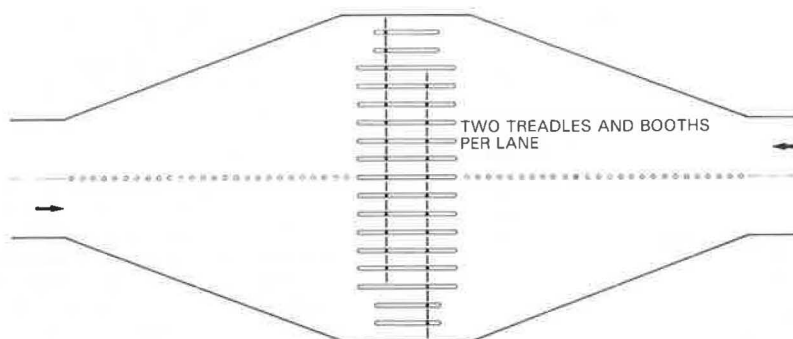
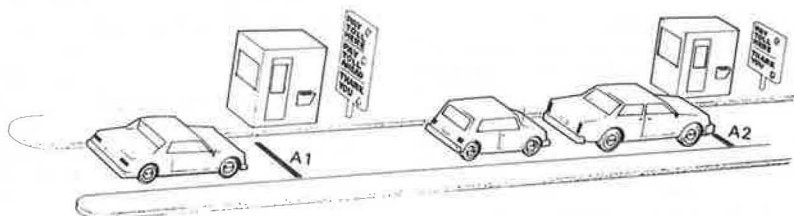


Figure 2. Motorist instruction signs in tandem lane.



tapers. Citing New York State Thruway standards, a report by the New York State Department of Transportation (7, pp. IV-4, IV-5) recommends a desirable taper of 20 to 1 and a minimum taper of 8 to 1. Below the latter value, it is stated that merging and weaving accidents will increase rapidly. Long tapers can be expensive in urban areas. A European study recommends a taper rate of 10 to 1 (8, p. 189). In addition, if the plaza taper is insufficient, turbulence caused by weaving and merging can cause the plaza to be the minimum-capacity section on the roadway. Several major bridges and tunnels have been built with inadequate tapers due to space constraints (9, p. 255). Inadequate tapers can also develop when toll plazas are widened to accommodate increased traffic.

The required taper rate can be estimated on the basis of its ability to prevent queues extending from the toll barrier from cutting access to the outermost toll lanes. The taper rate will depend on the expected queue length (QL) in front of the toll booth.

QL is given by (2)

$$QL = [\rho^2 / (1 - \rho)] (\text{number of vehicles}) \quad (1)$$

$$\rho = (\text{toll lane arrival rate}) / (\text{toll lane service rate}) < 1 \quad (2)$$

For example, if service rate is 1/(6 s/vehicle) and arrival rate is 1/(8 s/vehicle), ρ is 0.75.

For a vehicle of width CW to move from behind the queue to the adjacent outermost lane, a lateral distance along the roadway of $TPR \cdot CW$ is required, in which TPR is the taper rate.

If LW is the lane width, then by definition of taper rate the minimum lateral approach roadway is $TPR \cdot LW$. The queue length plus the minimum car lane-change length must be less than the approach roadway length.

$$QL \cdot VS + TPR \cdot CW < TPR \cdot LW \quad (3)$$

$$TPR > (QL \cdot VS) / (LW - CW) \quad (4)$$

By using $VS = 25$ ft, $LW = 15$ ft, and $CW = 6$ ft, $TPR =$

$2.77QL$. Corresponding values of ρ , QL, and TPR are indicated below.

ρ	QL	Taper Rate
0.95	18	50/1
0.90	8	22/1
0.85	4.8	13/1
0.80	3.2	9/1

The above shows the sensitivity of the plaza design guidelines to the toll processing rate. Most toll plazas during peak hours will have queues of at least five vehicles.

INCREASE IN CAPACITY OF TANDEM TOLLS

If the tandem toll-collection stations are separated by a buffer distance of several vehicle lengths, the capacity increase is (3)

$$ICAP = 2 [TC / (TC + TM)] 100 \quad (5)$$

where TC is the cycle time per vehicle in a nontandem system (e.g., 9 s/vehicle corresponds to a capacity of $3600/9 = 400$ vehicles/h) and TM is the move-up time, or time between when the paying preceding vehicle leaves the toll station and when the nonpaying following vehicle reaches the same point at the toll station.

High, intermediate, and low estimates of TM are 3.2, 2.8, and 2.4 s/vehicle. These estimates are based on analogies to velocity profiles of vehicles entering a signalized intersection after the aspect changes from red to green and from a study of the on-time duration of the braking lights of vehicles as they stop at a toll booth after approaching through a short queue (3).

If the collection stations were not buffered, the increase in capacity can be estimated by (2, p. 364)

$$ICAP = 2 \{ TC / [TC + TM + 0.4SD(TC)/\sqrt{N}] \} \quad (6)$$

where $SD(TC)$ is the standard deviation of the cycle times and N is the number of vehicles per batch if a batch-processing scheme is used.

Hall and Daganzo (1) present data collected at

the Golden Gate Bridge and use the previous equation to calculate a capacity increase of $ICAP = 18$ percent for the \$1.00 toll by using $TM = 2.7$ s and assuming that the booths are not buffered. When the booths are buffered, $ICAP = 34$ percent.

The result is consistent with a test of tandem tolls conducted at the Golden Gate Bridge in 1969. By using a makeshift arrangement where the second toll collector stood out in front of the islands, the flow rate was increased from 625 to 725 vehicles/h (16 percent) (2, p. 364).

Tandem tolls could also be used in a truck toll lane. Cycle component times for a tractor-trailer truck are $TC = 14$ s and $TM = 7.5$ s (8, p. 189).

The tandem move-up in time can be estimated as $TM = 5.0$ s.

The effectiveness of tandem tolls increases as the toll-collection cycle time increases. The previous equations were applied to derive Table 2.

APPLICATION TO REDUCE NEED FOR PLAZA WIDENING

I have presented an example that illustrates one of the situations in which a tandem toll system would be more economical than additional conventional toll lanes for increasing a toll plaza's capacity on weekends (3).

The cost parameters in the example are as follows:

1. Capital cost per additional booth:

Item	Cost (\$000s)
Toll booth	40
Toll registry equipment	30
Tapered approach road (1500 ft)	1500
	1640

2. Present worth of staffing: half-day/week, \$60 000.

By using these parameters, the capacity increase per unit of cost is

1. Tandem: 1.6 cars/(h*\$1000) and
2. Conventional: 0.6 car/(h*\$1000).

ACKNOWLEDGMENT

This report was supported by individual contribu-

tions of time, services, and a small amount of funds. No government or institutional funding was used. Several persons made noteworthy contributions: Linda Rubenstein for her encouragement and assistance in data collection; Ron Gordon, Redwood City, California, for supplying his collected information on tandem tolls; Nicolas Bellizi, Urbitrans Associates, for his observations and special surveys; and Mel Kohn of Parsons Brinkerhoff, for his review of the early drafts of this report.

REFERENCES

1. R. Hall and C. Daganzo. Tandem Toll Booths for the Golden Gate Bridge. Institute of Transportation Studies, Univ. of California, Berkeley, Res. Rept. UCB-ITS-RR-82-7, 1982.
2. M. Wohl and B. Martin. Traffic Systems Analysis for Engineers and Planners. McGraw-Hill, New York, 1967.
3. L. Rubenstein. Tandem Toll Collection Systems. Aug. 1982 (available from the author).
4. J.F. Curtin. Bridge and Tunnel Approaches. Proceedings of the American Society of Civil Engineers, 1940, p. 1821.
5. L.C. Edie. Traffic Delays at Toll Booths. Journal of Operations Research Society of America, Vol. 2, No. 2, May 1954, p. 121.
6. H.C. Wood and C.S. Hamilton. Design of Toll Plazas from the Operator's Viewpoint. HRB, Highway Research Board Proc., Vol. 34, 1955, pp. 127-139.
7. Final Report: Engineering, Environmental, and Socioeconomic Reevaluation of State Implementation Plan Strategy B-7: Tolls on East and Harlem River Bridges. New York State Department of Transportation, Albany, April 1977.
8. G. Roemer. The Design and Operation of Toll Booths. Oesterreichische Ingenieur Zeitschrift, Vol. 13, 1978, p. 189.
9. D.R. Culverwell, Freeman, Fox and Partners. Some Aspects of Toll Systems for Major Bridges and Tunnels. Planning and Transportation Research and Computation Co., London, England, 1974.

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

Reliability of Classified Traffic Count Data

PETER DAVIES AND DAVID R. SALTER

The reliability of classified traffic count data collected for the planning and operation of highway systems is examined. Manual classified count data are subject to serious errors, whereas automatic vehicle classification with modern microprocessor technology may have other accuracy problems. Accuracy checks carried out in the United Kingdom are described for two automatic classification systems—for simple classification by using inductive loops alone and for detailed classification by using loops and axle detectors in combination. An evaluation of automatic classification equipment, including these simple and detailed systems, has been carried out in the United States by the Maine Department of Transportation. The results of these studies are described. The accuracy of simple vehicle classification based on vehicle length alone is limited by the fundamental properties of inductive-loop sensors. However, at sites with good lane discipline, the accuracy of classification is likely to be sufficient for most routine purposes such as the measurement of passenger-car-equivalent flows. Tests in the United States have shown that the reduced reliability of pneumatic-tube sensors leads to poor classification accuracy when these sensors

alone are used for vehicle detection. More detailed vehicle classification methods can give greater accuracy, in excess of 90 percent, but as traffic conditions deteriorate, accuracies reduce. In the detailed classification method, there are difficulties in discriminating between certain cars, vans, and trucks, particularly where lane discipline is poor. Further developments of automatic classification techniques are currently in progress, and improvements are anticipated under urban traffic conditions and in the portability of detailed classification equipment. However, simple classified counters are already available and already have a part to play in displacing unreliable manual counts. Future trends in labor and microprocessor costs are anticipated to be such that as new developments become available, their rapid exploitation will become increasingly attractive.

Classified traffic counts have been carried out for decades to provide basic information used in the de-

sign, maintenance, and management of highway systems. In the past, manual counts have been the only source of classified flow data; automatic counts have been limited to the recording of axle pairs or total numbers of vehicles. More recently, the microprocessor revolution has changed this state of affairs, so that automatic vehicle classification and monitoring is now a practical proposition in many situations.

The demands for classified traffic count data are various. Simple classification into some five or six broad categories of vehicle is a common requirement for highway design or traffic signal-timing procedures based on passenger-car equivalents (PCEs). Longer-term monitoring of classified traffic flows provides the basis for forecasts of future traffic, disaggregated by type of vehicle. The economic appraisal of highway schemes may also require a knowledge of the mix of vehicle classes and their characteristically different operating costs, occupancies, and values of time.

A more detailed vehicle classification could also have a part to play in some areas of growing concern. Axle-weight distributions of different types of truck can be monitored at weighbridge sites and the data applied to pavement design or maintenance at other locations through detailed classified counts. The allocation of road damage costs to different classes of vehicle on toll highways or via general vehicle taxation again requires the detailed classification of freight vehicles. Other forms of classification, such as speed category, headway, and lane or turning movement, may also play important roles in special situations.

Relatively little is known about the sensitivity of design procedures or traffic control measures to errors in the classified count data. One study does suggest that highway scheme cost-benefit appraisal can be highly sensitive to the mix of vehicle classes assumed (1). Traffic forecasting could also be very sensitive, based as it is on the extrapolation of past trends from an assumed current situation; any errors in the base data may well be magnified in forecasts of the future. Finally, pavement design, with its high-order power-law relationship between axle weight and road damage, could eventually prove to be most sensitive of all to the basic traffic data input.

In this paper, we consider the reliability of classified traffic count data produced by manual and automatic means. Recent work on the accuracy of manual classified counts suggests that even for closely supervised, well-conducted surveys, results are much less reliable than might commonly be supposed. Automatic classification offers opportunities to overcome some accuracy problems but instead can lead to errors of a different nature than those resulting from manual enumeration.

We begin by reviewing available evidence on the reliability of manual classified traffic counts. Next, two automatic classification systems, for which results are presented in a number of accuracy studies, are described. Road-sensor design and software are two key areas in automatic classification, so the scope for their improvement is considered. Finally, the relative merits of automatic and manual classification are assessed in the context of the current state of the art.

MANUAL CLASSIFIED TRAFFIC COUNTS

At first sight, manual classified traffic counting appears a straightforward task. Passing vehicles are recorded for predetermined time periods either by marking different sections of survey forms or by hand-operated counters in order to build up a pic-

ture of the traffic flow disaggregated by vehicle class. In practice, however, this simple procedure gives rise to considerable scope for error. Vehicles can be missed, double counted, wrongly identified, or entered in the wrong place. There are many reasons why these mistakes occur. For example, enumerators are locally recruited temporary staff whose motivation and skill may vary considerably. Counting can be a tedious process; it requires extended concentration, which may easily be broken. The importance of the data may be far from obvious to temporary staff, so the apparently harmless invention of results may prove a strong temptation. Close supervision or performance checks are difficult and time-consuming, so sanctions against carelessness are rare and rewards for vigilance are generally nonexistent.

Even where counts are properly conducted and well supervised, there is evidence to suggest that results can be unreliable. The U.K. Department of Transport compared simultaneous counts by a dedicated full-time team with those of teams locally recruited for routine census work at three sites (2). Although no consistent biases emerged, there were considerable variations in both absolute totals and percentage of discrepancies. It was concluded that errors were apparently serious, both in absolute and percentage terms.

Further comparisons are described in an internal note of the U.K. Department of Transport, the results of which are summarized elsewhere (3). The results suggest that 95 percent confidence limits on 16-h total flows are probably within ± 10 percent but that considerably greater intervals apply to most individual vehicle classes:

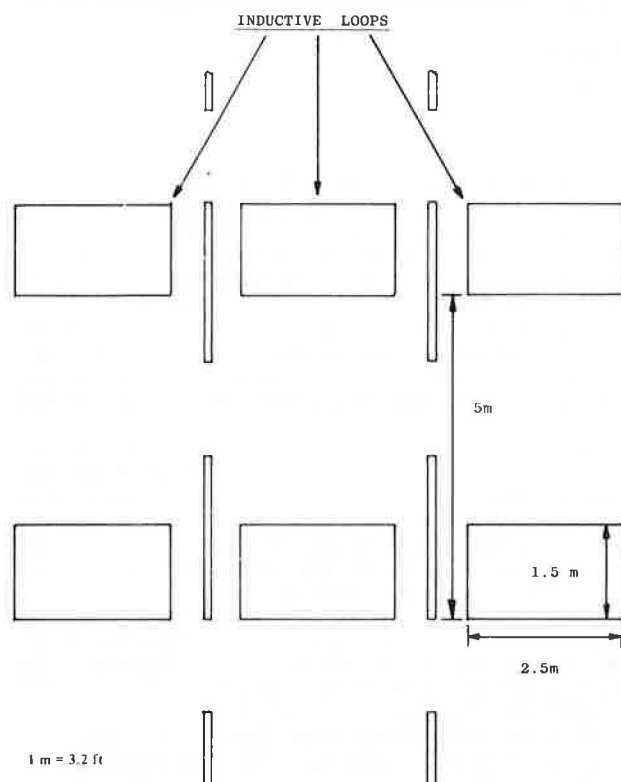
Vehicle Class	95 Percent Confidence
	Limit (%)
Two-wheeled motor vehicles	± 35
Cars and taxis	± 10
Buses	± 37
Light trucks	± 24
Other trucks	± 28

There was some evidence of difficulty in distinguishing "light" from "other" trucks, despite special markings carried by U.K. trucks; the interval fell to ± 18 percent for all freight vehicles combined. However, the greatest percentage of errors is seen in two distinctive categories--motorcycles and buses.

In view of these major discrepancies in 16-h counts, how good will peak-period data prove to be? The answer must be almost certainly that they will be worse, since there is less scope for compensating errors within shorter-duration counts. Moreover, at peak periods, enumerators will be fully stretched, and more vehicles may be missed or guessed. Our test with simultaneous film recording of traffic flows shows that even highly motivated research observers find it impossible to count with high accuracy and are often unaware that they have made mistakes. In less-controlled surroundings the problems are likely to be still more severe.

The quality of manual classified traffic counts can clearly give cause for concern. When coupled with the errors of sampling, scaling, and forecasting, the errors of manual enumeration may well be sufficient to produce suboptimal design or management decisions based on wrong information. Whether anything can be done about this at reasonable cost, for example, through the greater use of microprocessor-based automatic vehicle classification, is another question. This question is considered in the remainder of this paper.

Figure 1. Typical sensor configuration for simple classification.



SIMPLE AUTOMATIC CLASSIFICATION

For many routine purposes such as the determination of PCE traffic flows, simple classification into a small number of vehicle categories may well be sufficient. One type of automatic vehicle classifier already available counts vehicles into separate bins according to their overall lengths. Tests are described on the accuracy of such a system, a 12-bin speed or 4-bin length classifier manufactured by the Golden River Corporation.

When configured as a length classifier, this portable microprocessor-based system records classified traffic flows disaggregated into length categories specified by the user. The system's main component consists of a roadside processing unit sealed with its rechargeable battery pack into a cast aluminum case and linked to inductive-loop sensors in the road. A complementary retriever unit is connected to the roadside processor for the initial configuration of the system and for the recovery of data at intervals as required.

The road sensors consist of up to three pairs of matched inductive loops; each pair of loops is located in a single traffic lane (Figure 1). A wide range of loop dimensions will be accepted by the equipment, but typical loops would be 1.5 m long by 2.5 m wide (4 ft 11 in by 8 ft 2 in) spaced 5 m (16 ft 5 in) apart. The loops can either be cut into permanent slots or be attached temporarily to the road surface.

The classifier operates by timing vehicles between the two loops to give individual vehicle speeds. These data on speed and on the vehicle's presence time over each loop allow vehicle lengths to be calculated. Each vehicle is logged into one of four counting bins according to its estimated length. By the selection of appropriate length bands, simple vehicle classification is practicable

into categories of motorcycles, cars, small trucks, and heavy vehicles. Total flows in each class are recorded in memory at intervals of between 1 min and 24 h as preset by the user.

One optional output for use in setting up the system consists of individual vehicle speeds or lengths, which appear on the liquid crystal display on the front panel of the retriever unit. This facility was used in tests on the accuracy of the system for individual speed or length comparisons. Independent speed checks were carried out by precise timing of vehicle leading axles between pairs of pneumatic tubes, located next to the inductive loops of the speed and length classifier. Classifier length measurements were compared with vehicle manufacturers' data following manual identification of vehicle makes. The accuracy of the independent measurements has been assessed elsewhere (4).

Precise timing was carried out by using a portable roadside microcomputer. A machine code routine was written for a Golden River Environmental Computer to scan the sensors and increment a 32-bit counter between signals on successive tubes. The routine was calibrated by using a stopwatch over intervals of 30 min to 1 h, which gave a count rate of 23 485/s. Repeated short-duration checks against a microsecond-resolution advance timer showed no statistically significant systematic error and a random standard error of ± 0.11 ms (± 0.02 percent).

Hand-written recording of vehicle speeds, lengths, or makes was only possible at low flows. In other cases, individual vehicle results were dictated into a portable tape recorder or where possible were recorded automatically in the portable roadside computer memory. At the busiest site, immediate identification of vehicle makes was impractical, so a cinecamera was triggered by a road sensor to provide a photographic record of each vehicle. Vehicle makes were subsequently identified from the film.

The three sites selected for accuracy checks of the speed and length classifier were each of distinctive character. The first was a 6-m (20-ft) two-way internal-access road on the University of Nottingham campus, which provided a low-speed, low-volume site at which private cars predominate. The second site was an urban dual two-lane highway with a 65-km/h (40-mph) speed limit that carried fairly high volumes of general mixed traffic and buses. The final site was a rural high-speed single-lane highway with local dualing at intersections. Its modest traffic volumes included a higher proportion of commercial vehicles than those of the other sites.

SIMPLE CLASSIFICATION ACCURACY RESULTS

The results of individual vehicle speed comparisons are summarized below (1 km/h = 0.6 mph):

Site	No. of Vehicles	Mean Speed (km/h)	Systematic Difference (km/h)	Random Difference (km/h)
Low speed (1)	204	32.82	-0.32 \pm 0.08	\pm 1.10
Low speed (2)	215	31.29	-0.49 \pm 0.06	\pm 0.86
Urban	327	61.65	-0.09 \pm 0.05	\pm 0.87
High	161	71.54	-0.99 \pm 0.09	\pm 1.07

Speed measurement forms the first stage of length classification, so its accuracy is of considerable interest. The two sets of results presented for the low-speed site correspond to measurements on different days.

At the low-speed and the high-speed sites there were significant systematic differences between the speed measurements of the classifier and the independent system. A proportion of the differences may be due to systematic error in the independent speed measurements. Another factor, however, could be the matching of loop pairs for speed measurement; precise geometry and equality of loop feeder lengths appear to be of considerable importance. Systematic errors could if necessary be overcome by individual site calibration.

The random discrepancies in speed measurements do not vary greatly between sites. A proportion of the discrepancies is simply due to rounding to the nearest kilometer per hour on the liquid crystal display, which of itself would account for about ± 0.3 km/h. A smaller proportion will be due to errors in the independent speed measurements. The remaining random error in the speed measurements is unlikely to be of importance in vehicle classification.

The results of the individual vehicle length comparisons are given below (1 m = 3.2 ft):

Site	No. of Vehicles	Systematic Difference (m)	Random Difference (m)
Low speed	91	-0.70 ± 0.06	± 0.54
Urban	276	-0.90 ± 0.04	± 0.68
High speed	105	-0.17 ± 0.06	± 0.64

Lengths are systematically underestimated at all sites, apparently due to the use of two-turn loops instead of the three turns normally recommended by the manufacturer. The variations between sites are probably associated with different feeder lengths and indicate a need for individual site calibration.

Random errors are also significant; they are of a similar order of magnitude at each site. The main contributor to random error appears to be differences in the lateral position of vehicles. The tabulation below shows how vehicles passing over either edge of the loop have their lengths systematically underestimated in relation to those near the center. At this site the loop width was 2 m (6 ft 7 in) (1 m = 3.2 ft):

Distance from Curb to Nearside Wheel (m)	No. of Vehicles	Systematic Difference (m)
0.0-0.9	12	-1.36
0.9-1.2	22	-1.02
1.2-1.4	60	-0.74

Distance from Curb to Nearside Wheel (m)	No. of Vehicles	Systematic Difference (m)
1.4-1.6	58	-0.67
1.6-1.8	68	-0.89
1.8-2.1	43	-0.96
>2.1	5	-1.24

These results provided a strong indication that the variations in length measurement were associated principally with the characteristics of inductive loops rather than those of the measuring equipment. The effective length of loops appears to vary considerably with the lateral position of vehicles as well as with feeder lengths and number of turns. With this in mind, some laboratory and field trials were carried out at the University of Nottingham into the fundamental properties of inductive-loop layouts.

The laboratory tests set out to examine the magnetic fields of loops by breaking them down into three component parts at right angles. Experimental techniques have been described in detail elsewhere (5). The experiments aimed to produce contour maps that show the limits of the zone of detection for each component of the loop's magnetic field.

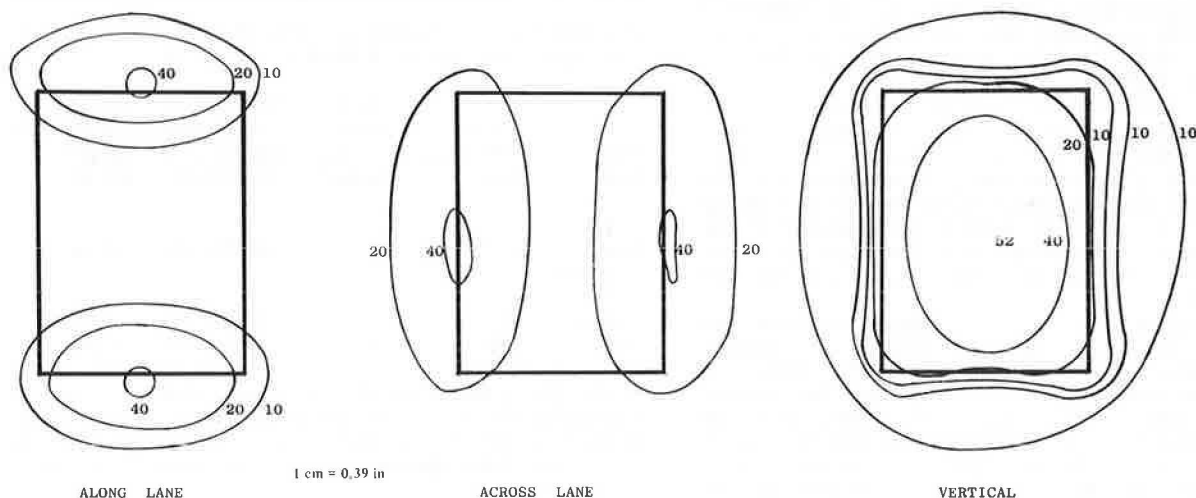
Typical contour plots for a three-turn rectangular loop are shown in Figure 2; the heights of the zone of detection are shown in centimeters. The loop was a one-third scale model of a 2.8-by-2.0-m configuration (9 ft 2 in by 6 ft 7 in).

For a two-dimensional body such as a thin steel plate, detection results when the component of the magnetic field cuts it at right angles. Thus detection begins when a vehicle's front panel enters the horizontal field running along the traffic lane. It ends as the rear panel leaves the equivalent downstream field. Vehicles crossing the edge of the loop can be picked up as their side panels cut the horizontal field running across the traffic lane. The curvature of these field boundaries is such that the loop's effective length changes continuously across its width. These findings were confirmed in a number of field trials.

Experiments with a wide variety of loop layouts indicated that although this basic problem cannot be wholly overcome, it can be reduced by the adoption of broader loops of rectangular outline. The overall width must be limited, however, by the need to prevent the zones of detection from spreading too far into adjacent lanes.

These basic limitations of the loop sensor con-

Figure 2. Zones of detection for three-turn rectangular loop.



strain the accuracy of simple vehicle classification by using loops alone. The effect of the errors will, however, depend on the vehicle length categories selected in relation to the distribution of lengths within the traffic stream. Misclassification will only affect vehicles whose lengths are similar to the category boundary values, and even here random errors will tend to cancel out. Classification accuracy will also be heavily dependent on lane discipline at the survey point, which varies considerably from site to site. Preliminary results from U.K. sites suggest that simple classification can be reasonably reliable based on vehicle length from loops alone.

One weakness of length classification is its inability to distinguish buses from long freight vehicles. An additional parameter, chassis height, can be estimated from the strength of the loop signal, providing opportunities for the extension of simple classification to this additional vehicle category. The problems of chassis height measurement are discussed in later sections.

DETAILED AUTOMATIC CLASSIFICATION

In cases where more detailed information is needed or where more accurate results are sought, a more complex form of automatic vehicle classification may be necessary. The accuracy tests described in this paper were carried out on a detailed classification system that was developed by the U.K. Transport and Road Research Laboratory (TRRL). It consists of permanent sensors in the road for vehicle detection and a roadside, mains-powered microprocessor system for the calculation of various parameters permitting detailed vehicle classification.

The road sensors consist of one inductive loop and two triboelectric axle sensors per lane, as shown in Figure 3. Sensor dimensions can be varied (these are specified as initial data to the microprocessor system), although standard layouts have been used at most classification sites to date. The

axle-detector spacing is usually 1 m (3 ft 3 in), and loops are typically 2.8 m long by 2.0 m wide (9 ft 2 in by 6 ft 7 in).

The roadside equipment includes loop-detector electronics and axle-detector signal-processing units, as well as the microprocessor and its peripherals. These initial interfaces transform raw pulses from the sensors into square-wave signals suitable for input to the microprocessor. The microprocessor itself is an RCA 1802 COSMAC single-board machine with 15K EPROM holding the classification software and 2K RAM for temporary data storage. Permanent recording is on magnetic cartridge. The system is shown in Figure 3.

The classification system resolves road sensor signals to 1-ms timing for the calculation of vehicle parameters. The first parameter, vehicle speed, is calculated from the times of the leading axle on successive axle detectors. Given the speed, wheelbase lengths are derived from the time intervals between axles on a single-axle detector. Overall length is estimated from the vehicle's presence time over the inductive loop. Finally, chassis height is estimated from the strength of the inductive loop signal to assist discrimination between certain classes of vehicle with similar wheelbase and overall lengths. The detector signals from a typical vehicle are shown in Figure 4.

The microprocessor compares the vehicle parameters of wheelbase, overhang, and chassis height with limiting values held in memory. When a parameter match is found, the vehicle class is identified. Twenty-five separate categories of vehicle are distinguished by the existing system, which allows considerable flexibility in modes of aggregation to suit the requirements of individual users. The vehicle categories are shown in Figure 5.

Output from the system is available in real time as a vehicle-by-vehicle listing. More commonly, summaries are produced and stored on magnetic cartridge at intervals specified by the user. The system is capable of producing data on vehicle flow, time headway, speed, class, wheelbase, and overall

Figure 3. Sensor configuration and roadside equipment for detailed classification.

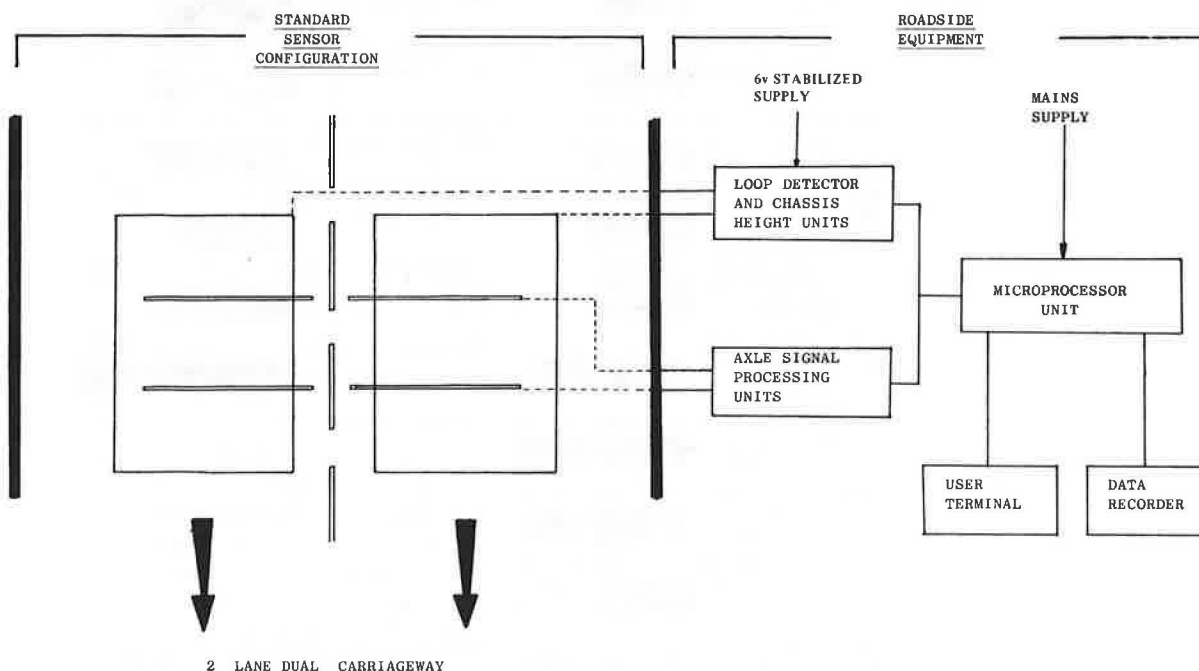


Figure 4. Sensor time sequence diagram.

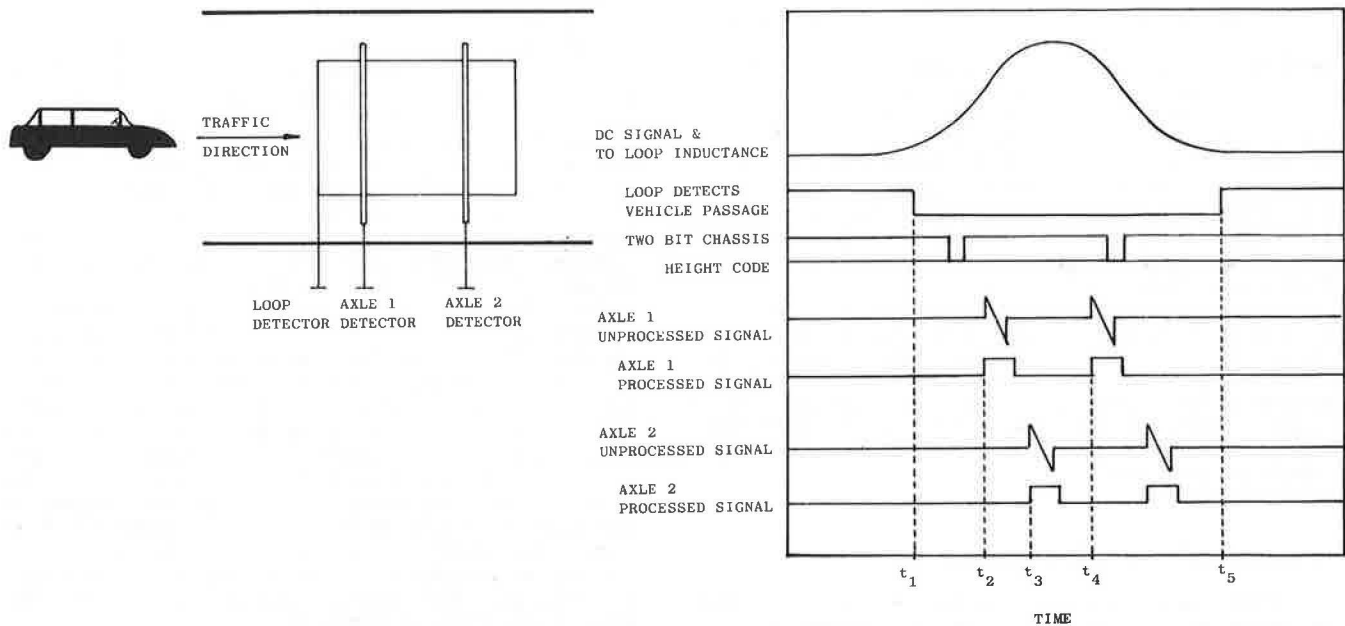



























Figure 5. Vehicle categories for detailed automatic classification.

Class no	Vehicle description	Class no	Vehicle description
0	Moped, scooter motorcycle 	45	Rigid 2 axle HGV +1 axle caravan or trailer 
1	Car, light van, taxi 	46	Rigid 2 axle HGV +2 axle (close coupled) trailer 
2	Light goods vehicle 	51	Artic, 2 axle tractor +1 axle semi-trailer 
21	Car or light goods vehicle+1 axle caravan or trailer 	52	Artic, 2 axle tractor +2 axle semi-trailer 
22	Car or light goods vehicle+2 axle caravan or trailer 	53	Artic, 3 axle tractor +1 axle semi-trailer 
31	Rigid 2 axle heavy goods vehicle 	54	Artic, 3 axle tractor +2 axle semi-trailer 
32	Rigid 3 axle heavy goods vehicle 	55	Artic, 2 axle tractor +3 axle semi-trailer 
33	Rigid 4 axle heavy goods vehicle 	56	Artic, 3 axle tractor +3 axle semi-trailer 
34	Rigid 3 axle heavy goods vehicle 	61	Bus or coach, 2 axle 
35	Rigid 4 axle heavy goods vehicle 	62	Bus or coach, 3 axle 
41	Rigid 2 axle HGV +2 axle drawbar trailer 	7	Vehicle with 7 or more axles 
42	Rigid 2 axle HGV +3 axle drawbar trailer 		
43	Rigid 3 axle HGV +2 axle drawbar trailer 		
44	Rigid 3 axle HGV +3 axle drawbar trailer 		

- 1N Vehicle with 1 axle counted
- 2N 2 axle vehicle not otherwise classified
- 3N 3 axle vehicle not otherwise classified
- 4N 4 axle vehicle not otherwise classified
- 5N 5 axle vehicle not otherwise classified
- 6N 6 axle vehicle not otherwise classified

length for four lanes of traffic with capacity flow in each lane.

As an example of the classification methodology, consider the problem of correctly classifying the following three types of four-axle vehicle:

1. Class 22, car + two-axle trailer;
2. Class 46, rigid two-axle truck + two-axle (closed-coupled) trailer; and
3. Class 52, two-axle tractor + two-axle semi-trailer.

The classification program contains the vehicle dimension reference table, which for these three categories consists of the values given in Table 1.

From Table 1 it can be seen that the class-46 vehicle can be identified without using the chassis height code. Separation of classes 46 and 52 occurs on the second wheelbase dimension, and similarly classes 46 and 22 are separated by the first wheelbase. The separation between classes 22 and 52, however, is based on the chassis height code if and only if the wheelbase or wheelbases actually overlap.

Similar data are held in memory for other categories of vehicle; limiting values for the appropriate parameters are specified. Should the software fail to find a match to any of the 25 recognized

classes, the vehicle is logged according to its number of axles but is otherwise unclassified. Further details of the system and its operation are given elsewhere (6,7).

ACCURACY OF DETAILED SYSTEM

In order to assess the accuracy of the detailed classification equipment, photographic logging of vehicles has been carried out at four test sites. Vehicle class, as identified from film, has been compared with the microprocessor class assignment for some 15 000 vehicles in rural and urban locations in order to determine the capabilities of the system in a range of traffic flow regimes. At other sites, classifier output has been compared with the results of manual classified counts.

Accuracy studies undertaken by TRRL at two rural sites, where free-flow conditions prevail, suggest that the overall accuracy of classification is about 92 percent (7). Accuracies are lower for certain classes of vehicle where particular problems are met in discriminating between similar parameters. University of Nottingham accuracy studies (8) indicate that under free-flow urban traffic conditions, overall accuracy remains quite high. However, as congestion levels, bringing slow-moving traffic and poor lane discipline, accurate classification becomes increasingly difficult.

The accuracy of each installation has been assessed by the compilation of accuracy matrices, which compare microprocessor and visual classification. Summaries of the matrices are provided in Tables 2 and 3, which indicate the accuracy of classification for the more common vehicle types. The results allow compensating errors between classes where they occur.

Table 2 shows that at the rural sites, despite

Table 1. Typical U.K. vehicle dimensions for detailed classification.

Vehicle Class	Wheelbase (m)			Chassis Height Code	Overhang (m)
	1	2	3		
21-22	1.90-2.95	1.90-6.00	0.50-1.30	1, 2, or 3	0-12.75
51-55	1.90-3.51	3.76-15.0	1.05-2.50	0	0-12.75
41-46	2.96-9.20	1.90-3.75	1.05-2.50	Not used	0-12.75

Table 2. Detailed classification accuracy under free-flow conditions.

Vehicle Class	Rural Site ^a				Urban Site ^b			
	Observed Total	Classifier Total	Error	Percentage of Error	Filmed Total	Classifier Total	Error	Percentage of Error
0	57	40	-17	30	127	61	-66	52
1	4159	3976	-192	5	5712	5979	+267	5
2	266	392	+126	47	781	508	-273	35
21-22	56	56	-	-	34	43	+9	26
31	319	357	+38	12	675	641	-34	5
32-35	43	43	-	-	121	128	+7	6
41-46	4	4	-	-	25	31	+6	24
51-55	71	70	-1	2	239	226	-13	5
61	39	41	+2	5	55	42	-13	24
N-classes	-	44	+44	-	1	99	+98	-
Missed	-	-	-	-	53	46	-	-

^aOverall accuracy = 92 percent.

^bOverall accuracy = 90 percent.

Table 3. Detailed classification accuracy for experimental installations.

Vehicle Class	Noncongested Conditions ^a			
	Filmed Total	Classified Total	Error	Percentage of Error
0	23	12	-11	49
1	2103	1879	-224	11
2	283	483	+200	71
21-22	21	21	0	-
31	336	347	+11	3
32-35	55	44	-11	20
41-46	3	9	+6	200
51-55	112	98	-14	13
61	15	17	+2	13
N-classes	1	100	+99	-
Missed	64	23	-	-

^aOverall accuracy = 80 percent.

the high overall accuracy, a major source of error lies in the classification of cars, vans, and light commercial vehicles. The physical similarity between cars and vans makes their separation difficult, so classification errors can be substantial. Under free-flow urban conditions, results are comparable with those for the rural situation, as indicated in Table 2. The problem of misclassification of cars, vans, and two-axle trucks is again apparent.

The serious underestimation of motorcycles (class 0) in Table 2 (urban site) is related to the use of commercial axle detectors that have low sensitivity and do not cover the full width of the lane. A new form of axle detector developed by TRRL has been used at other sites. This tends to reduce problems of motorcycle classification by its greater sensitivity and its coverage of the whole traffic lane.

Both urban and rural studies indicated that the most common causes of misclassification are related to vehicles changing lanes. Discrimination of cars, vans, and two-axle trucks is dependent on an estimation of chassis height; vehicles with higher chassis give weaker signals. The same effect can, however, result when vehicles cross the side of the loop rather than its center. Although special routines are provided in the detailed classifier to detect straddling vehicles and adjust their chassis-height values, the resulting classification was still not wholly satisfactory.

To help resolve these problems, reference was again made to the University laboratory study of loop zones of detection (5). Model tests and full-scale field trials led to the selection of a series-wound double rectangular loop in each lane, as shown in Figure 6. The tests indicated that more uniform length and chassis height measurements could be expected from these loops for a wider range of vehicle lateral positions. The new loops were installed at an experimental urban site with a high proportion of straddling vehicles following an upstream merge. Another feature of the site is congestion during busy periods.

Weaving at the two urban sites was compared by using the Golden River Environmental Computer system for precise timing of vehicles across three pneumatic tubes. Parallel tubes were used to measure vehicle speeds, and the time on a third, diagonal tube was used to indicate lateral position. Figure 7 shows the distribution of vehicle lateral positions at both installations; the higher proportion of straddlers is found at the experimental site.

Two separate accuracy studies were undertaken at this site during free-flow traffic conditions. The first, a routine comparison of microprocessor and film classification, indicated an overall accuracy of about 80 percent (Table 3). The reduced accuracy was clearly related to the high proportion of straddling vehicles, as shown for example in the number of cars wrongly classed as vans. It was not clear from these results whether the new loop design had helped to limit the problem of misclassification.

Further comparisons were therefore carried out by using a test vehicle at a range of lateral positions. The variations of loop output with lateral position for the conventional and experimental loops are shown in Figure 8. The experimental loops do give a more consistent signal over a wider range of lateral positions than the conventional loops. On the other hand, they do not cover as wide an area as had been expected from the experimental tests. One factor in this result appears to be interference between the fields of adjacent loops.

To summarize, the results of the tests on detailed automatic classification suggest that high overall accuracies can be obtained at many free-flow sites where the incidence of lane changing is low.

Some classes of vehicle still create problems, and work is continuing on further improvements. Accurate classification under urban traffic conditions presents greater difficulties, and research in this area is at an earlier stage.

EVALUATION OF U.S. AUTOMATIC CLASSIFICATION SYSTEMS

Six automatic classification systems have been tested by the Materials and Research Division of the Maine Department of Transportation for the Federal Highway Administration (FHWA) (10). Standard accuracy checks were applied to six currently available systems. Five of these, including the Golden River four-bin length classifier, provided simple classification on the basis of axle configuration or vehicle overall length. The detailed classification system developed at TRRL was also tested at the Maine facility.

Accuracy checks included the cross-comparison of automatic classification listings with film records and the collection of summary data over longer test periods (overnight and over weekends).

Results for the simple classification systems indicated that the poor reliability of both the vehicle sensors and the classification equipment was a major source of error in vehicle identification. Pneumatic tubes were subject to both accidental and purposeful damage, which resulted in axle undercounting, and in addition poor signal definition from the air-switch units led to axle undercounting even when pneumatic tubes were not damaged.

Systems that used inductive loops for vehicle sensing were found to be oversensitive to minor adjustments of, or variations in, the loop-detector units. The tests with the Golden River four-bin length classifier were limited by equipment faults, but a small sample of results suggested that loop problems were less noticeable and that the measurement of vehicle overall length was generally more accurate.

It was noted that the classification schemes used in the various simple classification systems did not offer adequate detail when compared with the data-collection requirements suggested by FHWA. Not only were the schemes limited by the number of categories used, but there could be misleading overlap between the categories that were defined.

Tests with the TRRL detailed classification system indicated that many of the above problems were overcome through the use of both loop and triboelectric sensors. Specific accuracy checks indicated that overall and axle-length measurements were extremely good and that speed measurements were to within ± 1.61 km/h (1 mph) of recordings made with radar equipment. In all, 98 percent of the 3000 vehicles checked were classified correctly, despite a much-reduced initial calibration procedure.

FURTHER DEVELOPMENTS IN DETAILED CLASSIFICATION

Additional research is currently in progress at both the University of Nottingham and TRRL to further improve the performance of the detailed vehicle classifier. Areas of interest include classification under congested traffic conditions, further improvements to sensor response for lightweight vehicles, further modifications to loop design for chassis height discrimination, and the classification and counting of bicycles.

Software development has been studied at the University by means of a classification simulation run on a minicomputer. Raw signals from the sensor array, recorded in conjunction with films of vehicle flow, provide data on which modifications to soft-

Figure 6. Experimental sensor configuration for detailed classification.

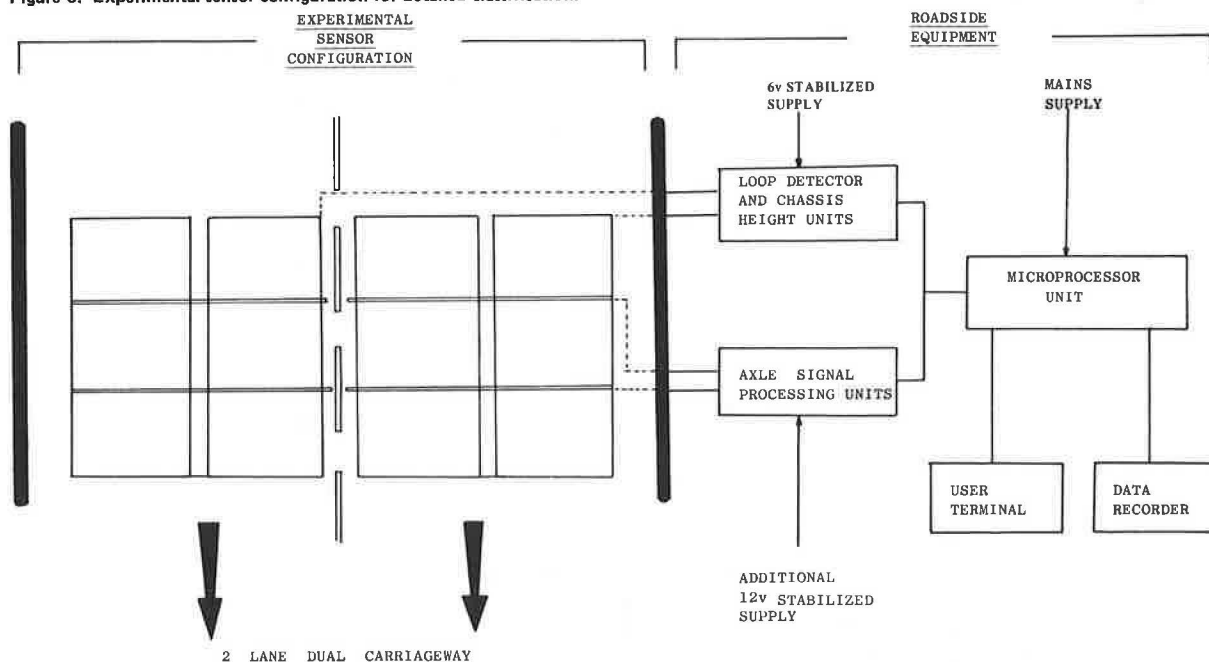


Figure 7. Distribution of vehicle lateral positions at two urban sites.

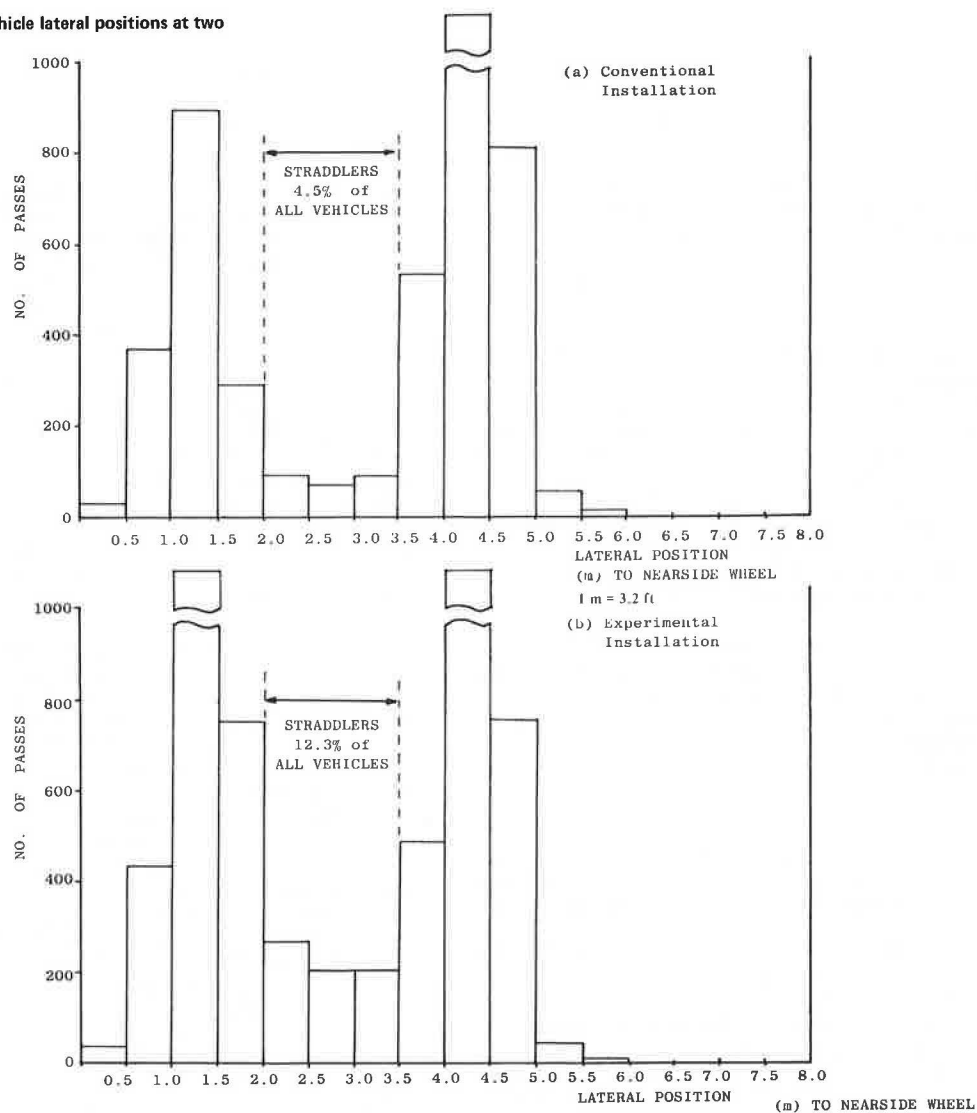
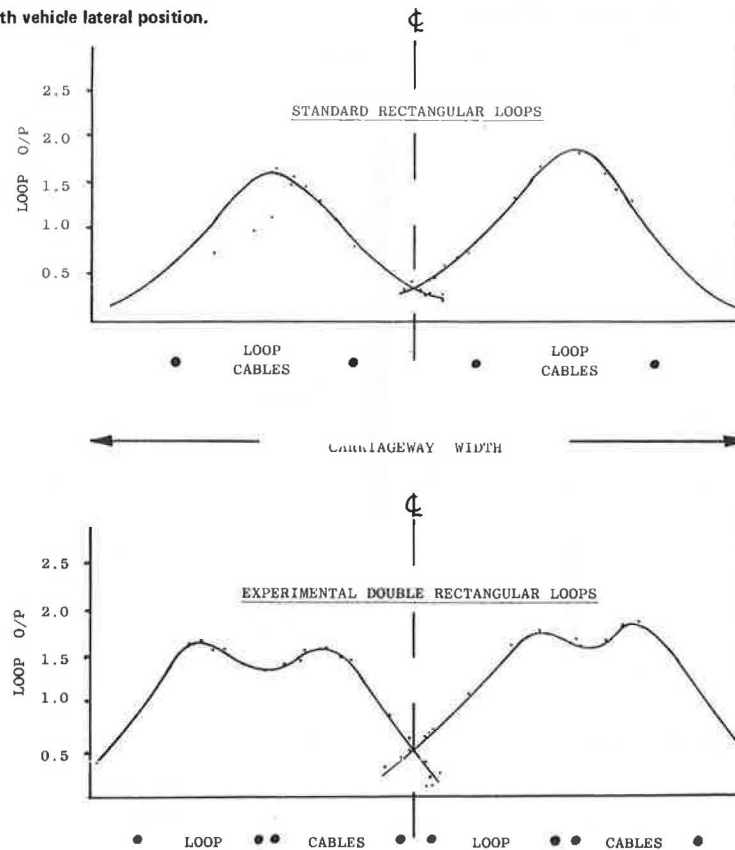


Figure 8. Variation of loop output with vehicle lateral position.



ware can be tested and appraised. Under experimental conditions, fine tuning of parameter boundary values and new procedures for dealing with unclassified vehicles have improved the accuracy of classification from 88 to 97 percent on a sample of 5000 vehicles. It remains to be seen whether these improvements will prove sufficiently robust to be transferable to routine conditions in the field.

The development of queue conditions over the vehicle sensors is known to cause a substantial reduction in the accuracy of classification. A major problem is the sensing of very slow-moving vehicles at the loops and axle detectors. Modifications to hardware are already in progress, and software development of special routines for congested conditions will follow in due course.

A final area of interest to many potential users of the system is the development of temporary sensors for a portable detailed classification system. Recent improvements to temporary loop sensors have resolved some problems in this area, but temporary axle detectors other than pneumatic tubes have yet to be developed to satisfactory standards. There are many possibilities for developments in this area and further work is currently proposed.

CONCLUSIONS

The evidence available on manual classification suggests that its accuracy is very much lower than might commonly be supposed. When coupled with the errors of sampling, scaling, and forecasting, the reliability of manual classified traffic count data must be seriously open to question.

The development of microprocessor equipment for traffic data collection makes long-term monitoring at increased numbers of survey points a practical proposition. Apart from the probability of improvements in classification accuracy by using automatic

equipment, the increased coverage through space and time should lead to smaller sampling errors, which gives a more reliable base for the forecasting of future traffic levels. Furthermore, in conjunction with axle-load data, the availability of more detailed classified count information might also pay dividends in the fields of pavement design and highway maintenance.

The accuracy of simple vehicle classification by using loops alone may be limited more by the inherent properties of loop sensors than by the microprocessor equipment or software adopted in any particular case. Nevertheless, provided that classification sites are chosen well and on-site calibration is carried out with care, the evidence suggests that classification can be sufficiently accurate for the determination of PCE flows. Simple vehicle classification equipment that uses only pneumatic-tube sensors has been shown to be less reliable due to the susceptibility of these sensors to accidental or purposeful damage.

The detailed vehicle classification equipment, which uses loop and axle sensors, is now commercially available from the Golden River Corporation. Further developments to the system are currently under way in a number of areas, and progress can be anticipated in the extension of the system to a full range of urban sites as well as to portable equipment for temporary census points. The widespread application of these techniques may still be some years ahead, although some applications for routine traffic monitoring are already under way.

In the longer term, the falling cost of microprocessor equipment together with increasingly high labor costs point to an inevitable growth in the attractiveness of automatic vehicle classification techniques. Some of those techniques will be applied in the future; others are available now. If properly used, their application should lead to more

reliable classified count data for highway planning and operation.

ACKNOWLEDGMENT

The cooperation of TRRL and of the Department of Transport and the Nottinghamshire County Council as highway authorities is acknowledged. The research is sponsored by the University of Nottingham Allied Breweries Bequest and is carried out by permission of P.S. Pell, Head of the Department of Civil Engineering. Responsibility for the paper and its contents, however, remains ours.

REFERENCES

1. Report of the Advisory Committee on Trunk Road Assessment. U.K. Department of Transport, London, 1977.
2. G. Penrice. Problems of Processing Data Collection and Analysis. Presented at Conference on Traffic Data Collection, Institute of Civil Engineers, London, 1978.
3. Traffic Appraisal Manual. U.K. Department of Transport, London, 1981.
4. P. Davies. The Accuracy of Automatic Speed and Length Classification. Department of Civil Engineering, Univ. of Nottingham, England, 1982.
5. P. Davies, D.R. Salter, and M. Bettison. Loop Sensors for Vehicle Classification. *Traffic Engineering and Control*, Vol. 23, No. 2, 1982, pp. 55-59.
6. An Automatic Method to Classify, Count, and Record the Axle Weight of Road Vehicles. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, TRRL Leaflet LF 639, 1981.
7. R.C. Moore, P. Davies, and D.R. Salter. An Automatic Method to Count and Classify Road Vehicles. Presented at International Conference on Road Traffic Signalling, Institution of Electrical Engineers, London, 1982.
8. D.R. Salter and P. Davies. The Accuracy of Automatic Vehicle Classification. Presented at 14th Annual Universities Transport Studies Group Conference, Univ. of Bristol, England, 1982.
9. Evaluation of Vehicle Classification Equipment. Maine Department of Transportation, Augusta, 1982.

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

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

Application of Counting Distribution for High-Variance Urban Traffic Counts

STEPHEN G. RITCHIE

This paper describes an application of the negative binomial counting distribution to high-variance, short-period traffic counts collected on urban arterial roads during peak-period flow conditions. The data were collected at four sites downstream from signalized intersections in metropolitan Melbourne, Australia, during 1977-1978. Alternative parameter estimation techniques are described as well as a simple method for dealing with transient traffic demand patterns. The results of these comparative evaluations suggest that the negative binomial distribution can be applied quite simply to commonly occurring problems that involve high-variance traffic counts and that results are often markedly better than those for other elementary counting distributions such as the Poisson distribution.

The inherent statistical variability of many flow-related attributes of urban transportation systems has important implications for the design, operation, and use of such systems, e.g., in transit service design (1), traveler mode choice (2,3), and delay at signalized intersections (4), to name only a few.

Moreover, continued emphasis on transportation systems management policies has increased the need for more accurate and realistic models that are useful for urban traffic systems analysis. Such models include basic statistical distributions of traffic characteristics such as vehicle headways, speeds, gap acceptances, and vehicle arrivals at a point, which are routinely used by traffic engineers and analysts in analyzing traffic system performance, designing and improving traffic facilities, and developing traffic simulation models.

This paper is concerned specifically with

traffic-counting distributions, which describe the distribution of vehicle arrivals at a point during a given time interval. Gerlough and Huber (5) have described the elementary traffic-counting distributions, namely, the Poisson distribution, the binomial distribution, and the negative binomial (NB) distribution. These statistical distributions have been known to traffic engineers and analysts for some time. However, the NB distribution has not enjoyed the same wide application as the Poisson distribution to traffic problems, despite its apparent superiority under fairly common variable flow-rate conditions where the variance of traffic counts is high.

In this paper an application is described of the NB counting distribution to high-variance traffic counts and more specifically to short-period counts collected on urban arterial roads during peak-period flow conditions. Alternative parameter estimation techniques are described as well as an explicit but simple method for dealing with transient traffic demand patterns (i.e., time-varying traffic flow rates). The results of these comparative evaluations are presented.

STOCHASTIC NATURE OF URBAN TRAFFIC FLOWS

In an urban arterial road network, one of the major factors that influences the nature of vehicle arrivals on a link is the presence of upstream signalized intersections. The cyclic interruption to flow pro-

duced by signals operating under peak-period traffic conditions tends to result in pulsed flows on the links of an urban network. At points downstream from a signal it can often be observed that during the early portion of a cycle, traffic flow is high and in the form of near-saturated platoons. Later in the cycle the flow is often light; it then consists of turning vehicles that have filtered through the upstream intersection and any other vehicles that might have entered the link at minor cross streets.

As Gerlough and Huber (5) have noted, if a traffic-counting interval corresponds to the green portion of the signal cycle or to the complete signal cycle, cyclic effects may be masked. However, if the counting interval is short (e.g., 10 s), there will be periods of high flow and periods of low flow. Thus, combining such short-period counts into one distribution results in a high variance, which produces a variance-to-mean ratio significantly greater than 1.0, the value expected if arrivals were random and Poisson in nature.

In practice, substantial nonrandomness can also be exhibited by urban traffic counts with counting intervals considerably longer than 10 s. For example, Newell (6) has asserted that typical variance-to-mean ratios of 60-s counts lie between 1.0 and 1.5. Williams and Emmerson (7) obtained an average variance-to-mean ratio of 1.51 in Newcastle, England, and Miller (8) noted that observations in Birmingham generally produced values between 1.0 and 2.0, though one set of data gave a ratio of just over 4.0. The data analyzed in this paper indicate that variance-to-mean ratios of 10-s counts taken on multilane urban arterial roads downstream from traffic signals can be at least as high as the values mentioned above. Also, invariably, as the variance of the counting distribution increases relative to the mean, the NB distribution provides a better fit to the data than a Poisson distribution.

However, a complication is that during peak periods, traffic demands are typically time dependent, so that significant degrees of nonstationarity are likely to exist in time-series short-period traffic counts collected in the field. In theory, this factor invalidates all elementary counting distributions that are appropriate only for time-stationary processes and suggests that more advanced time-series analyses may be necessary under such conditions (9,10). Such analyses, of course, probably exceed not only the experience and time constraints of practicing traffic engineers, but also their computational resources and accuracy requirements. In addition, elementary statistical distributions, and particularly the Poisson distribution, have for many years been applied widely by traffic engineers to problems such as the analysis and description of vehicle arrivals at a point, the design of turn pockets, warrants for traffic signals and pedestrian crossings, as well as accident analysis and simulation of traffic flows (11).

It is highly likely that such distributions will also prove valuable in the future, under a range of flow conditions, principally due to their relative simplicity. Accordingly, it is suggested in this paper that when an elementary counting distribution is required for a situation in which the variance of counts is high or is expected to be high, the NB distribution should be given greater consideration than in the past. This distribution will typically be more realistic and will provide a better fit to the data than a Poisson distribution. The remainder of the paper is concerned with empirical estimation and application of the NB distribution.

DATA

The data used in this analysis were collected in 1977-1978 at several sites in metropolitan Melbourne, Australia, on arterial roads downstream from signalized intersections. Four surveys were conducted, and data were collected manually by using digital counters. In each survey, a 10-s counting interval was used. This particular interval was chosen because the results of the analyses were to be used in a traffic simulation model, a description of which may be found elsewhere (12). The characteristics of each survey site are briefly described below.

The site for surveys 1 and 2 was about 10 km from the Melbourne central business district (CBD) on a major divided (three-lane) circumferential arterial road. Survey durations were 7:30-9:30 a.m. for survey 1 and 3:30-5:15 p.m. for survey 2. The counting station was located approximately 300 m downstream from a signalized intersection with cycle length of about 80 s.

Survey 3 was conducted about 19 km from the Melbourne CBD on a major divided (three-lane) radial arterial road. The counting station was located approximately 250 m downstream from an intersection with signal cycle length of 100 s. The survey duration was 7:15-10:00 a.m.

Survey 4 was conducted about 16 km from the Melbourne CBD on a four-lane undivided radial arterial road. The counting station was located approximately 600 m downstream from an intersection with signal cycle length of 90 s and about 250 m downstream from a signalized pedestrian crossing. The survey duration was 7:00-9:15 a.m.

At each site, the adjacent land use was residential. For surveys 1-3, the weather was cool with intermittent showers, whereas for survey 4, the weather was cool and dry.

As expected, time trends in the peak-period flows were evident from 15-min flow profiles for each survey. Analysis of the mean and variance of the 10-s counts from each survey was then undertaken for both individual 15-min intervals and each whole peak period of counts. These results are shown in Table 1. The high variance-to-mean ratios for these surveys, both for 15-min and whole-period counts, are notable, particularly for survey 3 (Table 1), where variance-to-mean ratios of the 10-s counts in some 15-min periods exceeded 5.0, whereas that for the whole period was 4.652. These results are consistent with the high degree of nonrandomness, which is to be expected in such urban traffic counts but which has often been unaccounted for in previous analyses and studies.

NB DISTRIBUTION

Description

Unlike the single-parameter Poisson distribution, the NB distribution is specified by two parameters, the mean m and a parameter k . Both these parameters must be estimated from data or at least from knowledge about the mean and variance of the data. The main issue, however, concerns the procedure for estimating the parameter k , as discussed in the next section. It can be shown that the Poisson distribution is obtained as a limiting form of the negative binomial when the parameter k approaches infinity. Furthermore, the NB distribution can be obtained from the Poisson distribution when the Poisson parameter is constant during each counting interval but varies between intervals with an Erlang (or

Table 1. Analysis of 10-s traffic-count results for each study.

Time	Avg Flow (vehicles/10 s)	Variance/ Mean Ratio	Time	Avg Flow (vehicles/10 s)	Variance/ Mean Ratio
Survey 1 ^a			Survey 2 ^a		
7:30 a.m.	2.522	1.142	3:30 p.m.	1.056	1.136
7:45 a.m.	2.689	1.577	3:45 p.m.	1.622	1.158
8:00 a.m.	2.856	1.311	4:00 p.m.	1.611	1.321
8:15 a.m.	3.233	1.675	4:15 p.m.	1.411	1.750
8:30 a.m.	3.022	1.480	4:30 p.m.	1.767	1.565
8:45 a.m.	2.822	1.916	4:45 p.m.	1.544	1.574
9:00 a.m.	2.000	1.494	5:00 p.m.	1.367	1.865
9:15 a.m.	1.714	1.765	Whole period	1.483	1.503
Whole period	2.615	1.613			
Survey 3 ^a			Survey 4 ^a		
7:15 a.m.	5.944	4.009	7:00 a.m.	3.478	2.728
7:30 a.m.	5.911	4.139	7:15 a.m.	5.133	3.000
7:45 a.m.	6.156	4.662	7:30 a.m.	5.256	2.662
8:00 a.m.	6.111	5.415	7:45 a.m.	5.433	3.412
8:15 a.m.	6.033	5.432	8:00 a.m.	4.889	2.971
8:30 a.m.	6.333	5.003	8:15 a.m.	4.189	3.154
8:45 a.m.	5.811	4.946	8:30 a.m.	4.211	3.130
9:00 a.m.	4.689	5.011	8:45 a.m.	4.200	3.420
9:15 a.m.	4.166	2.919	9:00 a.m.	3.600	2.483
9:30 a.m.	3.544	3.596	Whole period	4.488	3.084
9:45 a.m.	3.244	2.316			
Whole period	5.268	4.652			

^a15-min periods starting at times listed below.

Pearson Type III) density function (13).

The NB probability $P(n)$ of n arrivals in a given counting interval can be written in several forms, one of which is

$$P(n) = \binom{n+k-1}{k-1} [k/(m+k)]^k [m/(m+k)]^n \quad (1)$$

where

$n = 0, 1, 2, \dots$,
 m = mean arrival rate,
 k = distribution parameter ($k > 0$), and

$$\binom{n+k-1}{k-1} = (n+k-1)! / (n! (k-1)!)$$

Recursion equations have also been derived:

$$P(0) = [k/(m+k)]^k \quad (2)$$

$$P(n) = [(n+k-1)/n] [m/(m+k)] P(n-1) \text{ for } n \geq 1 \quad (3)$$

Parameter Estimation

Large-sample methods of parameter estimation for the NB distribution have been summarized by Williamson and Bretherton (14). The principal techniques are the method-of-moments (MM) procedure and the maximum-likelihood (ML) method. In the MM procedure, the first two distribution moments are simply estimated from the sample moments and then used to estimate the parameter k . On the other hand, although the ML method is generally more efficient, it is somewhat more complex computationally. It involves finding parameter estimates that maximize the sample-likelihood function, which gives the relative likelihood of observing the sample data as a function of the parameter values (15).

Let $f(n)$ be the observed frequency of n arrivals per counting interval ($n = 0, 1, 2, \dots$), z be the highest value of n observed, and N be the total number of observations (or counts). Then

$$N = \sum_{n=0}^z f(n) \quad (4)$$

and the sample mean m is given by

$$m = (1/N) \sum_{n=0}^z n f(n) \quad (5)$$

The sample variance v is given by

$$v = [1/N(N-1)] \left\{ N \sum_{n=0}^z n^2 f(n) - \left[\sum_{n=0}^z n f(n) \right]^2 \right\} \quad (6)$$

By using the MM procedure to estimate the NB parameters, the mean arrival rate m is obtained from Equation 5, whereas the parameter k is obtained from

$$k = m / [(v/m) - 1] \quad (7)$$

where m and v are obtained from Equations 5 and 6, respectively. From Equation 7, it is apparent that the variance-to-mean ratio must be greater than 1 to enable the NB distribution to be fitted. Also, as this ratio approaches unity (the value for a Poisson process), \hat{k} approaches infinity (also the limiting value for a Poisson process). In addition, it is clear that the MM procedure requires only the sample mean and either the sample variance or the variance-to-mean ratio. The importance of this is that a realistic counting distribution may be specified with only limited sample data, simply on the basis of a mean flow rate and a suitably chosen variance-to-mean ratio. Limited empirical evidence, to be described in the next section, suggests that the variance-to-mean ratio of 10-s traffic counts on urban arterial roads increases nonlinearly with the mean flow rate. The variance-to-mean ratios in Table 1 for surveys 1-4 fall in the range 1.1-5.4. Also, as noted earlier, typical values for 60-s traffic counts have been found to lie between 1.0 and 2.0. These values may provide some guidance in selecting a suitable default variance-to-mean ratio, especially when the variance of counts cannot be calculated directly from sample data in a specific application.

In the ML approach to estimating the distribution parameters, the mean arrival rate m is found, as before, from Equation 5. However, an ML estimator of the parameter k is found from the solution for k

(other than $k = \infty$) of the equation $g(k) = 0$:

$$g(k) = N \log [1 + (m/k)] - \{ [f(1) + f(2) + \dots + f(z)]/k \} - \{ [f(2) + f(3) + \dots + f(z)]/(k+1) \} - \dots - [f(z)/(k+z-1)] \quad (8)$$

The equation $g(k) = 0$ can be solved directly on many computers and on some programmable calculators by using a standard routine for finding the roots of an equation. Alternatively, an iterative solution procedure such as the Newton-Raphson method (16) may be readily programmed for a microcomputer as well as for some hand calculators. By using the Newton-Raphson method, the value of k for the $(j+1)$ st iteration is given by

$$k_{j+1} = k_j - [g(k_j)/g'(k_j)] \quad j = 0, 1, 2, \dots \quad (9)$$

where k_{j+1} , k_j are values of k for the $(j+1)$ st and j th iterations and $g(k)$ is as in Equation 8.

$$g'(k) = [-Nm/k(m+k)] + \{ [f(1) + f(2) + \dots + f(z)]/k^2 \} + \{ [f(2) + f(3) + \dots + f(z)]/(k+1)^2 \} + \dots + [f(z)/(k+z-1)^2] \quad (10)$$

In applying Equation 9, a starting value of k (k_0) can be found by using the MM procedure. However, if initial values of k much greater than the root are taken, spurious results can be obtained. In fact, the ML results reported in this paper were obtained by using Equation 9 with a fixed initial value of $k_0 = 1.0$ because, in some cases, convergence could not be obtained with an MM initial value.

To decide when an acceptably accurate estimate of k has been obtained from Equation 9, a decision rule in the form of a convergence criterion is necessary. The criterion used allowed a maximum relative error of 10^{-4} or 0.01 percent. Iterations were terminated when the following inequality was satisfied:

$$|k_{j+1} - k_j|/k_j < 5 \times 10^{-5} \quad (11)$$

Once convergence was achieved, the ML estimate of k was obtained from

$$k = k_{j+1} \quad (12)$$

Transient Traffic Demands

In view of the time dependencies exhibited by the peak-period traffic counts for surveys 1-4, a method was proposed that might account more satisfactorily, in a very simple fashion, for the underlying non-stationarity. This involved fitting separate distributions to the 10-s counts in each 15-min interval and aggregating the resulting frequencies to form an aggregated whole-period frequency distribution for each survey. The purpose of this procedure was to investigate whether the aggregated distribution of counts would fit the data better while at the same time explicitly recognize the transient nature of the traffic demand pattern for each survey. These results, together with those for the normal method of estimating a single distribution on the whole sample of peak-period counts, are presented in the next section.

RESULTS

A number of comparative analyses were conducted by using each of the four data sets from surveys 1 through 4. These analyses examined several issues, including methods of applying the theoretical counting distributions to peak-period counts, parameter-estimation techniques, and goodness-of-fit statistics. More specifically, for each of the four

surveys the overall goodness of fit of both Poisson and NB distributions to the collected data was determined, two methods of applying the Poisson and NB distributions to the peak-period counts were investigated, and in each analysis that involved the NB distribution, two methods of parameter estimation were used, namely, the MM procedure and the ML method. However, before further discussion of these analyses and results, the goodness-of-fit statistics are described.

All goodness-of-fit statistics were based on r pairs of observed and predicted frequencies. For both 15-min and whole-period frequency distributions, the cumulative predicted probability distribution (Poisson or NB as appropriate) was used to derive r . The minimum value of r was chosen that satisfied the following:

$$\sum_{r=0}^{r-1} P(r) > 0.99 \quad (13)$$

where $P(r)$ is the predicted probability of r vehicle arrivals in a 10-s counting interval.

Consideration was given to several goodness-of-fit statistics, in particular to Kolmogorov-Smirnov and chi-square statistics. However, the nonparametric Kolmogorov-Smirnov test was not employed because the predicted distribution parameters were estimated from the sample (17). Further, use of chi-square statistics would have involved several difficulties. These centered on the statistical requirement for a minimum predicted frequency per cell, which would necessitate combining some adjacent frequencies in the tails of distributions and would lead to difficulty or ambiguity in the interpretation of chi-square statistics for different models applied to the same data set.

This difficulty or ambiguity may arise not because of the minimum frequency requirement per se, but because different cell structures, different numbers of cells, different numbers of parameters estimated from the observed data and required for the calculation of predicted frequencies, and hence different degrees of freedom can be defined for different models. Also, the large sample strictness inherent in these statistics results in very strict tests and a tendency to reject the fit of many models unless they fit the data extremely well (18).

It was therefore judged desirable to use statistics that could be readily interpreted and used to compare the goodness of fit of different counting distributions to the same data set. Accordingly, a mean absolute deviation statistic (D) and coefficient of determination (R^2) were chosen.

The absolute deviation (d) between any observed and predicted frequency is given by

$$d = |O(n) - F(n)| \quad (14)$$

where $O(n)$ is the observed frequency of n arrivals per counting interval and $F(n)$ is the predicted frequency of n arrivals per counting interval.

The mean absolute deviation (D) between the observed and predicted frequency distribution is then given by

$$D = \sum_{n=0}^{r-1} |O(n) - F(n)|/r \quad (15)$$

with r defined as before. The main advantage of using absolute differences for D is that if the predicted frequency distribution underestimates some frequencies and overestimates others, positive and negative differences cannot cancel to give a false measure of the overall goodness of fit.

The coefficient of determination was obtained

Table 2. Fitted counting distributions and results for each survey.

Survey	Distribution	m^a (vehicles/10 s)	\hat{k}	Iterations	D	R^2
1	Poisson	2.615 (2.496, 2.734)	-	-	28.42	0.61
	NB-MM	2.615 (2.457, 2.772)	4.266 (2.377, 6.156)	-	13.58	0.91
	NB-ML	2.615 (2.457, 2.772)	3.439 (2.554, 4.323)	7	11.93	0.93
2	Poisson	1.483 (1.388, 1.578)	-	-	26.10	0.79
	NB-MM	1.483 (1.364, 1.601)	2.949 (1.823, 4.076)	-	6.73	0.99
	NB-ML	1.483 (1.364, 1.601)	2.639 (1.741, 3.538)	6	5.28	0.99
3	Poisson	5.268 (5.125, 5.411)	-	-	69.51	0.13
	NB-MM	5.268 (4.913, 5.622)	1.442 (1.232, 1.653)	-	14.79	0.79
	NB-ML	5.268 (4.913, 5.622)	1.024 (0.908, 1.140)	2	8.94	0.94
4	Poisson	4.488 (4.342, 4.634)	-	-	50.84	0.01
	NB-MM	4.488 (4.200, 4.776)	2.153 (1.778, 2.529)	-	13.63	0.76
	NB-ML	4.488 (4.200, 4.776)	1.547 (1.331, 1.763)	4	9.12	0.89

^a95 percent confidence intervals for m and \hat{k} are shown in parentheses and are based on Anscombe's study (19).

from a linear regression between the predicted and observed frequencies of arrivals for each survey.

As noted earlier, whole-period fits of the counting distributions were determined by estimating the distribution parameters from the whole sample of 10-s counts for each survey. The results are shown in Table 2 for the Poisson distribution, NB distribution with parameter estimation by the MM procedure (NB-MM), and NB distribution with parameter estimation by the ML method (NB-ML). It can be seen that the Poisson fits were inferior to the NB distribution, especially for the higher-flow, higher variance-to-mean ratio surveys, 3 and 4. For these two surveys the Poisson fits were very poor. For all surveys, the NB-ML distributions gave better results than the NB-MM distributions; the difference was most noticeable for surveys 3 and 4.

For each survey, distribution parameters were then estimated separately for each 15-min interval. The frequency distributions so formed were aggregated over each peak period to form aggregated whole-period fits, shown below:

Survey	Distribution	D	R^2
1	Poisson	25.32	0.66
	NB-MM	12.97	0.92
	NB-ML	11.88	0.93
2	Poisson	24.39	0.81
	NB-MM	7.05	0.98
	NB-ML	5.74	0.99
3	Poisson	59.44	0.05
	NB-MM	15.18	0.76
	NB-ML	9.60	0.92
4	Poisson	47.77	0.00
	NB-MM	13.52	0.77
	NB-ML	9.35	0.89

This approach resulted in some improvement in the Poisson fits, but these were still much worse than either NB fit. In general, there were only marginal changes in the NB fits compared with estimating the NB parameters on the whole sample as in Table 2, although surprisingly these changes indicated slight reductions in the goodness of fit of the NB distributions.

Because of the marked improvement in statistical goodness of fit of the NB distributions over the Poisson distributions, it was decided to inspect

some of the derived NB frequency distributions more closely. Figures 1 through 4 show the observed and aggregated NB-ML frequency distributions for each of the four surveys. At least two features emerge from these figures. First, although the overall fits appear very good, the NB distribution tended to underestimate slightly the number of intervals with zero arrivals, which were largely caused by the cyclic interruption of the upstream signals. Second, the tails of the fitted distributions were much longer than those of the observed distributions. One reason for this latter feature is probably that, like the Poisson distribution, the NB distribution treats vehicles as if they were points. Physical constraints on the maximum number of vehicles able to pass a point in a given time interval are consequently unaccounted for. Such constraints arise, for instance, from finite vehicle lengths, quite apart from any inability or unwillingness on the part of drivers to minimize the gap between their vehicle and the one in front of them. Figures 3 and 4 for surveys 3 and 4, the higher-flow multilane surveys, also show that the NB distribution tended to underestimate the frequency of arrivals in the mid- to higher range of observed arrivals. These arrivals probably consisted of reasonably compact platoons, which are characteristic of flows downstream from a signal during peak periods. However, the NB distribution also tended to overestimate the frequency of low (but greater than zero) arrival intervals.

Finally, it was observed that the sample variance-to-mean ratios for the 10-s counts in each 15-min interval for each survey tended to vary with the mean flow rate. The following relationship was therefore fitted (with t -statistics in parentheses):

$$v/m = 1.203 + 0.0956 m^2 \quad R^2 = 0.81 \quad (7.34) \quad (12.13) \quad (16)$$

where m and v are as defined previously in Equations 5 and 6 and shown in Table 1 for each survey. The constant in Equation 16, 1.203, is statistically not significantly different from 1.0, the variance-to-mean ratio value for a Poisson process. Equation 16 may thus be interpreted to suggest that as the mean flow rate approaches zero, the limiting counting distribution is a Poisson distribution, whereas the

Figure 1. Aggregated frequency distribution: Survey 1.

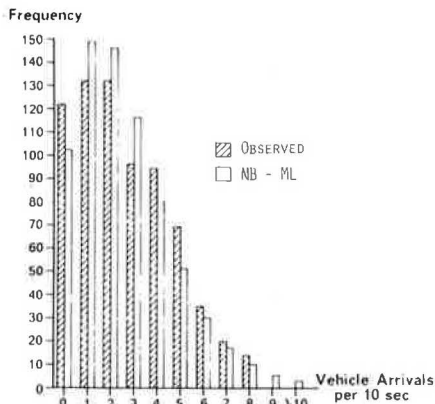
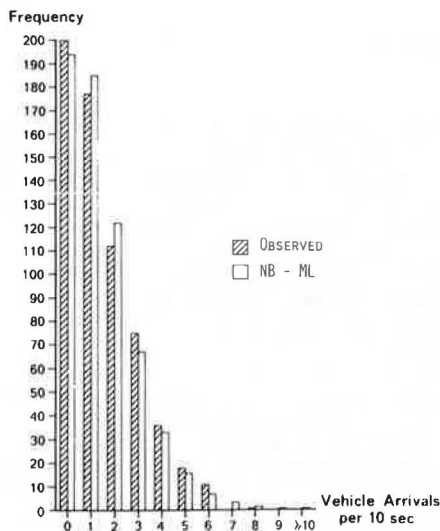


Figure 2. Aggregated frequency distribution: Survey 2.



departure from a Poisson process increases with the mean flow rate.

Although Equation 16 is based on results for the 15-min intervals in each survey, it was able to predict quite accurately (i.e., well within a 95 percent confidence interval) the whole-period variance-to-mean ratio for each survey. Clearly, if further traffic surveys under a broader range of geometric and traffic conditions, and counting intervals, than those in surveys 1-4 confirmed a simple functional relationship such as that in Equation 16, application of the NB distribution would be further simplified. In that case, the distribution parameters could be estimated (by using the MM procedure) solely from the mean flow rate, as for the Poisson distribution.

CONCLUSIONS

The analysis in this paper suggests that the NB distribution can be applied quite simply to commonly occurring problems involving high-variance traffic counts with results often markedly better than those for other elementary counting distributions, such as the Poisson distribution.

Several methods of fitting counting distributions to time-dependent peak-period counts were investi-

Figure 3. Aggregated frequency distribution: Survey 3.

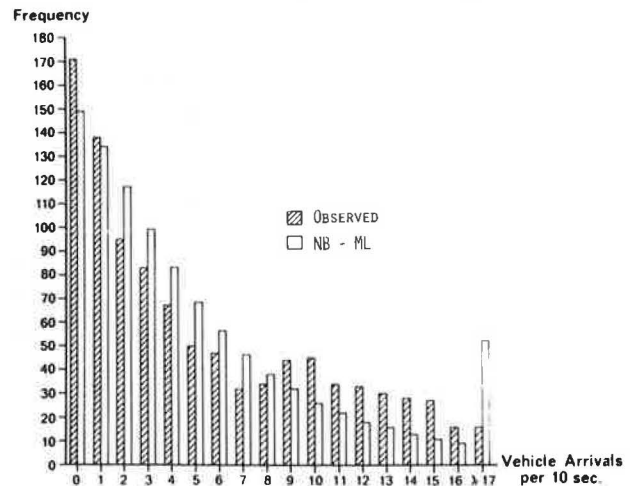
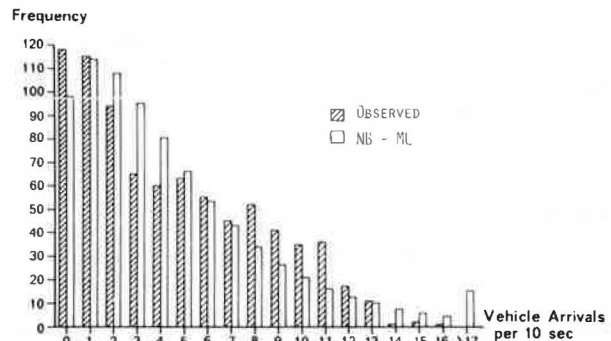


Figure 4. Aggregated frequency distribution: Survey 4.



gated, as well as alternative parameter-estimation techniques. From the results, it would appear that unless there is a particular reason to explicitly account for the temporal aspect of peak-period traffic demands, as may be the case in a traffic simulation study, the simpler method of estimating the NB distribution from the whole sample of peak-period traffic counts is to be preferred. Furthermore, from a practical viewpoint, the somewhat simpler MM parameter-estimation technique may often be preferred to the ML method, although the latter yields more efficient parameter estimates. In any case, it is clear that a relatively simple alternative to the Poisson distribution exists, which can yield a much better representation of high-variance, short-period urban traffic counts.

It is hoped that this paper might stimulate renewed interest in application of the NB counting distribution, particularly on the part of practitioners concerned with urban traffic systems analysis.

ACKNOWLEDGMENT

Much of the research reported in this paper was conducted while I was with the Transport Group, Department of Civil Engineering, Monash University, Australia. The encouragement and advice of M.A.P. Taylor, Commonwealth Scientific and Industrial Research Organization, Australia, in the early stages of this work is gratefully acknowledged.

REFERENCES

1. W.C. Jordan and M.A. Turnquist. Zone Scheduling of Bus Routes to Improve Service Reliability. *Transportation Science*, Vol. 13, No. 3, 1979, pp. 242-268.
2. M.D. Abkowitz. The Impact of Service Reliability on Work Travel Behavior. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA, Ph.D. dissertation, 1980.
3. J.N. Prashka. Development of a "Reliability of Travel Modes" Variable for Mode-Choice Models. Transportation Center, Northwestern Univ., Evanston, IL, Ph.D. dissertation, 1977.
4. R.E. Allsop. Delay at a Fixed Time Traffic Signal--I: Theoretical Analysis. *Transportation Science*, Vol. 6, 1972, pp. 260-285.
5. D.L. Gerlough and M.J. Huber. Traffic Flow Theory. TRB, Special Rept. 165, 1975.
6. G.F. Newell. Approximation Methods for Queues with Application to the Fixed-Cycle Traffic Light. *SIAM Review*, Vol. 7, 1965, pp. 223-237.
7. T.E.H. Williams and J. Emmerson. Traffic Volumes, Journey Times and Speeds by Moving-Observer Method. *Traffic Engineering and Control*, Vol. 3, 1961, pp. 159-162 and 168.
8. A.J. Miller. Settings for Fixed-Cycle Traffic Signals. Proc., Australian Road Research Board Conference, Vol. 2, No. 1, 1964, pp. 342-362.
9. G.E.P. Box and G.M. Jenkins. Time Series Analysis: Forecasting and Control. Holden-Day, New York, 1970.
10. N.W. Polhemus. On the Statistical Analysis of Non-Homogenous Traffic Streams. *Transportation Planning and Technology*, Vol. 6, 1980, pp. 45-54.
11. D.L. Gerlough and F.C. Barnes. Poisson and Other Distributions in Traffic. Eno Foundation for Transportation, Saugatuck, CT, 1971.
12. S.G. Ritchie. Bus and Car Pool Lanes on Arterial Roads--A Simulation Study. Proc., Transportation Research Forum, 21st Annual Meeting, Philadelphia, Oct. 1980, Richard B. Cross Co., Oxford, IN, 1980, pp. 280-287.
13. D.J. Buckley. Road Traffic Counting Distributions. *Transportation Research*, Vol. 1, 1967, pp. 105-116.
14. E. Williamson and M.H. Bretherton. Tables of the Negative Binomial Probability Distribution. Wiley, London, 1963.
15. J.R. Benjamin and C.A. Cornell. Probability, Statistics, and Decision for Civil Engineers. McGraw-Hill, New York, 1970.
16. E. Kreyszig. Advanced Engineering Mathematics, 4th ed. Wiley, New York, 1979.
17. P.G. Hoel. Introduction to Mathematical Statistics. Wiley, New York, 1966.
18. A. Karlquist and B. Marsjo. Statistical Urban Models. *Environment and Planning*, Vol. 3, 1971, pp. 83-98.
19. F.J. Anscombe. Sampling Theory of the Negative Binomial and Logarithmic Series Distributions. *Biometrika*, Vol. 37, 1950.

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

Delay Models of Traffic-Actuated Signal Controls

FENG-BOR LIN AND FARROKH MAZDEYASNA

Traffic-actuated signal controls have more control variables for engineers to deal with than a pretimed control. The increased sophistication in their control logic provides greater flexibilities in signal control but also makes the evaluation of their performance more difficult. At the heart of the problem is that traffic delays cannot be readily related to the control variables of a traffic-actuated control. This prompts practicing engineers to rely mostly on short-term, subjective field observations for evaluation purposes. To provide an improved capability for evaluating alternative timing settings, delay models are developed in this study for semiactuated and full-actuated controls that employ motion detectors and sequential phasing. These models are based on a modified version of Webster's formula. The modifications include the use of average cycle length, average green duration, and two coefficients of sensitivity that reflect the degree of sensitivity of delay to a given combination of traffic and control conditions. Average cycle length and average green duration are dependent on the settings of the control variables and the flow pattern at an intersection. They can be estimated by existing methods.

Traffic-actuated controls employ relatively complex logic to regulate traffic flows. This type of logic infuses a much-needed flexibility into signal control, but it also makes the performance evaluation of a traffic-actuated control difficult. A major problem is that traffic delays resulting from such a control cannot be readily related to the settings of the control variables and the flow pattern at an intersection.

Current understanding of traffic delays at a traffic-actuated signal is obtained only through sensitivity analyses with the aid of computer simu-

lation models. Tarnoff and Parsonson (1) have provided a detailed review of the findings of these simulation studies. Computer simulation models, however, have significant limitations.

For one thing, practicing engineers may not be familiar with the nature and the capability of such models. Furthermore, to ensure broad applicabilities, simulation models are often difficult to use in terms of data needs, requirements of computer facilities, and the time one has to spend to learn how to use them. As a result, practicing engineers still rely mostly on short-term, subjective field observations in evaluating timing settings.

An alternative to the use of computer simulation is to develop a model in the form of a formula or a set of formulas. Such a model would allow expedient evaluation of a large number of alternatives and would be particularly useful in searching for optimal ways of using a signal control. This in turn could encourage practicing engineers to improve the efficiency of existing signal controls.

To partly satisfy this need, this paper presents a set of delay models for semiactuated controls and full-actuated controls that employ motion detectors. These delay models are calibrated in terms of simulation data. They are applicable to signal controls at individual intersections when single-ring, sequential phasing is used. The traffic flows con-

sidered in this study include only straight-through passenger vehicles. Other types of vehicles can be transformed into equivalent straight-through passenger vehicles for analysis (2).

GENERAL FORMULATION

The delay models assume the following form:

$$D = 0.9 \{ [C(1 - Ax)^2 / 2(1 - AxBy)] + [3600 (By)^2 / 2Q(1 - By)] \} \quad (1)$$

where

- D = average delay (s/vehicle),
- C = average cycle length (s),
- x = ratio of effective green to average cycle length = G_e / C ,
- G_e = effective green = $G + Y - k$ (s),
- G = average green duration (s),
- Y = yellow duration (s),
- k = loss time per phase,
- Q = traffic volume in a lane (vph),
- y = saturation ratio = $QC / (Q_s G_e)$, and
- Q_s = saturation flow rate, which is approximately 1800 vph of effective green.

Equation 1 is a modified version of Webster's formula (3). Its unique feature is the inclusion of the two coefficients of sensitivity, A and B. The reason for including these coefficients is that a vehicle subjected to a traffic-actuated control can exert an influence in the transfer of right-of-way. Consequently, traffic-actuated delays and pretimed delays can be expected to have different sensitivities to both the ratio of effective green to cycle length and the saturation ratio. A larger value of A represents a lesser degree of sensitivity of the average delay to the ratio of effective green to cycle length (x). In contrast, a larger value of B implies a higher degree of sensitivity of the average delay to the saturation ratio (y).

SIMULATION MODEL

The simulation model used in this study comprises a flow processor and a signal processor. The flow processor generates vehicle arrivals and processes the vehicles through an intersection according to the signal indications. The speed, location, and acceleration rate of each vehicle are updated by this processor once every second. The operation of this processor is a function of the signal indications. It is independent of the type of signal control. Based on the control logic of a given type of signal control, the signal processor uses flow data provided by the flow processor to determine the signal indications for each 1-s scanning interval.

Under various combinations of flow and pretimed signal settings, average delays obtained from the simulation model are generally within 15 percent of the values estimated from Webster's formula. This indicates that the flow processor can move vehicles downstream in a reasonably reliable manner. The average greens of individual phases of semiactuated controls and full-actuated controls as generated from the signal processor are found to be within 2-3 s of estimates obtained from analytical models (4,5). In addition, the simulated delays for a full-actuated control under near-optimal control conditions agree very well with estimates obtained by Morris and Pak-Poy (6). These tests do not constitute a complete validation of the simulation model. Nevertheless, they confirm the ability of the simulation model to produce satisfactory data.

DELAY MODELS

Semiactuated controls may be used when a lightly traveled side street intersects a major street. Detectors are installed on the side street to collect flow data for making signal-timing decisions. When motion detectors are used, the key control variables of this type of control usually include minimum green (G_{min}) for the major street and initial portion (I), unit extension (U), maximum allowable green (G_{max}), and detector setback (S) for the side street. Without vehicle actuation of the detectors, the green light is always given to the major street.

In contrast, full-actuated controls require the use of detectors on all approaches that are subjected to signal control. When motion detectors are used, the duration of each green phase in a given cycle is governed by the same set of control variables, which usually includes I, U, G_{max} , and S as defined previously.

To calibrate Equation 1 for either type of the control, three levels of the initial portion were considered: 5, 8.5, and 12.5 s. At each level more than 90 different combinations of flow conditions and settings of the control variables were examined through computer simulation. To avoid unnecessary complications, only two-phase controls were dealt with. Each phase involves two traffic lanes with equal or unequal volumes. The unit extension is confined to a value between 3 and 6 s, and the detector setback is from 50 to 120 ft.

Under semiactuated controls, the minimum green for the major street varies from 20 to 50 s and the maximum green for the side street is limited to 30 s.

For each combination of the flow and signal control conditions, the simulation model generates average delays, average cycle length, and average green duration related to each signal phase. The generated data were used in Equation 1 to determine the combination of A and B that best duplicates the simulated delays.

For semiactuated controls, the data generated for the side-street traffic were analyzed separately from those for the major-street traffic. The reason for this is that semiactuated controls are only responsive to the side-street vehicles and thus the side-street delays and the major-street delays are likely to be affected by the controls in different ways.

The results of the model calibration are shown in Figures 1 and 2 for semiactuated controls and in Figure 3 for full-actuated controls. Figure 1 shows that for semiactuated control, both A and B for the side-street delays are greater than or equal to 1.0. This indicates that the side-street delays are less sensitive to the ratio of effective green to cycle length (x) and more sensitive to the saturation ratio (y) than pretimed delays. Since B decreases with respect to the initial portion, it can also be said that a shorter initial portion gives rise to a greater sensitivity of the delays to the saturation ratio. In other words, when a short initial portion is used, the average side-street delay increases rapidly with respect to the saturation ratio. When the G/G_{max} ratio increases, however, A and B approach 1.0. This implies a convergence of semiactuated controls to pretimed controls.

For the major-street traffic, A has a constant value of 1.0 (Figure 2) as in the case of a pretimed control. The value of B increases with the initial portion used for the side-street traffic. Therefore, a longer initial portion for the side street could cause the major-street delays to rise rapidly with the saturation ratio. The value of B for the major-street traffic also varies with the ratio of average

Figure 1. Calibrated values of A and B for side-street traffic under semiactuated control.

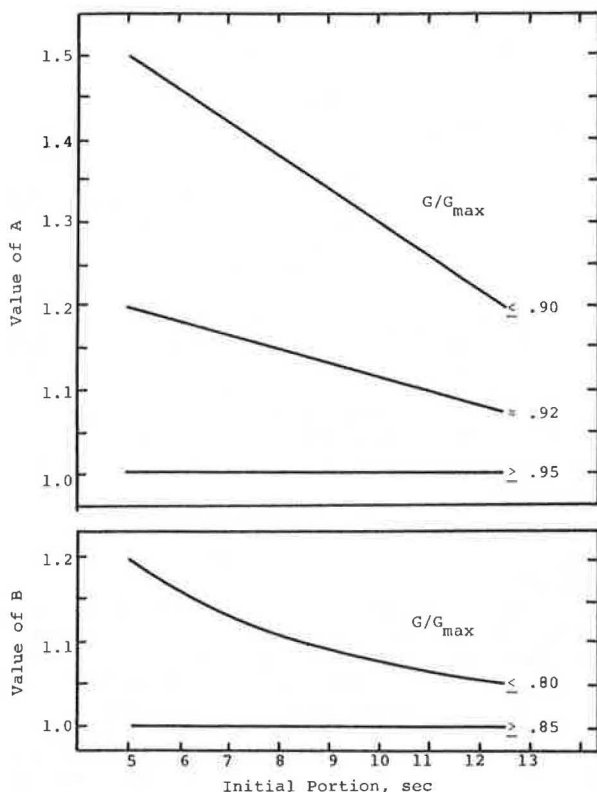
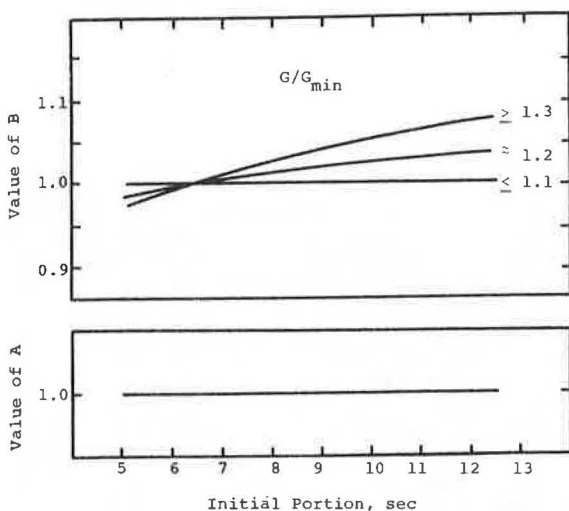


Figure 2. Calibrated values of A and B for major-street traffic under semiactuated control.

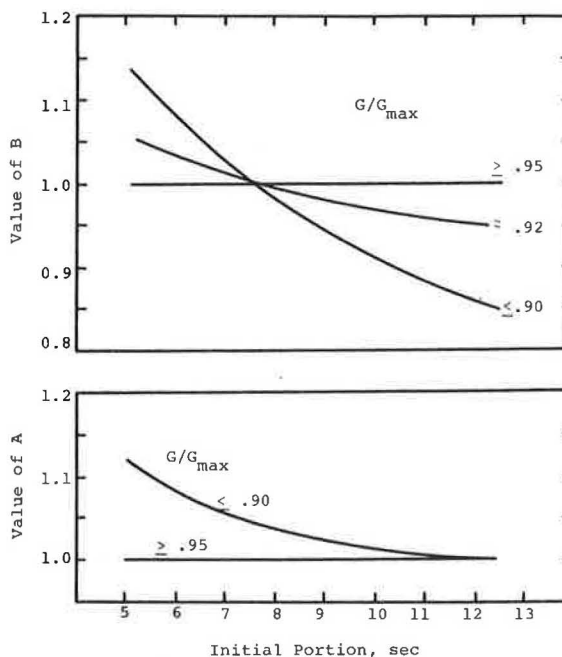


green to minimum green (G/G_{\min}). When this ratio is less than 1.1, the variation of the major-street green from one cycle to another is small and B approaches 1.0.

Under full-actuated controls, the value of A is at least as large as 1.0, and it can be concluded that average delays resulting from a full-actuated control are less sensitive to the G_e/C ratio than pretimed delays.

The value of B exceeds 1.0 when the initial portion is below 7.5 s and drops under 1.0 at a longer initial portion. This feature signifies that the

Figure 3. Calibrated values of A and B for full-actuated control.



sensitivity of the average full-actuated delays to the saturation ratio increases when the initial portion decreases. It follows logically that a heavier flow should be given a longer initial portion. Also, the average full-actuated delays are not so adversely affected by the saturation ratio as pretimed delays when the initial portion is greater than 7.5 s. The advantage of full-actuated controls, however, disappears when the G/G_{\max} ratio is in excess of approximately 0.95. At this level of the ratio, full-actuated controls behave more or less like pretimed controls.

To be useful as a tool for evaluating alternative timing settings and detector setbacks, Equation 1 has to be used in conjunction with a method for estimating G and C. A reasonable option is to use the methods presented by Lin (4,5) in two recent studies. These methods relate the average green of a signal phase of a semiactuated control or a full-actuated control to the control variables and the traffic flow pattern at an intersection. The formulations of the methods are not simple because of the complex logic of the traffic-actuated controls. Nevertheless, the methods can be applied manually or be implemented in the form of short computer programs with about 60 FORTRAN statements for semiactuated controls and about 100 statements for full-actuated controls.

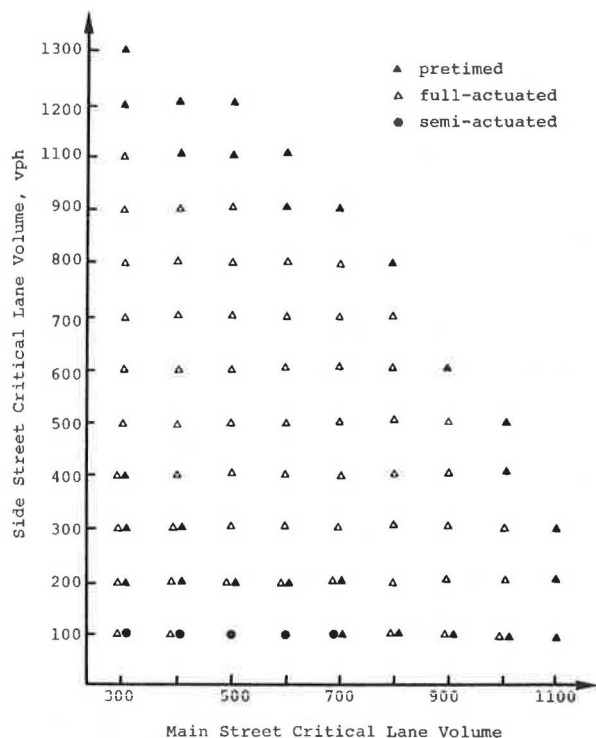
It should be noted that, in their existing forms, these estimation methods are generally applicable only when a unit extension of greater than approximately 3-3.5 s is used. With a shorter unit extension, premature termination of green phases is quite likely and the methods will require modifications.

Based on these methods for estimating G and C, the average delays obtained from the delay models agree reasonably well with the values generated from the simulation model. The differences are within 3 s in more than 85 percent of the cases examined in this study.

APPLICATIONS

The primary applications of the delay models are in

Figure 4. Domains of preferred types of two-phase signal control derived from delay models and Equation 1.



the evaluation of alternative timing settings. By using the models as a tool for sensitivity analysis, one can examine relatively easily how changes in the timing settings and detector setbacks may affect the efficiency of a signal control. The models are particularly useful for searching for optimal controls when a microcomputer is available to implement the methods for estimating G and C .

The delay models can also be used to assist in the selection of alternative types of signal control. For example, by using the delay models and Webster's formula, one can obtain Figure 4. This figure shows the most efficient types of signal control along pretimed, semiactuated, and full-actuated controls for various combinations of flows.

CONCLUSIONS

Use of simulation models for evaluating a large number of alternative timing settings and detector setbacks is usually cumbersome and requires substantial resources. The delay models described in this paper provide a more efficient alternative. These models, however, are applicable only when motion detectors are used and when sequential phasing is in effect. Similar models may also be developed for other types of traffic-actuated control. The availability of such simplified models could encourage practicing engineers to make an effort to improve existing controls.

Discussion

Kenneth G. Courage

Within the scope stated by the authors, the study

appears to be sound. The methodology is scientific and the results are reasonable.

The authors propose an extension to Webster's delay model to deal explicitly with certain operating parameters (minimum and maximum green) for a traffic-actuated controller. They suggest that a model of this form is preferable to existing models that treat these parameters implicitly. The results of the proposed model are not compared with those of the existing models. Such a comparison would have made the results more credible. Greater credibility might also have been achieved by starting with the TRANSYT modification to Webster's model, which deals with oversaturated as well as undersaturated operation.

The applicability of these results is constrained by the scope of the study, which was limited to exclude volume density operation and presence detection on the approaches. These features are both very common and both have been shown to produce a more efficient signal operation than the conventional actuated controller with motion detection.

Another limiting factor in the results is that the coefficients A and B are shown to vary with the parameter G/G_{\max} . In other words, the optimal setting of the operating parameters varies with traffic volume; therefore no permanent controller settings can be developed by using the proposed model. This variation has been recognized in the past and was the primary motivation behind the development of the volume density controller.

It must be recognized that the most successful modes of locally actuated intersection control are based on intuitive mechanical models. These models are primitive and they defy purely analytical treatment. Their popularity is derived from the fact that they can be fully implemented on the street, whereas theoretical models, such as the one discussed in this paper, cannot.

Authors' Closure

The operation of the pulse-mode traffic-actuated control has not been properly modeled and analyzed in the past. A comprehensive discussion of this issue is not appropriate for this closure. Nevertheless, an existing model discussed by Courage and Papapanou (7) can be used for a short discussion. This earlier model is based on a control strategy that has the following characteristics: (a) it distributes available green time in proportion to demand on critical approaches, and (b) it minimizes wasted time by terminating each green interval as soon as the queue of vehicles has been properly serviced. We indicated that this control strategy closely approximates the operation of the traditional traffic-actuated controller that has been properly timed and that the delay estimates will therefore reflect the best operations that can be expected from traffic-actuated control. The validity of these claims aside, this earlier model is aimed at estimating the minimum delays. In contrast, the proposed model provides a mechanism for estimating delays under various combinations of traffic and control conditions.

Furthermore, a close examination of the earlier model will reveal that the only control variable considered in the model is the maximum green of a signal phase. Unfortunately, this variable is of concern only under very limited conditions. The model also uses average cycle length (C_a), which

is calculated as $C_a = L/(1.0 - Y)$, where L is the total loss time per cycle and Y is the overall degree of saturation of critical movements. Again, under a traffic-actuated control, the average cycle length cannot be adequately estimated from such an equation. The result is a model that has little to do with the actual timing settings and detector setback.

In short, the proposed and the earlier models have distinct characteristics. A comparison of the two models is really meaningless and will not make the proposed model either more or less credible.

The discussant suggested that the simulation model used to develop the delay models should have been calibrated on the basis of the TRANSYT-7F model (8) rather than on the basis of Webster's delay model (3). The contention was that Webster's model gives unrealistically high estimates of delay when the saturation ratio approaches or exceeds 1.0 (Figure 5). Such a contention reflects a general lack of understanding, not only about the nature of Webster's model, but also about the flow characteristics at saturation ratios near or exceeding 1.0.

Figure 5. Comparison of Webster's model with TRANSYT-7F model.

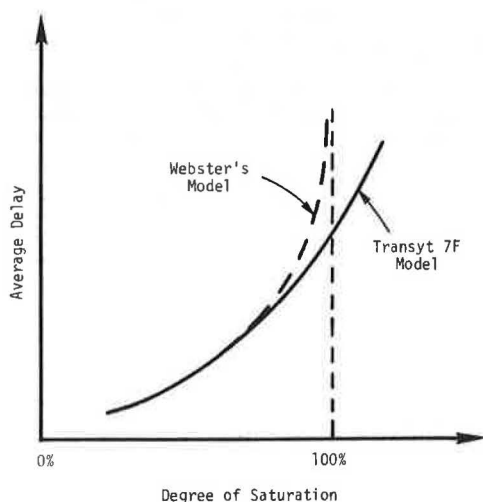
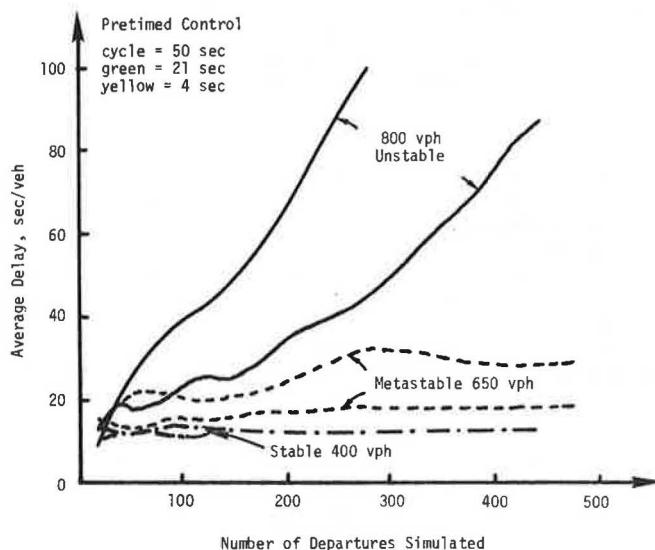


Figure 6. Stability of performance of signal control.



To resolve this issue, it is necessary to point out that in terms of delays, the operation of a signal system can be classified into the following states:

1. Stable state: In this state the average delay of a flow is primarily a function of the flow rate. Variations in the delays from one field observation or simulation run to another are small as long as the flow rate and the control conditions remain unchanged.

2. Metastable state: In this state the average delay depends not only on the flow rate but also on the sequence of the arriving headways. If the same flow rate persists, a stable value of the average delay can still be obtained. However, the variations in the delays become substantial at a given flow rate.

3. Unstable state: In this state the average delay depends not only on the flow rate and the sequence of the arriving headways but also on the time period in which a given flow rate persists. In other words, the average delay is time dependent. The longer the flow rate persists, the higher the average delay becomes.

The existence of these states can be identified from computer simulation. Figure 6 shows an example. There are no clear-cut boundaries between the various states. For pretimed control, the metastable state may arise when the saturation ratio is in the range of 0.8-0.9; the unstable state may emerge when the ratio is about 0.9 or greater.

Webster's model gives the estimated delay for a flow that persists indefinitely. Naturally, when the saturation ratio approaches 1.0, the estimated delay approaches infinity. If this nature of the model is not recognized, the comparison between Webster's model and the TRANSYT-7F model is just like the comparison between apples and oranges.

Since in the real world a flow will never persist long enough to induce an infinite average delay, the TRANSYT-7F model attempts to account for this fact by using a delay function with finite delay values (Figure 5). In so doing, it only gives one a false sense of security. The reason is that at high saturation flow rates, delay is time dependent. Therefore, in a finite time frame, the single delay function of the TRANSYT-7F model should be replaced with a set of delay functions, each of which is associated with a given time period of signal operation. In reality, this is extremely difficult if not impossible to accomplish.

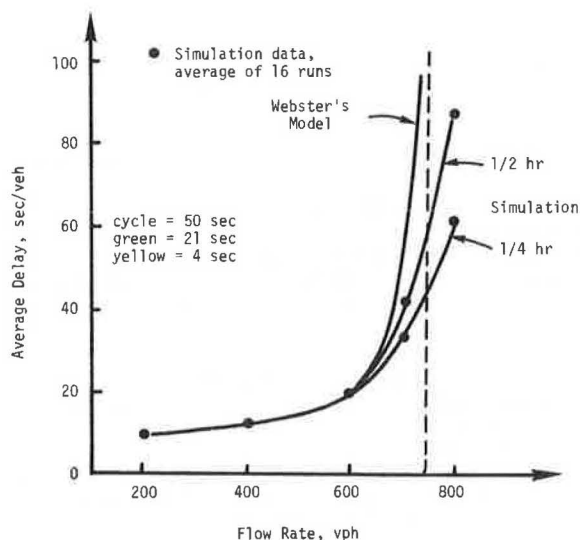
In the absence of better information, Webster's model is a reasonable basis for calibrating a simulation model. The calibration, of course, should be based on stable or metastable operations. A simulation model calibrated in this manner will not result in unrealistically high average delays over a finite time period. Figure 7 illustrates this feature.

In summary, the TRANSYT-7F model does not have real advantages over Webster's model. Significant improvements can be made only if the delay function at high saturation ratios can be explicitly and reliably related to time.

We recognized that the proposed models are not applicable to volume-density operation and presence-mode operation. However, the same methodology can be employed to develop a model for either one of such operations.

The discussant seemed to object to the fact that the model showed that the optimal settings of the operating parameters varied with traffic volume. Such variations not only are inherent to pulse-mode operation but also exist in presence-mode operation and volume-density operation. In fact, as long as a signal control requires predetermined settings, it

Figure 7. Comparison of simulated delays with Webster's delay under pre-timed control.



will not always result in an optimal control under varying traffic-flow conditions. Precisely for this reason, it is desirable to have a reasonably simple yet reliable model to determine the trade-off of timing settings with respect to a typical daily flow pattern. Such a trade-off analysis would enable one to select permanent settings. The use of volume-density control alleviates but does not eliminate this problem.

It would certainly be a blessing if the operation

of a signal control could be adequately represented by a primitive and intuitive model. In reality, such intuitive models for analyzing traffic-actuated controls are often misleading and are usually not better than practicing engineers' intuitive judgments. In anticipation of increased use of microcomputers, there is room for developing models that are more reliable than intuitive models but less difficult to use than most existing simulation models.

REFERENCES

1. P.J. Tarnoff and P.S. Parsonson. Selecting Traffic Signal Control at Individual Intersections. NCHRP, Rept. 233, 1981.
2. Interim Materials on Highway Capacity. TRB, Transportation Research Circular 212, Jan. 1980.
3. F.V. Webster. Traffic Signal Settings. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, TRRL Rept. 39, 1961.
4. F.B. Lin. Predictive Models of Traffic-Actuated Cycle Splits. Transportation Research, Vol. 16, Part B, No. 5, 1982, pp. 361-372.
5. F.B. Lin. Estimation of Average Phase Duration of Full-Actuated Signals. TRB, Transportation Research Record 881, 1982, pp. 65-72.
6. R.W.T. Morris and P.G. Pak-Poy. Intersection Control by Vehicle-Actuated Signals. Traffic Engineering and Control, Oct. 1967, pp. 288-293.
7. K.G. Courage and P.P. Papapanou. Estimation of Delay at Traffic-Actuated Signals. TRB, Transportation Research Record 630, 1977, pp. 17-21.
8. TRANSYT-7F User's Manual. FHWA, Dec. 1981.

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

Another Look at Bandwidth Maximization

KARSTEN G. BAASS

One solution to the problem of fixed-time traffic signal coordination is the provision of a large green band that allows road users to drive at a reasonable speed without stopping. This solution is popular with drivers, although it does not necessarily lead to delay minimization except in special cases. A method for deriving the globally maximal bandwidth together with all possible suboptimal values is described. The programs WAVE1 and WAVE2 can also be used to generate curves that show the continuous relation between uniform progression speed and corresponding maximal bandwidth over a wide range of speeds and cycles. The typical shape of this bandwidth-speed relationship is explained theoretically, and the theory is used in the development of the algorithm. It is shown that bandwidth varies greatly with progression speed and it is suggested that setting bandwidth at the globally optimal value may not always be the best choice. The decision to adopt a progression speed, a bandwidth, and a cycle time should take into account a range of values of speed and cycle. The proposed method was applied to 18 data sets of up to 24 intersections taken from the published literature and the results obtained were compared with those given by the mixed-integer linear-programming approach. Computer execution time is extremely short and the storage space required is negligible, so the method could be of interest in practical applications.

The maximization of bandwidth is one of the two approaches used for determining offsets between fixed-time traffic lights on an artery. There are a number of fairly restrictive hypotheses related to this approach, e.g., the assumptions of a uniform pla-

toon, no platoon dispersion, low volumes, and no or very few cars entering the artery from side streets. Situations corresponding to these assumptions are rare. Nevertheless, the bandwidth-maximizing approach is psychologically attractive to the user, who is unable to distinguish between a non-synchronized artery and one that is perfectly synchronized for delay and stop minimization but does not allow the user to pass at a reasonable speed through the artery without stopping.

Little and others (1) and Morgan (2) were the first to suggest a mathematical formulation for the bandwidth-maximizing problem, and more recently Little and others (3) published a program called MAXBAND. This program is based on a mixed-integer linear-programming approach and determines the speeds that give the overall maximum bandwidth over a range of acceptable speeds. The linear-programming approach also allows for variations in speed between intersections and enables new constraints to be easily introduced.

This paper describes an algorithm that determines the overall maximum bandwidth together with all sub-optimal values, if they exist, for a wide range of speeds. At this time, only two-phase fixed-time

traffic light operation is considered and a constant speed on the artery is assumed, but the algorithm could be generalized to changing speeds between intersections.

The solution approach is essentially geometric, and the algorithm is extremely fast for arteries of up to 25 intersections. The limit of 25 is not due to the algorithm nor to computing time but was set because this number is quite high enough for practical synchronization problems. The approach also provides insights into the theoretical relationships and may explain why the bandwidth approach works better in certain cases than in others.

THEORETICAL CONSIDERATIONS

Morgan (2) has shown that for a given speed there is a set of half-integer offsets that gives maximal equal bandwidths. It is shown also that one can derive another optimal solution with unequal speeds and bands from this initial solution. The following discussion will thus pertain only to equal maximal bandwidths.

Consider first one of the possible half-integer sets of offsets, represented for simplicity on a time-space diagram as in Figure 1 (offset scheme 0-1-0-0). For the present, we consider only speeds between a minimum and a maximum speed, as shown in Figure 1 by the dotted and interrupted lines. In the following, when necessary, the constant $K = V \cdot C$

is used instead of a fixed speed and a fixed cycle. This relation is given by a simple scale transformation of the time-space diagram. Figure 1 shows that there are two extremal values of speed possible. The first (V_d) occurs when the speed line is tangential to two of the lower reds and the second (V_u) when the speed line is tangential to two of the upper reds. These two speeds will be called possible optimal speeds. They give rise to two parallel bands of different bandwidths as shown in Figures 2 and 3.

As progression speed decreases in Figure 2, the bandwidth decreases until a speed is reached where bandwidth is zero. As speed increases in Figure 3 (or the slope of the speed line decreases), the bandwidth will ultimately become zero if the speed does not reach infinity first. These relations can be expressed by the following equations. For the case in which the speed line is tangential to lower reds and i, j are the critical lights, the slope of the speed line is given by

$$s = 360/V_d \cdot C = (d_j - d_i)/(x_j - x_i)$$

$$d_m = d_i + s(x_m - x_i) \quad (1)$$

and the bandwidth at all m intersections is given by

$$\begin{aligned} b_{md} &= (r_m/2) + \text{INT} \cdot 50 + g_m - d_m \\ b_{md} &= \text{INT} \cdot 50 - [(r_m + r_i)/2] - s(x_m - x_i) \end{aligned} \quad (2)$$

Figure 1. Geometry of a progression.

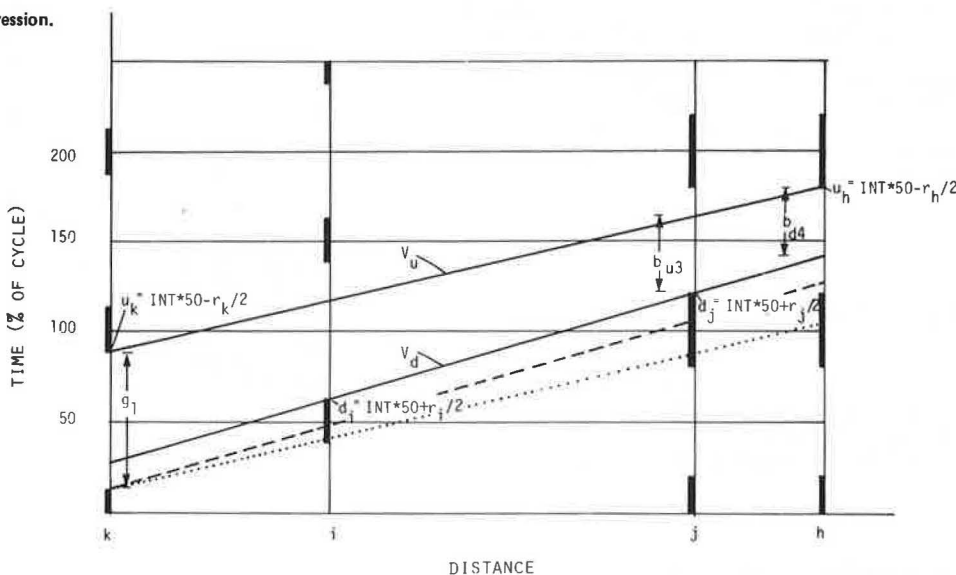


Figure 2. Band corresponding to speed V_d in Figure 1.

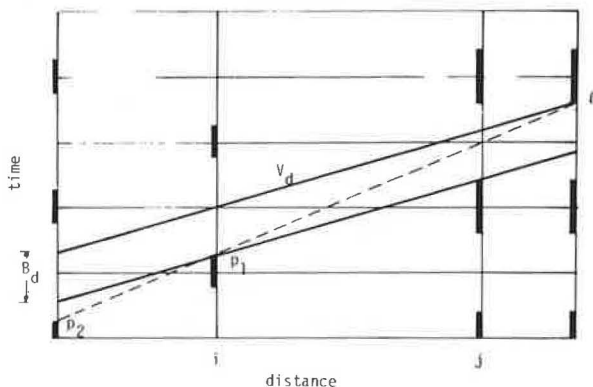
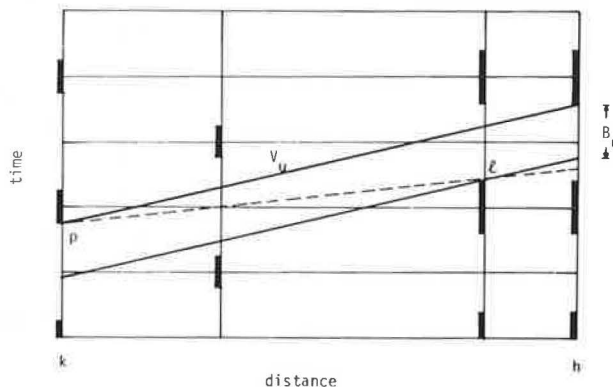


Figure 3. Band corresponding to speed V_u in Figure 1.



where

x_m = cumulative distance from first intersection,
 d_m = ordinate of upper edge of lower red,
 u_m = ordinate of lower edge of upper red,
 r_m = percentage of red at intersection m ,
 g_m = percentage of green at intersection m , and
 C = cycle length.

INT is an integer such that

$$0 \leq b_{md} \leq g_m$$

where b_{md} is as large as possible within these limits. The maximal bandwidth becomes

$$B_d = \min \{b_{md}\}$$

For the case in which the speed line is tangential to the upper reds and h and k are the critical lights,

$$V_u = (360/C) [(x_h - x_k)/(u_h - u_k)] \quad (3)$$

$$b_{mu} = \text{INT} \cdot 50 - [(r_m + r_k)/2] + s(x_m - x_k) \quad (4)$$

$$B_u = \min \{b_{mu}\}$$

The speed line can now be pivoted about p_1 in Figure 2 and speeds and corresponding bandwidths up to a bandwidth of zero can be determined. Clearly, the upper red that limits the bandwidth will have to be found, and it will in certain cases also happen that the speed line touches a red for $m < p$ before $B = 0.0$ is obtained. In this case, the pivot will have to be changed. The formula that gives the bandwidth for a known pivot point and a known limiting red becomes

$$B_d = \text{INT} \cdot 50 - [(r_p + r_q)/2] + (360/VC)(x_p - x_q) \quad (5)$$

$$B_u = \text{INT} \cdot 50 - [(r_p + r_q)/2] - (360/VC)(x_p - x_q) \quad (6)$$

where p is the number of the pivot intersection and q is the number of the limiting intersection.

We now consider an artery of four intersections in Laval, Quebec, as an example to illustrate these

equations. The data are given below and the cycle time is 80 s:

Artery	Distance (m)	Red (%)
1	0.00	25
2	297.18	24
3	803.15	40
4	987.55	40

Figure 5. Possible speed bands from intersection i for speeds between V_{\max} and V_{\min} .

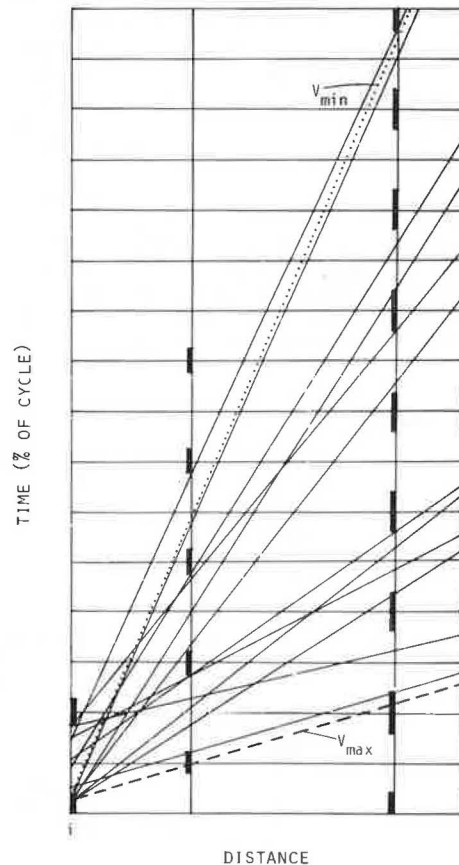


Figure 4. Curve relating speed to bandwidth for offset scheme and speeds V_u and V_d shown in Figure 1 (Laval example).

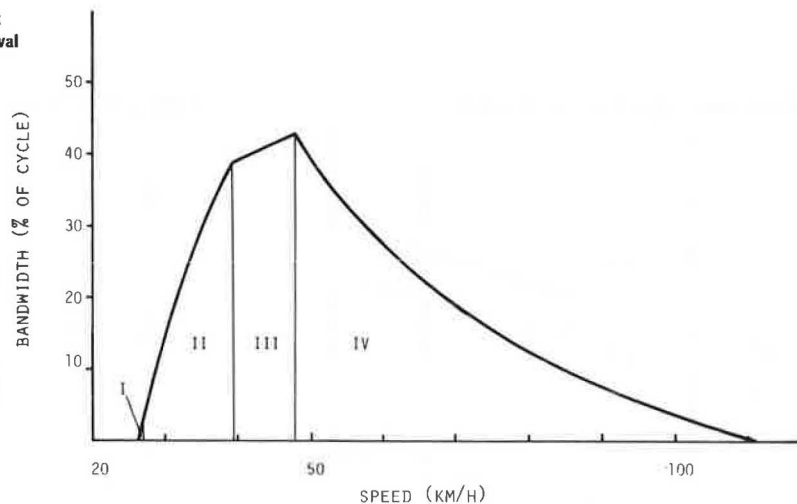


Figure 6. Curves relating speed to bandwidth for all possible offset schemes for Laval example ($n = 4$).

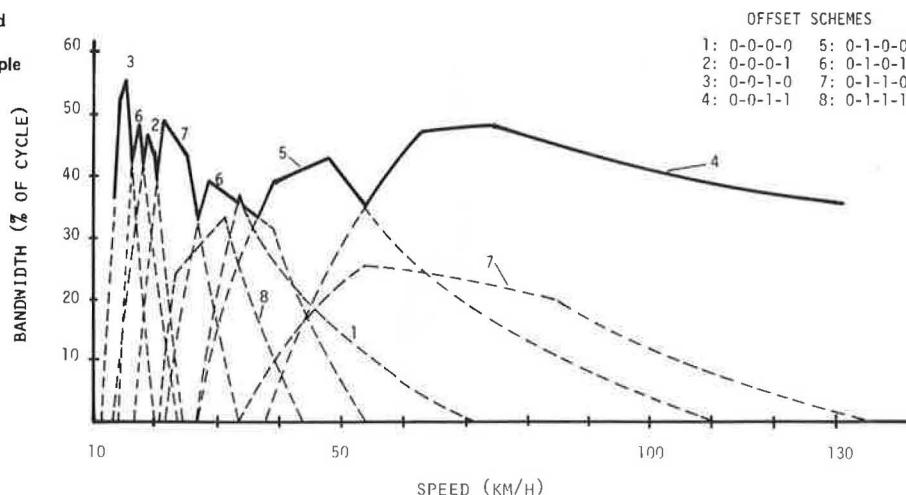


Figure 7. Example of V-B curve with major oscillations at lower speeds (data set 2, 10 intersections, 52 percent minimum green).

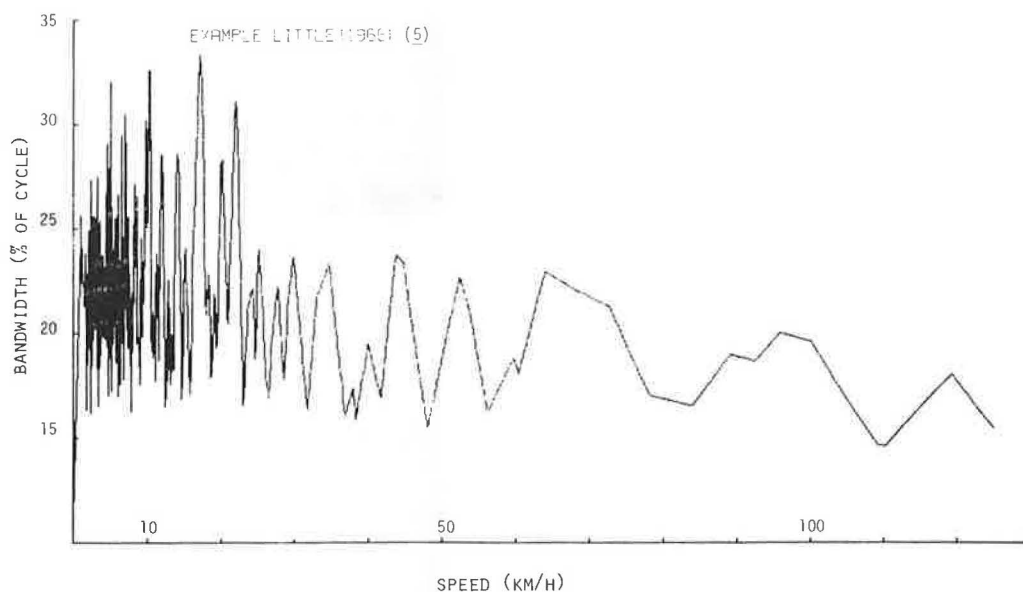


Figure 1 shows this artery and from Equations 5 and 6 the relationships tabulated below can be derived:

V_1	V_2	Bandwidth
26.53	27.02	I: $B = \text{INT} \cdot 50 - 32.5 - 4443.975/V$
27.02	39.26	II: $B = \text{INT} \cdot 50 - 32.0 - 3106.665/V$
39.26	48.04	III: $B = \text{INT} \cdot 50 - 40.0 - 829.800/V$
48.04	111.21	IV: $B = \text{INT} \cdot 50 - 32.5 + 3614.175/V$

A pivot change from intersection 2 to 1 has to be taken into account at a speed of 27.02 km/h. The equations above give the relation between B and V (or $K = V \cdot C$ and B) as shown in Figure 4. This curve is typical; it has a concave part up to the lower optimal speed and a convex section for speeds greater than the highest optimal speed. There is only one optimal speed in the case in which the critical lights for V_d and V_u are the same and have the same amount of green.

If speeds are allowed to vary to a larger extent, other speed bands will become possible for the same set of offsets shown in Figure 1. This is depicted in Figure 5. Each of these speed bands will produce a V-B curve similar to the one in Figure 4. There are $2^{**}(n - 1)$ possible half-integer sets of offsets

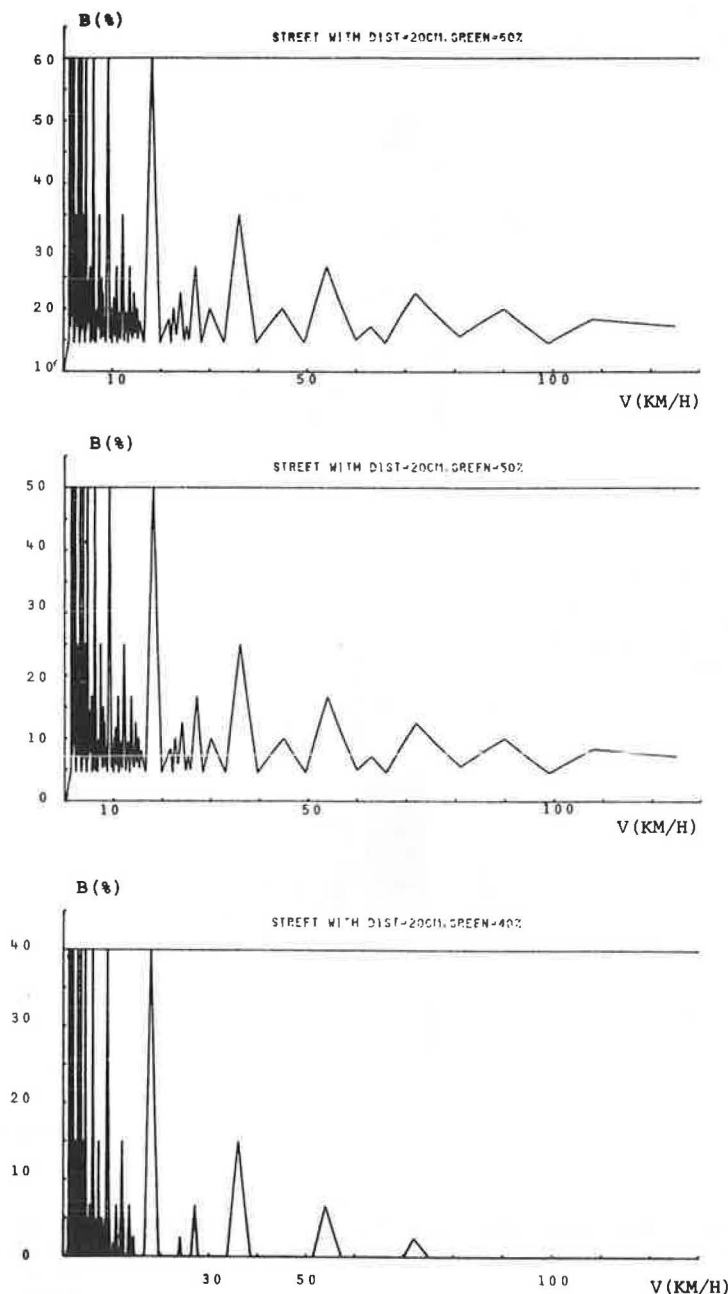
to be considered in this way, and the resulting curves of speed against bandwidth can be drawn together in one graph as shown in Figure 6. The envelope of these curves will be the curve of maximal bandwidth for each and every speed for a given cycle length. And since $V \cdot C$ is a constant, one can also derive all possible combinations of V and C that have the same bandwidth by using this envelope curve.

The extremal point of this curve between a minimum and a maximum speed is the same as the one obtained by the linear-programming approach of MAXBAND for the special case of uniform speeds. Some interesting remarks can be made by analyzing different V-B envelope curves.

1. There are many optimal values at low speeds and less at higher speeds as in Figure 7. This will be explained later by a simple formula, but it is also intuitively clear from Figure 5.

2. For the special case of equal distances between intersections and equal green times, maximum bandwidths corresponding to the minimum green may not be obtained at reasonable speeds. As the green time is decreased (as in Figure 8) from 60 percent to 50 percent and to 40 percent, the same envelope curve is displaced on the ordinate by 10 percent.

Figure 8. Comparison of speed-bandwidth curves for different minimum greens and equal distances between intersections.



Increasing the distance between lights (four times, for example) does not change the envelope curve. The range of speeds in Figure 9 (top) between 0 and 30 km/h is merely stretched four times to speeds between 0 and 120 km/h. This entails an increase in oscillations of the V-B curve.

3. As average distance between intersections increases, the envelope curve becomes more unstable. The stability of the band with respect to speed decreases. Figure 6 gives an example of a relatively stable situation at least between speeds of 60-90 km/h with an 80-s cycle. Stability should also be taken into account in the choice of a band and progression speed. If, for example, in Figure 9 (bottom) an optimal speed of 36 km/h and a band of 50 percent is chosen and speed increases or decreases by only 2 km/h, the bandwidth will fall to only 5 percent. This may explain why certain bands are apparently very inefficient as volume increases slightly; a slight reduction in speed is produced,

but there is a major decrease in bandwidth. It may not always be good to adopt the extremal value of the V-B curve if this value is on a steep part of the envelope curve.

4. As the number of intersections increases, the envelope curve becomes more and more unstable, as is shown in Figure 10. In this case, oscillations increase and attainable bandwidth is small.

Clearly, the discussion up to now is purely theoretical and certainly impractical, since it is impossible to enumerate all possible combinations of half-integer offsets in order to produce the envelope curve that gives the extremal values. In the case of 24 intersections there would be 8 388 608 offset schemes to be investigated and for each offset scheme there would be a certain number of possible speeds, depending on the range of speeds to be studied.

THE ALGORITHM

The aim is to determine all possible extremal points on the V-B envelope curve for up to 25 intersections and also for a reasonable range of speeds and cycles.

It appears impractical to enumerate all possible

half-integer offset combinations. But this enumeration can be avoided. It is obvious that the speeds giving rise to the extremal points on the V-B curve must be straight lines in the time-space diagram, i.e., lines that correspond to V_d and V_u . These possible optimal speeds can be calculated as

Figure 9. Arteries with equal distance between intersections: top, 100 m; bottom, 400 m.

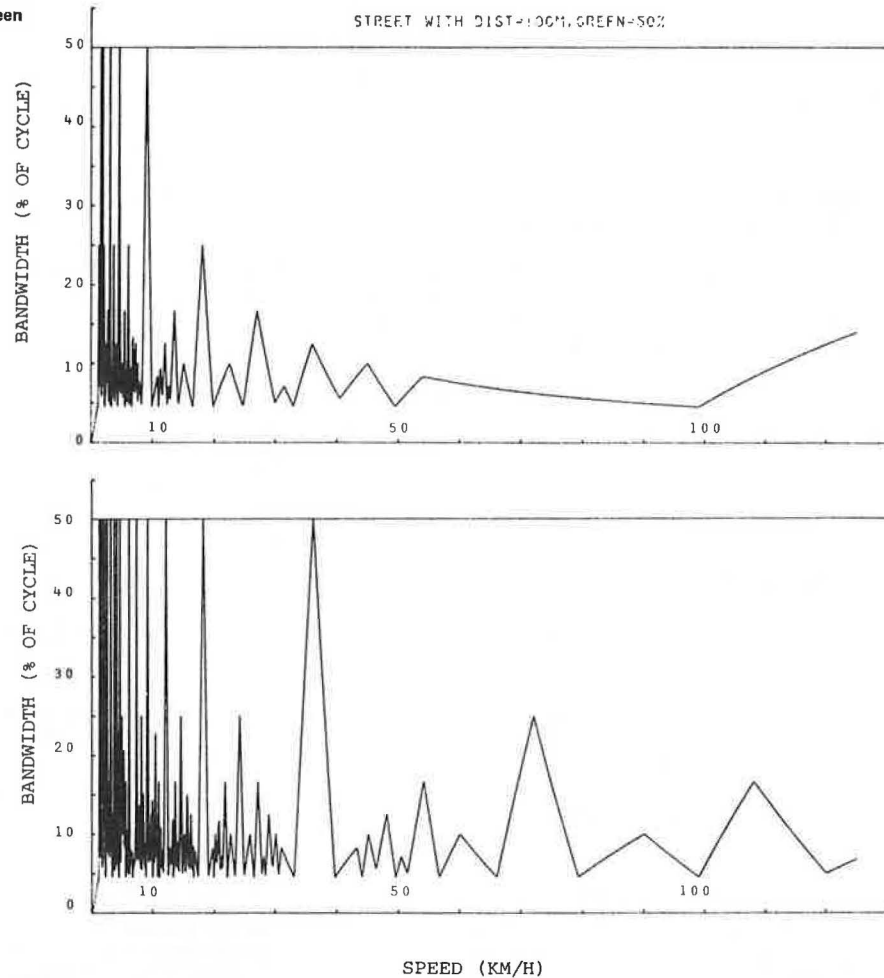
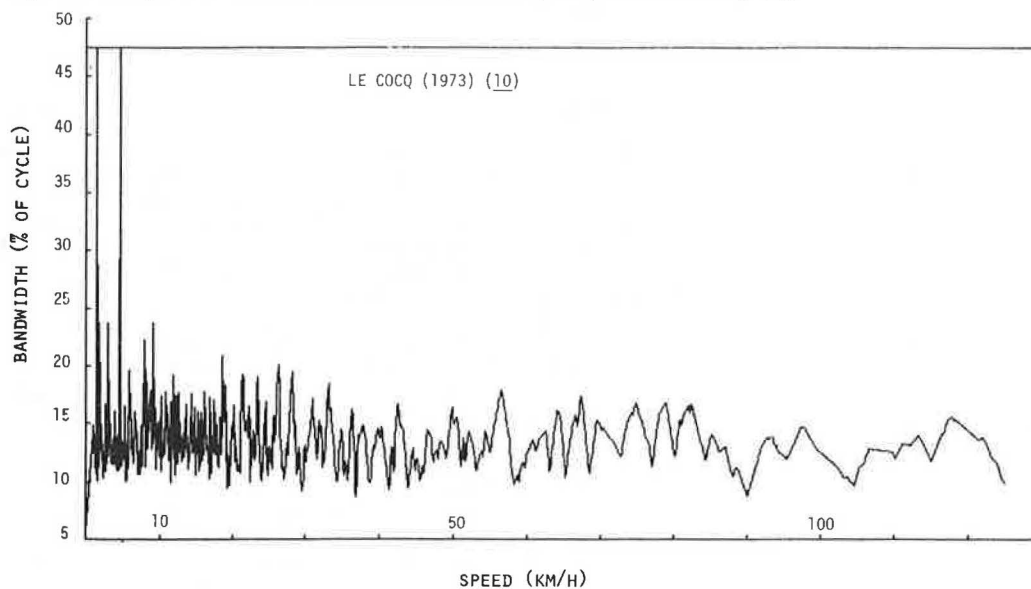


Figure 10. Example of V-B curve for 24 intersections (data set 15, 47.5 percent minimum green).



straight lines between any two critical lights i and j for all $i = 1 \dots n$ and all $j = (i + 1) \dots n$. Furthermore, as Figure 11 shows, there may be several possible optimal speeds between each pair of i and j .

For V_d (referring to Figure 11), we have

$$V_d = (720/C) [(x_j - x_i)/(r_j - r_i + 100\ell)] \quad (7)$$

For V_u (referring to Figure 12), we have

$$V_u = (720/C) [(x_j - x_i)/(r_j - r_i + 100k)] \quad (8)$$

where $\ell, k = 0 \dots m$ such that

$$V_{\min} \leq V \leq V_{\max}$$

The resulting speeds are speeds that are extremal points on the subset of V-B curves and that may be extremal points of the envelope curve.

We now show that the number (N) of speeds to be calculated in this way is small or at least computationally feasible for all cases in which the number of intersections (n) is less than 25. With V_{\min} and V_{\max} , the limits of the speed range to be investigated, fixed, we have from Equation 7 for V_d

$$\ell_{i1} \leq (720/100CV_{\min})(x_j - x_i) + (r_i - r_j)/100 \quad (9)$$

$$\ell_{i2} \geq (720/100CV_{\max})(x_j - x_i) + (r_i - r_j)/100 \quad (10)$$

Figure 11. Geometric relationships for V_d .

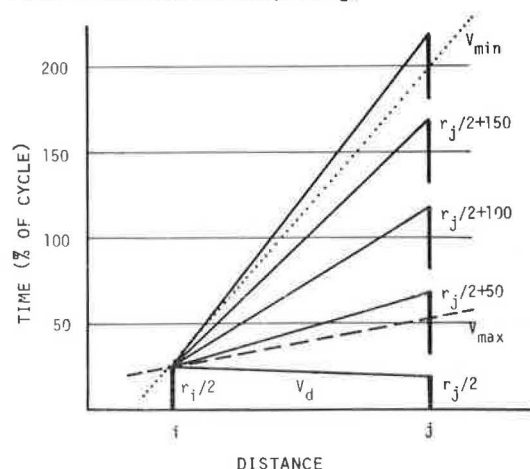
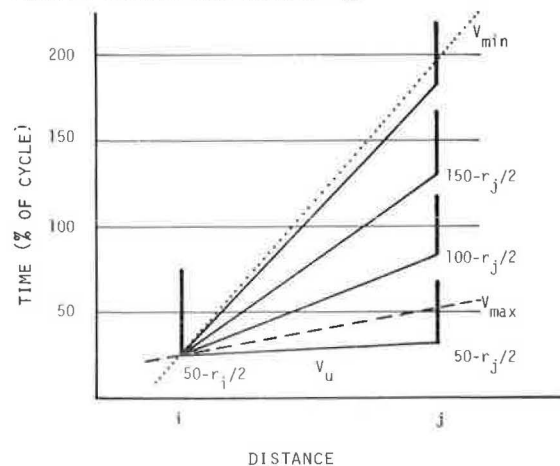


Figure 12. Geometric relationships for V_u .



If we take the same number for V_u , the number of speeds N_i to be calculated for each pair i, j would be

$$N_i = 2[IFIX(\ell_{i1}) - IFIX(\ell_{i2})]$$

where $IFIX(\ell)$ denotes the integer part of ℓ . An upper bound for N_i can be obtained by using

$$N_i \approx 2(\ell_{i1} - \ell_{i2}) \approx (14.4/C)(x_j - x_i) [(1/V_{\min}) - (1/V_{\max})]$$

$$N_i \approx F(x_j - x_i)$$

$$F = (14.4/C) [(1/V_{\min}) - (1/V_{\max})] \quad (11)$$

This result is not surprising. If the range of speeds to be considered between V_{\min} and V_{\max} is small, very few possible speeds will fall between these two extremes. Also, as V_{\min} becomes small many speeds are possibly optimal, which explains the many oscillations in the V-B curve at low speeds or low K-values. From the tabulation below, the importance of the lower speed range can be seen. In fact, 50 percent of possible speeds between 10 and infinity and 20 and infinity lie between 10 and 20 km/h:

Range of Speed

V_{\min}	V_{\max}	F	Percentage
10	∞	0.1	100.00
10	100	0.09	90.00
15	∞	0.067	67.00
20	∞	0.05	50.00
30	∞	0.033	33.33
40	∞	0.025	25.00
30	60	0.0167	16.66
20	80	0.0375	37.50
15	125	0.05867	58.67

N_i depends approximately only on the distances between intersections and one can develop a summation formula for N over all intersections i , which constitutes an upper bound for the number N :

$$N \leq F \left[\sum_{i=1}^n (2i - n - 1) x_i \right] \quad (12)$$

Eighteen data sets were analyzed. These data sets were taken from the published literature, and the second column in Table 1 gives the references. An approximate lower bound for N was obtained by considering that distances between intersections are, on average,

$$L = x_n/n$$

In this case formula 12 becomes

$$N \geq FL(n^3 - n)/6 \quad (13)$$

The usefulness of Equations 12 and 13 can be verified from the data in Table 1. In fact, the formulas correspond fairly well to the number (N) of speeds actually calculated. Figure 13 shows the relation between N and n . It may also be seen in Table 1 that the mean distance between intersections alone explains more than 80 percent of the number N . The remaining percentage is due to the differences in reds and the variation of distances about the mean. The rapidity of the algorithm (called WAVE1) is further increased because in many cases the configuration of the reds is such that the same speeds V_d and V_u are generated repeatedly and the bandwidth corresponding to these speeds does not have to be determined several times. Equations 7 and 8 show also that $V_d = V_u$ when $r_i = r_j$. This further reduces the number N .

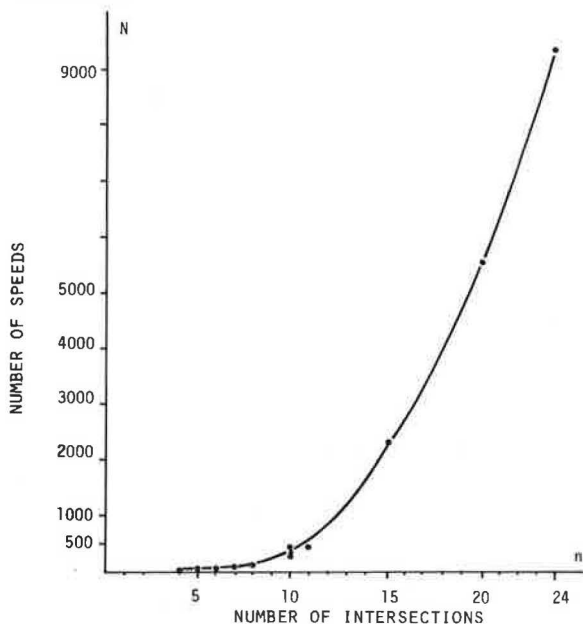
The main steps in the algorithm for finding the extremal points of the envelope curve and the cor-

Table 1. Number N for 18 data sets.

Data Set	Reference	No. of Intersections	N Exact	Lower Bound	Upper Bound	Avg Distance (m)
1	MAGTOP (1975) (4)	5	82	64	88	301.8
2	Little (1966) (5)	10	367	321	373	184.4
3	Le Cocq (1971) (6)	10	468	427	472	245.0
4	Institute of Traffic Engineers (40 percent) (1950) (7, pp. 229-239)	8	136	132	142	148.6
5	Institute of Traffic Engineers (1950) (7, pp. 229-239)	8	139	132	142	148.6
6	Davidson (1960) (8)	7	138	116	138	195.9
7	Purdy (1967) (9)	6	88	77	88	208.3
8	LeCocq (1973) (10)	4	61	42	61	401.3
9	Kell (1956) (11)	10	340	303	344	173.7
10	Laval	4	34	26	36	246.9
11	Pignataro (1973) (12, pp. 372-381)	6	38	38	46	104.1
12	Woods (1960) (13, pp. 7-40 to 7-43)	8	164	157	168	177.1
13	Morgan (1964) (2)	9	169	170	193	135.5
14	Kelson (1980) (14)	5	58	53	63	253.0
15	Le Cocq (1973) (10)	24	9384	9057	9394	364.6
16	Le Cocq (1973) (10)	20	5513	5021	5522	357.5
17	Le Cocq (1973) (10)	15	2387	2148	2396	363.3
18	200 m, 50 percent green	11	450	464	464	200.0

Note: All data sets in metric system; $C = 80$ s; $V_{\min} = 15$ km/h and $V_{\max} = 125$ km/h.

Figure 13. Number of possible optimal speeds N versus number of intersections n.



responding offsets, if they are required, over all ranges of K are described below. Note that certain steps and tests that are crucial for rapid execution and efficient storage are omitted here for conciseness. However, these steps are not essential for an understanding of the proposed procedure.

1. Do for $i = 1 \dots n$; if $i = n$, go to step 6.
2. Do for $j = (i + 1) \dots n$; if $j = n$, go to step 1.

$$C1 = r_j - r_i$$

$$C2 = (720/C) (x_j - x_i)$$

3. Do for all $\ell = 0 \dots m$ so that

$$V_{\min} \leq V_{dij} \leq V_{\max}$$

If no more ℓ satisfy the condition, go to step 2.

$$V_{dij} = C2 / (C1 + \ell 100)$$

$$V_{uij} = C2 / (\ell 100 - C1)$$

Do not retain speed if it hits a red light for all $k \neq i, j$.

4. Calculate for ℓ and do for $k = 1 \dots n$; if $k = n$, go to step 5.

$$C3 = 360 (x_j - x_k) / V_{dij} C \text{ or } V_{uij}$$

$$C4 = (r_j + r_k) / 2$$

$$b_{kd} = C3 - C4 \pm \text{INT} * 50 \text{ if } V_d$$

$$b_{ku} = -C3 - C4 \pm \text{INT} * 50 \text{ if } V_u$$

INT is such that

$$0 \leq b_{kd} \leq g_k \text{ or } 0 \leq b_{ku} \leq g_k$$

Calculate offset, if required. If not, go to step 4.

$$\text{TET} = C3 - b_{kd} - C4 \text{ if } V_d$$

$$\text{TET} = C3 + b_{ku} + C4 \text{ if } V_u$$

If TET is negative, change the sign.

5.

$$\text{THETA}_k = \text{AMOD}(\text{TET}, 100)$$

$$B_d = \text{MIN}_k(b_{kd})$$

$$B_u = \text{MIN}_k(b_{ku})$$

Go to step 3.

6. The extremal points are found by a simple search algorithm that eliminates extremal points not on the envelope curve.

7. Choose speed and cycle over all acceptable extremal points by using the relationship $K = VC$.

Another algorithm based on a modified Brooks (15) algorithm, which gives the envelope curve and the extremal points, is called WAVE2. This algorithm produces the same extremal points as WAVE1 but to a lesser degree of accuracy. All data sets except 15 to 18 were also tested by Couture (16), who used the MAXBAND program of Kelson (14). The accuracy of these approaches is compared in Table 2. For the purposes of comparison only, the extremal point found by MAXBAND is given. WAVE1 determines all extremal points and WAVE2 produces also the continuous V-B envelope curve. All data sets were tested over a range of $K_{\min} = 640$ to $K_{\max} = 10,000$, which represents the practical limits of V and C .

Table 3 compares execution times in seconds for WAVE1 and WAVE2 on an IBM 4341 computer over all ranges of K and n and for up to 24 intersections. Execution time for WAVE1 is highly dependent on the

average distance between intersections. Figure 14 shows this comparison and it can be seen that the 20-intersection case would be the cutoff point between the two programs.

Figure 15 gives the graphic output of program WAVE2 and Figure 16 is an output listing of the program WAVE1.

CONCLUSION

Maximum bandwidth varies with speed, especially at

lower speeds. The extent of this problem depends mostly on the distances between intersections. In certain cases (when the average distance is great), maximum bandwidth changes so rapidly with speed that no stability can be expected, which may partly explain the unreliable functioning of certain progressions. It would seem important not only to consider the extremal points of bandwidth but also to take into account their stability with regard to changing speeds. The best choice of speed and bandwidth may in fact be lower than the overall maximum. The proposed method generates the entire relationship be-

Table 2. Comparison of precision for MAXBAND, WAVE1, and WAVE2.

Data Set	MAXBAND		WAVE1		WAVE2	
	K	B	K	B	K	B
1	2156.00	0.3000	2156.00	0.3000	2152.00	0.3000
2	1364.80	0.3364	1364.80	0.3363	1368.00	0.3333
3	745.89	0.4698	745.89	0.4698	744.00	0.4597
4	1410.21	0.2724	1410.21	0.2724	1416.00	0.2712
5	1410.21	0.3724	1410.21	0.3724	1416.00	0.3712
6	704.20	0.4297	704.20	0.4297	704.00	0.4295
7	1060.80	0.3775	1060.80	0.3775	1064.00	0.3750
8	3549.70	0.4444	3549.70	0.4444	3552.00	0.4444
9	1563.41	0.2369	1563.62	0.2368	1560.00	0.2293
10	1215.52	0.5539	1215.20	0.5538	1216.00	0.5527
11	1426.41	0.3393	1426.46	0.3392	1424.00	0.3388
12	2506.67	0.3500	2506.50	0.3500	2512.00	0.3500
13	1096.95	0.6000	1096.93	0.6000	1096.00	0.5953
14	3017.50	0.4000	3017.60	0.4000	3024.00	0.4000
15			1488.00	0.2100	1488.00	0.2093
18			1440.00	0.5000	1440.00	0.5000

Table 3. Central-processing-unit time for programs WAVE1 and WAVE2.

Data Set	WAVE1	WAVE2	n	Data Set	WAVE1	WAVE2	n
1	0.49	3.82	5	10	0.40	3.49	4
2	0.79	5.61	10	11	0.39	3.72	6
3	0.94	5.39	10	12	0.60	4.46	8
4	0.61	4.58	8	13	0.56	4.76	9
5	0.56	4.71	8	14	0.41	3.40	5
6	0.54	4.24	7	15	33.9	21.5	24
7	0.44	3.52	6	16	16.6	12.1	20
8	0.44	3.47	4	17	5.72	7.78	15
9	0.98	5.57	10	18	1.54	4.93	11

Figure 14. Comparison of central-processing-unit time for programs WAVE1 and WAVE2.

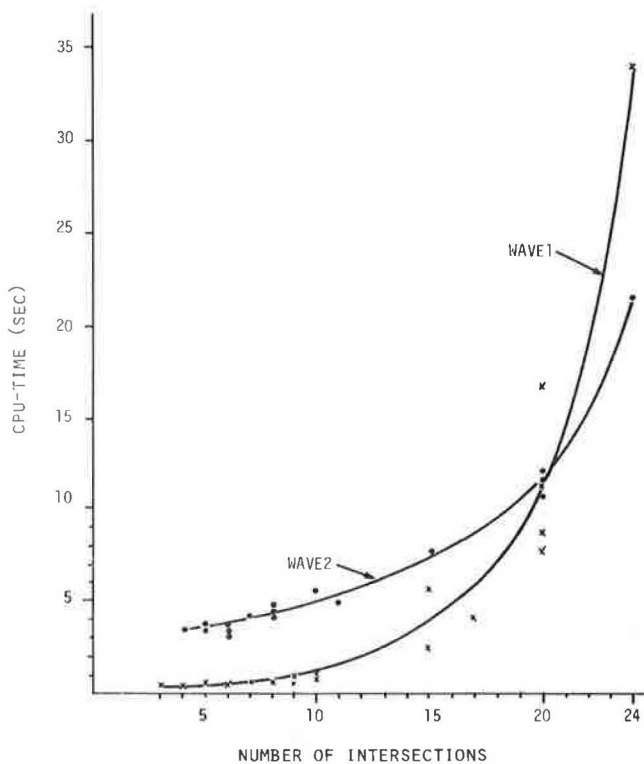


Figure 15. Graphic output of WAVE2.

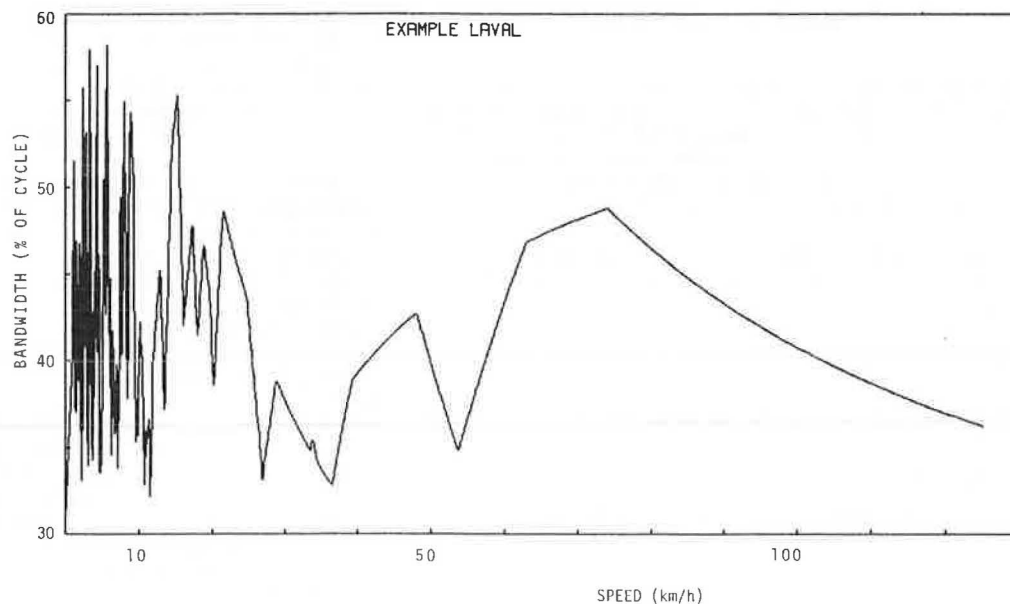


Figure 16. Output of WAVE1.

EXEMPLE LAVAL

* CARREFOUR *	ABSCISSE	* ROUGE *
1	0.0 MEIRES	25.00 POURCENI
2	297.18 MEIRES	24.00 POURCENI
3	803.15 MEIRES	40.00 POURCENI
4	987.55 MEIRES	40.00 POURCENI

GMIN= 60.00 SEC DUREE DU CYCLE= 80.00 SEC VMIN= 15 KM/H VMAX= 125 KM/H

EXEMPLE LAVAL

VIIESSSES OPTIIMALES POSSIBLES 34

VII. BANDE	VII. BANDE	VII. BANDE	VII. BANDE	VII. BANDE	VII. BANDE	VII. BANDE	VII. BANDE	VII. BANDE	VII. BANDE	VII. BANDE
15.19 55.38	16.03 11.75	16.18 42.85	16.60 35.27	17.26 48.01	17.42 12.36	18.33 14.72	18.77 46.73	19.66 43.51	21.08 20.64	
21.42 48.75	21.88 22.07	22.95 23.84	23.09 24.06	24.75 43.53	25.36 28.23	26.48 35.32	27.02 33.72	28.22 28.10	28.77 38.85	
31.19 33.39	33.62 35.32	33.77 35.43	39.07 31.24	39.26 38.86	41.34 30.07	48.04 42.73	53.56 25.49	54.21 25.31	62.86 46.80	
73.97 48.78	77.29 20.74	85.04 19.76	104.56 38.29							

EXEMPLE LAVAL

VIIESSSES OPTIIMALES SUR LA COURBE D'ENVELOPPE

VII. BANDE PCENI	VII. BANDE PCENI	VII. BANDE PCENI	VII. BANDE PCENI	VII. BANDE PCENI	VII. BANDE PCENI
15.19 55.38 *****	16.18 42.85 77.37	17.26 48.01 86.69	18.77 46.73 84.37	21.42 48.75 88.02	24.75 43.53 78.60
26.48 35.32 63.77	28.77 38.85 70.14	33.77 35.43 63.97	39.26 38.86 70.17	48.04 42.73 77.15	62.86 46.80 84.50
73.97 48.78 88.08	104.56 38.29 69.14				

tween speed and bandwidth in the form of an envelope curve, or it may generate all extremal points of this curve in extremely short execution times for arteries with up to 25 intersections. The proposed procedure may be practically useful since it provides more than just a single point solution and may contribute to a better understanding of the basic relationships.

REFERENCES

1. J.D.C. Little and others. Synchronizing Traffic Signals for Maximal Bandwidth. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA, Rept. R 64-08, March 1964, 54 pp.
2. J.T. Morgan. Synchronizing Traffic Signals for Maximal Bandwidth. Operations Research, Vol. 12, 1964, pp. 896-912.
3. J.D.C. Little and others. MAXBAND: A Program for Setting Signals on Arteries and Triangular Networks. TRB, Transportation Research Record 795, 1981, pp. 40-46.
4. MAGTOP: User's Manual Sample Input and Output. FHWA, 1975, 291 pp.
5. J.D.C. Little. Synchronization of Traffic Signals by Mixed Integer Linear Programming. Operations Research, Vol. 14, 1966, pp. 568-594.
6. J.P. Le Cocq. Coordination des feux sur un itinéraire: ACHILLE II--Maximisation de l'onde verte. Services d'Etudes Techniques des Routes and Autoroutes (SETRA), Ministère de l'Équipement et du Logement, Paris, France, 1971, 57 pp.
7. H.K. Evans, ed. Traffic Engineering Handbook. Institute of Traffic Engineers, New York, 1950.
8. B. Davidson. Design of Signal Systems by Graphical Solutions. Traffic Engineering, Vol. 30, Nov. 1960, pp. 32-38.
9. F.G. Purdy. The Arithmetic of a Balanced Two-Way Signal Progression. Traffic Engineering, Vol. 37, No. 4, Jan. 1967, pp. 48-49.
10. J.P. Le Cocq. Coordination des feux sur un itinéraire: ACHILLE. SETRA, Paris, France, 1973, 64 pp.
11. J.H. Kell. Coordination of Fixed Time Traffic Signals. Institute of Transportation and Traffic Engineering, Univ. of California, Berkeley, Course Notes, Aug. 1956, 19 pp.
12. L.J. Pignataro. Traffic Engineering: Theory and Practice. Prentice-Hall, New York, 1973.
13. K.B. Woods and others. Highway Engineering Handbook. McGraw-Hill, New York, 1960.
14. M.D. Kelson. Optimal Signal Timing for Arterial Signal Systems: Vol. 2--MAXBAND User's Manual. FHWA, Rept. FHWA/RD-80/083, Dec. 1980, 230 pp.
15. W.D. Brooks. Vehicular Traffic Control--Designing Arterial Progression Using a Digital Computer. IBM Data Processing Division, Kingston, NY, 1964.
16. L. Couture. La coordination des feux de circulation sur une artère pour optimiser la bande verte en considérant les retards. Ecole Polytechnique, Université de Montréal, Montreal, Canada, master's thesis, 1982.

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

Abridgment

Calibration of TRANSYT Platoon Dispersion Model for Passenger Cars Under Low-Friction Traffic Flow Conditions

PATRICK T. McCOY, ELIZABETH A. BALDERSON, RICHARD T. HSUEH, AND ABBAS K. MOHADDES

The calculation of delay and stops by the TRANSYT program and in turn the effectiveness of the signal timings resulting from its optimization procedure depend on the ability of its platoon dispersion model to accurately predict traffic flow patterns from one signal to another. Therefore, calibration of the dispersion factor α and travel-time factor β in the model is important to the successful implementation of the TRANSYT program. However, because of limited research, a definitive description of the relationship between the appropriate values of α and β and roadway conditions does not exist. The objective of this research was to contribute to the ultimate development of a definitive description of this relationship by calibrating this model for passenger cars under low-friction traffic flow conditions. Platoon dispersion studies were conducted on six arterial street segments (2 two-way two-lane segments and 4 four-lane divided segments). Traffic flow patterns of nearly 1700 platoons were analyzed. The results indicated that less platoon dispersion was observed in this study than has been found in other studies of low-friction traffic flow conditions. It was concluded that appropriate values of α and β for passenger cars under these conditions are α equal to 0.21 and β equal to 0.97 on two-way two-lane streets and α equal to 0.15 and β equal to 0.97 on four-lane divided streets.

The traffic simulation model in the TRANSYT signal-timing optimization program is recognized as one of the most realistic in the family of macroscopic computerized traffic simulation models (1). Rather than assume uniform traffic flow within platoons traveling from one signal to another, the TRANSYT model accounts for the effects of the dispersion of platoons as they travel between signals. The model uses the following recurrence relationship developed by Robertson (2) to simulate platoon dispersion:

$$q^1(i + \beta t) = F \times q(i) + (1 - F) \times q^1(i + \beta t - 1) \quad (1)$$

$q^1(i)$ = flow in i th time interval of predicted platoon;

$q(i)$ = flow in i th time interval of initial platoon;

t = average travel time over distance for which platoon dispersion is being calculated;

β = empirical travel-time factor expressed as ratio between average travel time of leading vehicle in platoon and average travel time of entire platoon ($0 < \beta \leq 1.0$);

F = empirical smoothing factor, which controls rate at which platoon disperses, expressed as $F = 1/(1 + \alpha \beta t)$; and

α = empirical dispersion factor between 0 and 1.0.

The amount of dispersion in the traffic flow pattern predicted by Equation 1 is determined by the parameters α and β . A value of 0.0 for α and a value of 1.0 for β represents no platoon dispersion over the distance for which the platoon dispersion is being calculated. Values of 1.0 for both α and β represent a maximum platoon dispersion over the distance for which the platoon dispersion is being calculated.

The calculation of delay and stops by the TRANSYT program and in turn the effectiveness of the signal timings resulting from its optimization procedure depend on the ability of the recurrence relationship in Equation 1 to accurately predict the traffic flow pattern from signal to signal (2). Therefore, selection of appropriate values for α and β is important to the successful implementation of the

TRANSYT program. From an analysis of traffic flow patterns observed by Hillier and Rothery (3) at four sites in West London, England, Robertson (2) determined that the values of these parameters that resulted in the best fit between the actual and calculated traffic flow patterns were α equal to 0.5 and β equal to 0.8. The sites where the observations were made had different traffic conditions that ranged from single-lane flow with heavy parking and very restricted overtaking to multilane flow with no parking and relatively free overtaking. However, Robertson (2) cautioned users of TRANSYT that one might expect the appropriate values for α and β to be a function of site factors such as roadway width, gradient, parking, opposing traffic volume, traffic composition, and others.

Since the development of the recurrence relationship in Equation 1, only a limited number of studies to evaluate its parameters have been documented. Sneddon (4) applied Equation 1 to traffic flow data collected at two sites in Manchester, England, to determine the values of α and β that provided the best fit to the observed traffic flow patterns. One of the sites was a three-lane dual carriageway with 10-15 percent commercial vehicles in the peak hours and reasonable freedom for overtaking. The other site was a two-way road 35 ft wide with 2-3 percent commercial vehicles, two narrow lanes in the direction studied, and severely restricted overtaking. For the three-lane dual carriageway, the best-fit value of α was 0.40, and for the two-way road, the best-fit value of α was 0.63. For both sites, the best-fit value of β was 0.8.

In another study, Collins and Gower (5) observed the dispersion of platoons of passenger cars on a three-lane dual carriageway in the suburbs of London, England. They found that the values of α and β that provided the best fit between the observed and calculated traffic flow patterns were 0.20 and 0.80, respectively.

In Toronto, Canada, Lam (6) conducted platoon dispersion studies on a four-lane two-way suburban arterial street with left-turn bays. In using Equation 1 with the parameter values suggested by Robertson (i.e., $\alpha = 0.5$ and $\beta = 0.8$), he found an average error of 13.8 percent in the computed value of delay. Therefore, he calibrated Equation 1 by using his observed traffic flow data. He found that the parameter values that provided the best fit between the observed and predicted platoon dispersion were α equal to 0.24 and β equal to 0.8. As a result of this calibration, the average error in the computed delay was significantly reduced to 8.2 percent.

El-Reedy and Ashworth (7) conducted a study of platoon dispersion along a single carriageway in Sheffield, England. The road was 33 ft wide. The traffic flow observed was on a 5 percent downgrade. It was subject to a 30-mph speed limit, and it had a bus volume of 12/h. Clearway regulations applied during the morning peak period when the observations were made. Data were collected at three different positions on three different days at distances of 1082 ft (position A), 1378 ft (position B), and 1837 ft (position C) downstream from a signal. The good-

ness of fit of traffic flow patterns predicted by Equation 1 with the observed traffic flow patterns was evaluated separately for each position. No platoon dispersion was found at position A. The best-fit parameter values at position B were α equal to 0.60 and β equal to 0.63. At position C, these values were α equal to 0.70 and β equal to 0.59.

To validate the accuracy of Equation 1 for conditions found in the United States, traffic flow data were collected by Alan M. Voorhees and Associates (8) on Route 7 east of its intersection with Towlston Road in Fairfax County, Virginia. Data were collected 100, 400, and 800 ft downstream from the intersection. Vehicle speeds were 55 mph. Equation 1 was applied to the average traffic flow pattern observed at the 100-ft station to predict the traffic flow patterns at the 400- and 800-ft stations. The parameter values determined by Lam (i.e., α equal to 0.24 and β equal to 0.8) were used to predict these patterns, because the data collected closely matched the free-flow, suburban arterial case evaluated by Lam. At both the 400- and 800-ft stations, close agreement was obtained between the observed and predicted traffic flow patterns. Although it was acknowledged that additional research should be conducted to further refine the relationship between the parameters of Equation 1 and roadway conditions, it was concluded that it was possible to develop the general recommendations regarding this relationship shown in the first two columns of Table 1 (8).

Based on the results of the study conducted by Alan M. Voorhees and Associates (8) and preliminary work performed by the University of Florida, parameter values similar to those presented in the first two columns of Table 1 are suggested in the user's manual for TRANSYT-7F (1). A comparison of these values, which are also presented in Table 1 (1), with those from the NCHRP report shows that these two sets of suggested parameter values are identical except in the case of α for moderate and low friction. However, as implied by the discussion in the user's manual, this agreement is more of an indica-

tion of the extent to which the parameter values from the TRANSYT-7F manual were based on those given in the NCHRP report than it is a validation of the relationship shown between the parameter values and roadway conditions.

A summary of the platoon dispersion studies cited above and conducted outside the United States is presented in Table 2. Together the findings of these studies indicate that there is a relationship between platoon dispersion and roadway conditions. In general, the greater the apparent level of friction to traffic flow implied by the description of roadway conditions, the greater is the platoon dispersion as reflected by a higher best-fit value of α . But although there seems to be a general agreement between the nature of the relationship shown in Table 2 and that presented in Table 1, the ranges of the parameter values are greater in Table 2. Not only does α vary over a wider range, but also best-fit values other than 0.8 are indicated for β . In fact, the limited amount of platoon dispersion research conducted in the United States has not investigated values other than 0.8 for β , probably because the TRANSYT signal-timing optimization program allows the user to specify an input value only for α and not for β . The value of 0.8 for β is fixed in the program. Thus, there is a need to develop a more definitive description of the relationship between these parameters and roadway conditions.

The primary objective of the research reported in this paper was to provide information that would contribute to the ultimate development of a definitive description of the relationship between the appropriate parameter values of the TRANSYT platoon dispersion model presented in Equation 1 and roadway conditions. The specific objectives of the research were (a) to observe the dispersion of platoons of passenger cars on urban arterial streets under low-friction traffic flow conditions and (b) to calibrate the TRANSYT platoon dispersion model for the conditions observed. This paper presents the procedure, findings, and conclusions of this research.

Table 1. Parameter values recommended in NCHRP Report 233 and in TRANSYT-7F manual.

NCHRP 233		TRANSYT-7F Manual		Roadway Characteristic	Description of Conditions
α	β	α	β		
0.5	0.8	0.5	0.8	Heavy friction	Combination of parking, moderate to heavy turns, moderate to heavy pedestrian traffic, narrow lane width; traffic flow typical of urban CBD
0.37	0.8	0.35	0.8	Moderate friction	Light turning traffic, light pedestrian traffic, 11- to 12-ft lanes, possibly divided; typical of well-designed CBD arterial
0.24	0.8	0.25	0.8	Low friction	No parking, divided, turning provisions, 12-ft lane width; suburban high-type arterial

Table 2. Summary of platoon dispersion studies conducted outside United States.

Best-Fit Parameter Value		Description of Conditions	Reference
α	β		
0.20	0.80	Three-lane dual carriageway; suburban high-type arterial	Collins and Gower (5)
0.24	0.80	Typical suburban arterial roadway with two lanes in each direction; turn lanes provided	Lam (6)
0.40	0.80	Three-lane dual carriageway with 10-15 percent commercial vehicles; reasonable freedom for overtaking	Sneddon (4)
0.63	0.80	Two-way road 35 ft wide with two narrow lanes in the direction studied; 2-3 percent commercial vehicles; severely restricted overtaking	Sneddon (4)
0.60 ^a	0.63 ^a	Single carriageway 33 ft wide on 5 percent downgrade; subject to 30-mph speed limit and clearway regulations during peak periods; bus volume of 12 vehicles/h in direction studied	El-Reedy and Ashworth (7)
0.70 ^b	0.59 ^b	Characteristics ranging from single-lane flow with heavy parking and very restricted overtaking to multilane flow with no parking and relatively free overtaking	Robertson (2), Hillier and Rothery (3)
0.50	0.80		

^a For traffic flow pattern 1378 ft downstream.

^b For traffic flow pattern 1837 ft downstream.

Table 3. Study sites.

Site	Type of Arterial Street	Lanes Observed		Alignment	Gradient	Parking	Driveway Access	Speed Limit (mph)	Peak Period Studied
		No.	Width (ft)						
1	Two-way two-lane	1	13	Tangent	Level	None	None	35	p.m.
2	Two-way two-lane	1	13	Tangent	Level	None	Limited	35	p.m.
3	Four-lane divided	2	12	Tangent	Level	None	None	45	a.m.
4	Four-lane divided	2	12	Tangent	Level	None	None	45	a.m.
5	Four-lane divided	2	13	Tangent	Level	None	None	45	a.m.
6	Four-lane divided	2	12	Tangent	Level	None	Limited	45	a.m.

PROCEDURE

Platoon dispersion studies were conducted on six arterial street segments in Lincoln, Nebraska, during the summer of 1981. At each location, the traffic flow patterns of platoons of passenger cars discharging through a signalized intersection were recorded at four points over a distance of 1000 ft downstream from the signal. Two of the segments were on two-way two-lane streets and the other four were on two-way four-lane divided streets. Each segment was a level, tangent section and each had low-friction-type roadway characteristics. All of the studies were conducted during peak periods under fair weather and dry pavement conditions.

The traffic flow patterns recorded during these studies were analyzed to determine the values of the parameters α and β of the TRANSYT platoon dispersion model that provided the best agreement between observed traffic flow patterns and those predicted by the model. The Kolmogorov-Smirnov (K-S) test was then applied to evaluate the goodness of fit of the observed patterns with the patterns predicted by the calibrated model. Also, in order to provide a basis of comparison with previous research, the best-fit values of α with β equal to 0.8 were determined and tested.

Study Sites

In the selection of the study sites, an effort was made to find arterial street sections downstream of signalized intersections that were typical of low-friction traffic flow conditions. Therefore, all six sites selected were level, tangent sections with 12- to 13-ft lanes, and there was no parking at any time along them. Four of the study sections did not have any driveways along them, and the driveway volumes on the other two sections were negligible and did not interfere with the platoon flow on the sections during the study periods. Two of the sections were on two-way two-lane streets with 35-mph posted speed limits. The other four sections were on four-lane divided streets with 45-mph posted speed limits. Characteristics of the study sites are summarized in Table 3.

The approaches to the signalized intersections upstream of the study sites on the two-way two-lane arterial streets had right-turn lanes of sufficient length so that right-turn movements did not interfere with the discharge of through-vehicle platoons in the adjacent lane. These approaches did not have left-turn lanes. However, the left-turn volumes on these approaches were very low during the study periods so that the left-turn movements seldom interfered with the discharge of through-vehicle platoons. During the conduct of the field studies at these locations, the occurrence of left-turn movements was noted so that data collected for the platoons with which they were associated could be excluded from the subsequent data analysis.

The approaches to the signalized intersections upstream of the study sites on the four-lane divided streets had left-turn and right-turn lanes. These lanes were of sufficient length so that the turning movements on these approaches did not interfere with the discharge of vehicle platoons from the two through lanes.

Data Collection

At each study site, four observers were stationed downstream from the signalized intersection. The first observer was located at a point immediately downstream from the intersection. The other three observers were located downstream from the first observer at distances of 300, 600, and 1000 ft, respectively. At each of these points, the observer recorded the time of arrival of every vehicle coming from the intersection by pressing a switch connected to a 20-pen recorder.

A fifth observer was stationed at the site of the 20-pen recorder, which was located near the intersection. This observer made notations directly on the recorder chart paper of the following items: (a) the beginning and ending of the green times on the street being studied, (b) the first vehicle of each platoon passing through the intersection, (c) platoons that contained any vehicles other than passenger cars, and (d) platoons that were delayed and/or influenced by turning movements, pedestrians, stalled vehicles, or any other circumstances. These notations were made to facilitate the identification of passenger-car platoons to be studied in the subsequent data analysis.

Data Analysis

The 20-pen recorder charts were examined. Data for platoons that contained vehicles other than passenger cars and for platoons that were delayed or influenced by turning movements, pedestrians, stalled vehicles, or other circumstances were identified and excluded from the analysis. For each of the remaining platoons, the record of its traffic flow pattern at each of the four downstream observation points was identified and expressed as a histogram relating the number of vehicles that passed the point in each 2-s interval to time referenced to the instant at which the first vehicle in the platoon passed the first observation point. This histogram representation of platoon flow past each observation point is shown in Figure 1.

The histograms of all platoon flows at each study site were combined to determine the average platoon flow pattern at each observation point. The TRANSYT platoon dispersion model (Equation 1) was then applied to the average platoon flow pattern at the first observation point to predict an average platoon flow pattern at each of the other three observation points. Average platoon flow patterns were predicted for each combination of α and β values

where α and β were varied in increments of 0.01 over the ranges of 0.00-1.00 and 0.50-1.00, respectively. Then, the combination of α and β values that minimized the sum of squares of the differences between the observed and predicted average platoon flow patterns at the 300-, 600-, and 1000-ft observation points was selected as the best-fit α and β values for the study site. The K-S test was then applied to evaluate the goodness of fit of the observed patterns with those predicted by the calibrated model.

Similarly, to provide a basis of comparison with other platoon dispersion research cited earlier, the best-fit value of α with β equal to 0.8 was determined for each study site. Again, the K-S test was applied to evaluate the goodness of fit of observed patterns with the predicted patterns.

FINDINGS

Traffic flow patterns were analyzed for nearly 1700 passenger-car platoons, which ranged in size from 5 to 20 vehicles on the two-way two-lane study sites (sites 1 and 2) and from 5 to 38 vehicles on the four-lane divided study sites (sites 3, 4, 5, and 6). The best-fit values of the dispersion factor α and the travel-time factor β for the average platoon flow pattern at each study site are presented in Table 4. Also presented in this table are the best-fit values of α for β equal to 0.8.

Comparison of the best-fit values of α and β indicates that under the low-friction traffic flow conditions studied there was more platoon dispersion on the two-way two-lane sections than there was on the four-lane divided sections. Also, comparison of these values with those in Tables 1 and 2 indicates

that less platoon dispersion was observed in this research than was found in other research on low-friction traffic flow conditions. Of course, it must be remembered that the results of this research apply only to passenger-car platoons. The inclusion of platoons that had vehicles other than passenger cars in them, which was reported to be the case in some of the previous research (4,5,7), would have probably tended to increase the amount of platoon dispersion observed in this research. However, since the percentage of vehicles other than passenger cars observed at the study sites was very low (less than 3 percent), it is expected that only a slight increase would have resulted.

In comparison with the results of previous research, not only are the best-fit values of α found in this research lower, but also the best-fit values of β are much higher than 0.8, which is the travel-time factor used in the TRANSYT program and which is the value found, or apparently in some cases used, in previous research (2,4-6,8). However, the comparison of the best-fit values of α for β equal to 0.8 shown in Table 4 with the α values for the β equal to 0.8 shown in Table 1 for low-friction roadway characteristics indicates that more rather than less platoon dispersion was observed in this research than in the previous research. In fact, these higher α values are indicative of moderate-friction conditions rather than low-friction ones. However, this was to be expected, because with the travel-time factor β equal to 0.8, the TRANSYT platoon dispersion model (Equation 1) predicts platoons to arrive earlier at points downstream than if the best-fit values were used. Consequently, the resultant best-fit values of α for β equal to 0.8 were higher, which spread the platoons more and tended to compensate for the early arrivals. It should be noted that in this research the average travel time (t) used in Equation 1 was the average of the travel times of all vehicles in the platoon.

In all cases, the results of the K-S tests, conducted at a 10 percent level of significance, indicated that the observed platoon flow patterns fit those predicted by the calibrated TRANSYT platoon dispersion model (i.e., Equation 1 with the best-fit values of α and β) at all downstream observation points. This was also true for the best-fit values of α for β equal to 0.8.

The relationship between platoon dispersion and platoon size was also investigated. The best-fit values of α and β for different platoon sizes at each study site are presented in Table 5. In general, these results indicate that larger platoon sizes experience slightly more dispersion than do smaller platoon sizes.

CONCLUSIONS

Based on the findings of this study, it was concluded that the dispersion of passenger-car platoons on urban arterial streets under low-friction traffic flow conditions is less than that determined for

Figure 1. Histogram representation of platoon flow pattern.

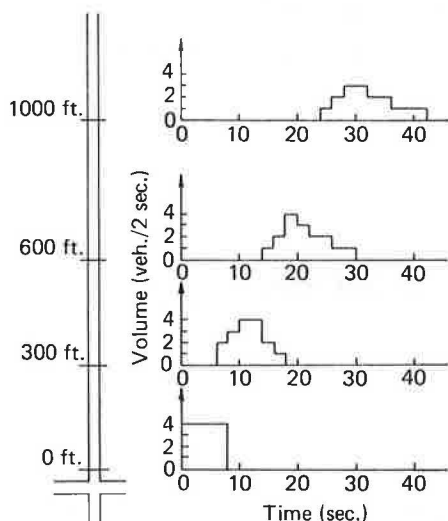


Table 4. Summary of results.

Site	Type of Street Section	Best-Fit Parameter Value		Value of α for $\beta = 0.8$	Range of Platoon Size	No. of Platoons
		α	β			
1	Two-way two-lane	0.22	0.99	0.51	5-15	294
2	Two-way two-lane	0.20	0.96	0.35	5-20	319
3	Four-lane divided	0.16	0.95	0.38	5-38	309
4	Four-lane divided	0.13	0.97	0.35	5-23	303
5	Four-lane divided	0.14	0.99	0.36	5-23	286
6	Four-lane divided	0.16	0.96	0.38	5-15	180

Table 5. Best-fit parameter values versus platoon size.

Site	Platoon Size Range									
	5-10 Vehicles		11-15 Vehicles		16-20 Vehicles		21-25 Vehicles		26-30 Vehicles	
	α	β	α	β	α	β	α	β	α	β
1	0.21	0.99	0.24	0.99	NA	NA	NA	NA	NA	NA
2	0.21	0.92	0.24	0.94	NA	NA	NA	NA	NA	NA
3	0.08	0.99	0.14	0.97	0.12	0.97	0.14	0.98	0.16	0.99
4	0.09	0.98	0.11	0.98	0.13	0.97	NA	NA	NA	NA
5	0.06	0.99	0.10	0.98	0.16	0.96	NA	NA	NA	NA
6	0.13	0.97	0.16	NA	NA	NA	NA	NA	NA	NA

Note: NA = sufficient data not available.

low-friction roadway characteristics by previous research. In addition, the dispersion of passenger-car platoons is less on a four-lane divided arterial street than it is on a two-way two-lane arterial street. Also, larger platoons experience more dispersion than do smaller platoons.

In regard to calibration of the TRANSYT platoon dispersion model, the average results of this research indicate that appropriate values of the dispersion factor α and the travel-time factor β for passenger-car platoons under low-friction traffic flow conditions on urban arterial streets are as follows:

Type of Street	α	β
Two-way two-lane	0.21	0.97
Four-lane divided	0.15	0.97

Thus, in order to more accurately account for the patterns of passenger-car platoon flow for these conditions, the input to the TRANSYT program should be revised to enable the user to specify the travel-time factor β as well as the dispersion factor α .

REFERENCES

1. TRANSYT-7F User's Manual. Office of Traffic Operations, FHWA, Feb. 1981.
2. D.I. Robertson. TRANSYT: A Traffic Network Study Tool. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, Rept. LR 253, 1969.
3. J.A. Hillier and R. Rothery. The Synchronization of Traffic Signals for Minimum Delay. Transportation Science, Vol. 1, No. 2, May 1967, pp. 81-94.
4. P.A. Sneddon. Another Look at Platoon Dispersion--3: The Recurrence Relationship. Traffic Engineering and Control, Feb. 1972, pp. 442-444.
5. J.F. Collins and P. Gower. Dispersion of Traffic Platoons on A4 in Hounslow. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, TRRL Rept. SR 29UC, 1974.
6. J.K. Lam. Studies of a Platoon Dispersion Model and Its Practical Applications. Proc., Seventh International Symposium on Transportation and Traffic Theory, Institute of Systems Science Research, Kyoto, Japan, 1977.
7. T.Y. El-Reedy and R. Ashworth. Platoon Dispersion Along a Major Road in Sheffield. Traffic Engineering and Control, April 1978, pp. 186-189.
8. P.J. Tarnoff and P.S. Parsonson. Selecting Traffic Signal Control at Individual Intersections. NCHRP, Rept. 233, June 1981.

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

Evaluation of Dynamic Freeway Flow Model By Using Field Data

N.A. DERZKO, A.J. UGGE, AND E.R. CASE

An attempt to calibrate and validate a dynamic freeway model by using real data from Queen Elizabeth Way in Ontario, Canada, is described. The model used in this research is the one developed by H. Payne; one of the Phillips kinetic models was also applied for comparison purposes. The overall conclusion is that the models exhibit instabilities in their behavior and do not track real road data correctly.

Traffic simulation models are playing an increasingly important role in the development of urban freeway traffic management systems because they provide an economical and safe way to evaluate alternative system designs and control strategies prior to implementation. Freeway models in common use today are adequate for simulating traffic conditions over

a period of hours but are not sufficiently realistic for the research and development of new surveillance and control techniques for real-time applications. For such applications, the model must have the ability to realistically represent the shorter-term dynamic phenomena (e.g., shock waves) characteristic of traffic flow. These considerations led to a review a few years ago of the state of the art of traffic flow models and eventually to the conclusion that the Payne model (1-6) seemed to be the most realistic and the most developed of the very few dynamic models available at the time. Unfortunately, although it had been tested to some degree, the model had never undergone a comprehensive validation

with actual freeway data. As a result, the decision was made to attempt to calibrate and validate the Payne model by using data from the Queen Elizabeth Way (QEW) Freeway Surveillance and Control System (7). That effort is described in this paper.

Shortly after work started on validation of the Payne model, we became aware of a Boltzmann-type statistical model that was being developed by Phillips (8-10). It is a significant improvement over an earlier model developed by Prigogine (11) and was particularly interesting for two reasons. First, Phillips shows that a family of continuum models of varying levels of refinement can be derived by taking the various moments of the governing stochastic partial-differential equations. One of these continuum models is virtually identical to the Payne model but has some important differences (which affect model performance at high densities) that were later incorporated into the Payne model for comparison purposes. The second reason the Phillips formulation was interesting was because it focused attention on the statistical nature of the calibration and validation process, which in turn provided the basis for developing a realistic methodology.

The results so far have not been encouraging, for reasons that are not yet entirely understood. Nevertheless, it was felt that our experience would possibly be of interest to others working in traffic flow model research and development.

MACK AND FREFLO MODELS

The MACK model was developed by Payne (1-3) as an analytical tool for evaluating ramp control plans and strategies for freeways. It is a macroscopic model that represents traffic flow in terms of aggregate measures such as density, speed, and flow rate.

The FREFLO model (4-6) is a successor to MACK. Both models have the same theoretical foundation.

In both models, the freeway is divided into sections. The time period is divided into uniform time intervals. Each model consists of a set of vehicle equations and a corresponding set of dynamic speed-density equations. Traffic performance data are accumulated for traffic flow in each freeway section defined. The first type of equation expresses the conservation of vehicles:

$$\rho_1^{n+1} = \rho_1^n + (\Delta t / \Delta x_j) (\ell_{j-1} q_j^{n+1} - \ell_j q_{j+1}^{n+1} + f_j^{on, n+1} - f_j^{off, n+1}) \quad (1)$$

where

- Δt = time interval,
- ρ = section density [vehicles/(lane * mile)],
- n = time index,
- q = flow rate across upstream boundary [vehicles/(h * lane)],
- ℓ = number of lanes,
- Δx = section length (miles),
- f_{on} = on-ramp volume (vehicles/h), and
- f_{off} = off-ramp volume (vehicles/h).

The dynamic speed-density equation is

$$u_j^{n+1} = u_j^n - \Delta t \{ u_j^n (u_j^n - u_{j-1}^n) / x_j + (1/T) [u_j^n - u_c(\rho_j^n) + (\nu/\rho_j^n) (\rho_{j+1}^n - \rho_j^n) / \Delta x_j] \} \quad (2)$$

where

- u = section mean speed (miles/h),
- T, ν = relaxation and anticipation parameters, and
- $u_c(\rho)$ = equilibrium speed-density curve.

The three groups of terms express three physical

processes. The first is convection; i.e., vehicles traveling at speed u_{j-1} in the upstream section will tend to continue to travel at that speed as they enter the next section. The second term represents the tendency of drivers to adjust their speeds to the equilibrium speed-density relationship. The third term expresses anticipation of changing travel conditions ahead, i.e., tendency to slow down if the density is perceived to be increasing.

The conservation and dynamic equations are used together to update values from time n to time $n + 1$. Under conditions of uniform flow,

$$q_{j+1}^{n+1} = \rho_j^n u_j^n \quad (3)$$

and this relation is used when updating densities by means of the conservation equations.

PHILLIPS MODELS AND THEIR COMPARISON WITH PAYNE MODEL

The work of Phillips (8-10) introduces a description of traffic analogous to that used in the kinetic theory of gases.

We begin with a brief review of this description.

Let x denote position along the highway and v denote vehicle speed. At each time t , we define a function (essentially a probability density function) $\phi(t, x, v)$ such that $\phi(t, x, v) dx dv$ gives the average number of vehicles found in the interval $(x, x + dx)$ of road with speed in the interval $(v, v + dv)$. We call ϕ the traffic-distribution function. The description of phase space (x, v) is flexible. We could, for example, talk about a distribution function for each lane of a multilane highway or for the truck component of traffic.

The traffic density $K(t, x)$ is found by integrating:

$$K(t, x) = \int_0^\infty \phi(t, x, v) dv \quad (4)$$

For many purposes, it is convenient to work with the function $f(t, x, v)$, which gives the probability that a vehicle randomly chosen at point x will have speed in the range $(v, v + dv)$. It is clear that the following relationship holds:

$$\phi(t, x, v) = K(t, x) f(t, x, v) \quad (5)$$

Phillips (8-10) derives a number of partial-differential equations (or models) satisfied by the moments of the traffic-distribution function. The Phillips model that most closely resembles Payne's is

$$\begin{aligned} (\partial K / \partial t) + (\partial K u / \partial x) &= 0 \\ (\partial u / \partial t) + u(\partial u / \partial x) &= \lambda(u_c - u) - (1/K)(dP/dK)(\partial K / \partial x) \end{aligned} \quad (6)$$

where

- u = mean speed,
- $u_c(K)$ = equilibrium speed-density relation,
- $\lambda(K)$ = delay coefficient, and
- $P(K)$ = traffic-pressure function.

There are both strong similarities and very significant differences between the Payne and the Phillips models. Although these have been discussed by Phillips, we present a more quantitative comparison here.

Since the continuity equation is identical in the two models, we begin by rewriting the u -equations in a way that suggests finite difference conversions. Phillips' equation becomes

$$(\partial u / \partial t) = -u(\partial u / \partial x) - [(dP/dK)(\partial K / \partial x)] - \lambda(K)[u - u_c(K)] \quad (7)$$

Figure 1. Comparison of $\lambda(K)$ and $1/T$.

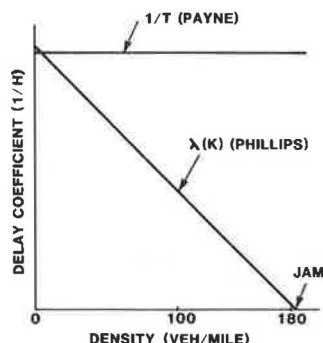


Figure 2. Speed-density curves.

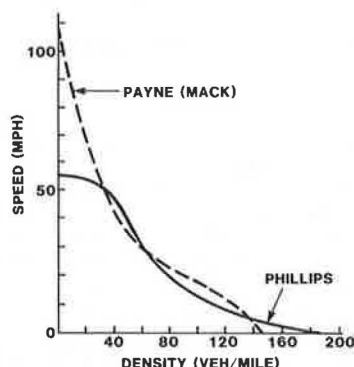
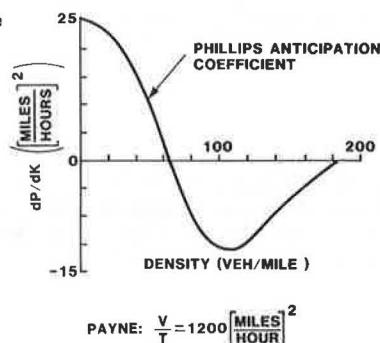


Figure 3. Phillips pressure coefficient.



Payne's formulation is

$$(\partial u / \partial t) = -v(\partial u / \partial x)(v/T)(\partial K / K \partial x) - (1/T)(u - u_e) \quad (8)$$

The tendency-to-equilibrium term is the last one in each of the above equations. A graph showing $\lambda(K)$ and $1/T$ is given in Figure 1. The value of $1/T$ shown is the MACK default value. The Phillips coefficient $\lambda(K)$ is calculated by using the following:

$$\lambda(K) = [u_1(K_0 - K)] / \{L_0[C_0K + C_r(K_0 - K)]\} \quad (9)$$

where K_0 , L_0 , C_0 , and C_r are constants given by Phillips (9, p. 12). The equilibrium speed-density curves for Phillips and Payne are compared in Figure 2. Again, for Payne's curve we use the MACK default curve:

$$u_e(K) = 107K^3 - 2.31K^2 + 0.0215K - 7.4 \times 10^{-5}$$

For Phillips, we use

$$u_e(K) = (u_d u_1^7 + u_1^7 u_d^7 \zeta^6) / (u_1^7 - 9\sigma_d^2 u_1^3 u_d^3 \zeta^3 + u_d^7 \zeta^7) \quad (10)$$

where $\zeta = K/(K_0 - K)$ and u_1 , u_d , and σ_d are constants given by Phillips (9, p. 27).

The two curves are qualitatively similar, but Payne's curve suffers a bit from the limitations imposed by being cubic. In particular, $u_e = 0$ for $K > 145$ is unrealistic.

The second-to-last terms in Equations 7 and 8 also warrant comparison. In Phillips' equation, the term $(dP/dK)(\partial K/K \partial x)$ arises out of traffic-pressure considerations, where (8, p. 64)

$$P(K) = (\sigma_d^2 u_1^2 K) / \{u_1^2 + 14.64 \sigma_d^2 [K/(K_0 - K)]^2\} \quad (11)$$

In Payne's work, the counterpart $(v/T)(\partial K/K \partial x)$ is called the anticipation term. The Phillips pressure coefficient is plotted in Figure 3. We note that $|dP/dK| \leq 25$. Payne's anticipation term is enormous by comparison: $v/T = (3600 \times 5)/15 = 1200$.

With Phillips, it must be noted that $dP/dK < 0$ for $K > 62$ vehicles/mile. This property will cause the mean traffic speed to increase for increasing density in this region. This surely cannot be correct.

This completes the quantitative comparison of Payne's and Phillips' first-order mean speed equations. Later in this paper we discuss the results of replacing the subroutine in MACK, which embodies a finite-difference version of Payne's differential equation. The consequences of the differences in the two equations become evident there.

FREEWAY DATA

The QEW Freeway Surveillance and Control System, located west of Toronto, provides continuous traffic-data collection, traffic-responsive control based on both mainline and ramp conditions, incident detection, hardware-status monitoring, a performance evaluation, and reporting capability. The system has been described in detail by Case and Williams (7). Extensive data and information supplied from 10 mainline detector stations and several ramp-metering installations served as an excellent input for our study of a dynamic freeway flow model.

Specifically, the data and information collected during weekday peak hours in October 1978 have been used.

PRELIMINARY STATISTICAL ANALYSIS

The purpose of this section is to understand the substantial statistical fluctuations in real road data. The Phillips description of traffic is the one we adopt as the basis for our analysis.

Flow Rate

The distributions of detector output given here are for stationary conditions. If the expected flow rate is λ vehicles per unit time, then the probability of counting a vehicle in a small time interval T is λT . This type of situation leads to the Poisson distribution. Its properties are well known. The probability of k vehicles in time T is

$$p(k) = (\lambda T)^k \exp(-\lambda T) / k! \quad (12)$$

The mean of this distribution is

$$E(k) = \sum_{k=1}^{\infty} k p(k) = \lambda T \quad (13)$$

and the variance is

$$\text{Var}(k) = \sum_{k=0}^{\infty} (k - \lambda T)^2 p(k) = \lambda T \quad (14)$$

The traffic-distribution function affects the counterdistribution solely through the constant λ .

We proceed to find the relation of λ to the traffic-distribution function $K(x)f(x,v)$. Since the analysis is done at the fixed point x where the detector is located, we shorten the notation to $Kf(v)$. During a time dt , $v dt Kf(v)dv$ vehicles are counted with speeds in the interval $(v, v + dv)$. The weighting factor v enters because a faster vehicle will be picked up from farther upstream during any given counting interval Δt . Hence, $K dt \int_0^\infty v f(v) dv$ vehicles are counted altogether, so that

$$\lambda = K \int_0^\infty v f(v) dv \quad (15)$$

Speed

The reasoning in the previous paragraph produces the speed distribution as a by-product. The speed-distribution function is $\alpha v f(v)$, where α is chosen to make

$$\int_0^\infty \alpha v f(v) dv = 1 \quad (16)$$

The mean speed observed at the detector counter is

$$v_0 = E(v) = \alpha \int_0^\infty v^2 f(v) dv \quad (17)$$

It is important to note that this number may be different from the actual mean speed, defined as

$$\bar{v} = \int_0^\infty v f(v) dv \quad (18)$$

We note that

$$v_0 = (\bar{v}^2 + \sigma^2)/\bar{v} = \bar{v} + (\sigma/\bar{v})\sigma \quad (19)$$

where σ is the standard deviation of the random variable v . The variance of the observed mean speed is defined to be

$$\alpha \int_0^\infty (v - v_0)^2 v f(v) dv$$

Occupancy

Suppose that $p(l)dl$ gives the fraction of vehicles on the road with measured lengths in $(l, l + dl)$. Since l is the length as measured by detectors, it must include any component contributed by the detector zone. We make the (not-altogether-justified) assumption that length is independent of speed. Then the probability that a passing vehicle has length in $(l, l + dl)$ and speed in $(v, v + dv)$ is $\alpha v f(v)p(l)dvdl$. Such a vehicle contributes an amount

$$L = \max \{ (l/v), \Delta t_{\text{counter}} \}$$

to the occupancy counter, where $\Delta t_{\text{counter}}$ is the counting interval. We make the simplifying assumption that conditions are such that $l/v < \Delta t_{\text{counter}}$ always; that is,

$$f(v) = 0 \quad \text{for } v < l_{\text{max}}/\Delta t_{\text{counter}}$$

With this simplifying assumption, the expected occupancy contribution per vehicle becomes

$$E(L) = \alpha \int_0^\infty dl \int_0^\infty dv (l/v) v f(v) p(l) = \alpha \bar{l} \int_0^\infty f(v) dv = \alpha \bar{l} \quad (20)$$

The variance of L is

$$V(L) = \int_0^\infty dl \int_0^\infty dv [(l/v) - \alpha \bar{l}]^2 v f(v) p(l) \quad (21)$$

The calculations are simplified if we note that the assumed independence of v and l together with $l/v = (1/v)l$ lead to the formulas

$$E(l/v) = E(l)E(1/v) \quad (22)$$

$$E[(l/v)^2] = (E(l^2)E(1/v^2)) \quad (23)$$

from which

$$\begin{aligned} \text{Var}(L) &= E(l^2)E(1/v^2) - [E(l)E(1/v)]^2 \\ &= \text{Var}(l)\text{Var}(1/v) + E(l)^2\text{Var}(1/v) + E(1/v)^2\text{Var}(l) \end{aligned} \quad (24)$$

It can be checked that the standard deviation of a single length measurement is small relative to the mean of such a measurement.

We shall see shortly that the dominant cause of variation in the output of the occupancy counter is the fact that the number of vehicles counted tends to follow the Poisson distribution, except at very high density.

The occupancy counter records the sum S of a random number N (having Poisson distribution) of identically distributed length measurements L_i . This situation leads to the following formulas:

$$\begin{aligned} E(S) &= E(L)E(N) \\ \text{Var}(S) &= \text{Var}(L)E(N) + E(L)^2\text{Var}(N) \\ &= \text{Var}(L) + E(L)^2E(N) \end{aligned} \quad (25)$$

when N has a Poisson distribution. The mean occupancy reading, which uses Equations 13 and 15, is

$$E(S)/T = \lambda E(L) \quad (26)$$

Density/Occupancy Ratio

In using road data to check speed times density = flow rate, we first calculate speed times occupancy. If we substitute mean values for each factor on the left, we should obtain from Equations 19, 20, and 26

$$[\int_0^\infty v^2 f(v) dv / \int_0^\infty v f(v) dv] (\lambda \alpha \bar{l})$$

The estimate of mean flow rate is

$$\lambda = K \int_0^\infty v f(v) dv \quad (27)$$

from Equation 15. Now, consider the following estimate:

$$\text{Flow rate}/(\text{speed} \times \text{occupancy}) = [\int_0^\infty v f(v) dv]^2 / \int_0^\infty v^2 f(v) dv \quad (28)$$

which is a useful estimate for the density/occupancy (d/o) ratio l/\bar{l} only for very narrow distributions $f(v)$ (which is often not the case). The degree of bias in Equation 28 can be better judged from the following formula:

$$\begin{aligned} \text{Density/occupancy} &= (\text{speed} \times \text{density})/(\text{speed} \times \text{occupancy}) \\ &= (1/\bar{l})[\mu^2/(\mu^2 + \sigma^2)] \end{aligned} \quad (29)$$

where μ and σ are the mean and standard deviation of v with respect to the density function $f(v)$. This means that the d/o estimates based on Equation 28 are too low. The correction factor is

$$(\mu^2 + \sigma^2)/\mu^2 = 1 + (\sigma/\mu)^2 \quad (30)$$

Calculations of this correction for road data we have been using yield correction factors < 1.05 in most cases and < 1.10 for very broad distributions.

Figure 4. Evaluation procedure.

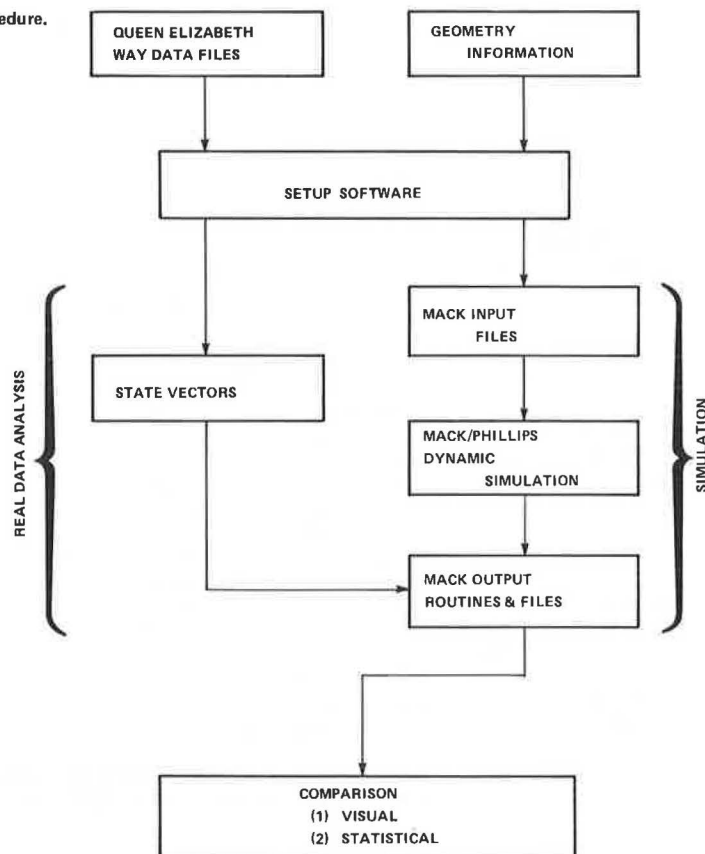
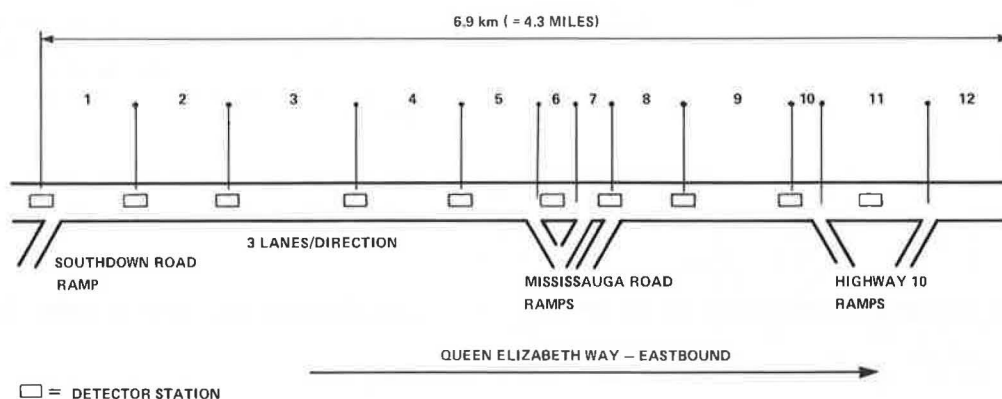


Figure 5. QEW geometry.



EVALUATION PROCEDURE

The evaluation was done by using MACK, a program that incorporates a finite-difference version of Payne's model. Currently, MACK has been superseded by FREFLO. However, the underlying finite-difference scheme describing the traffic dynamics is identical, as can be seen by a comparison of the UPDATE sub-routines in each. Consequently, we would expect similar results for FREFLO.

MACK calculates the traffic-state functions (density and mean speed) along the road at each time interval by using as input the initial state and the upstream on-ramp and off-ramp flow rates as functions of time. This corresponds to an initial-boundary-value problem for the underlying partial-differential equation.

MACK is linked to the QEW road data files by means of a program that calculates the MACK input

information and produces a MACK input data set. The input of road geometry is also done in MACK input format to minimize modifications to MACK software.

The evaluation procedure is summarized in Figure 4, and Figure 5 shows the sectioning geometry and ramp configurations.

STATISTICAL AND VISUAL COMPARISONS

If we assume that the section deviations between MACK output and road data have approximately normal distributions and are independent for different road sections, then the normalized sum of squares of section deviations has a chi-square distribution and standard statistical theory can be used.

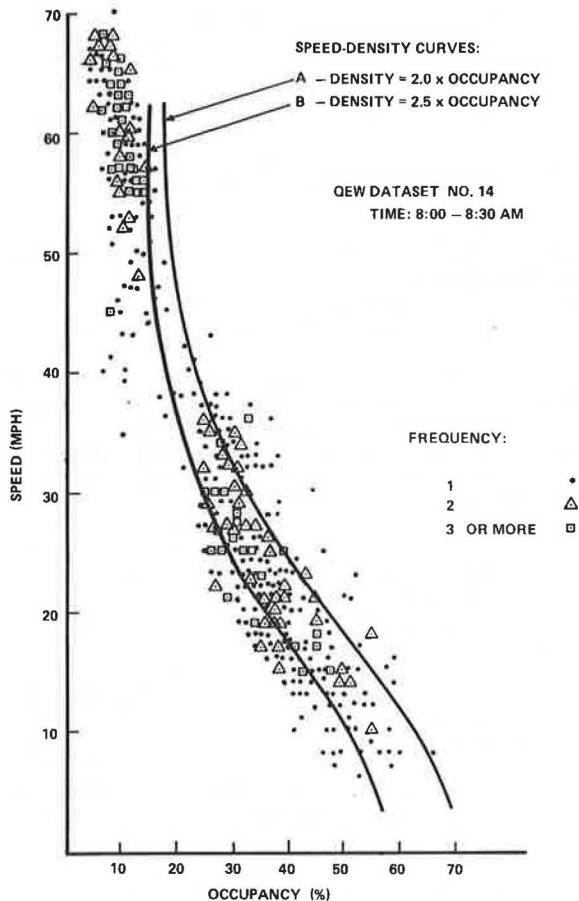
As an example, we apply the test to evaluate how closely the density calculated by MACK agrees with the density derived from occupancy observations for one of our runs. The relevant figures are contained

Table 1. Calculated versus observed densities.

Road Section	Density Observed (vehicles/mile)	Density Calculated (vehicles/mile)	Estimated Standard Observed Density
1	14.0	15.7	4.0
2	18.0	16.3	4.0
3	25.0	17.3	4.1
4	53.0	18.1	4.2
5	79.3	18.6	4.2
6	86.4	18.7	4.3
7	85.0	20.0	4.5
8	76.9	21.2	4.6
9	67.0	21.3	4.6
10	62.6	21.4	4.6
11	60.5	19.4	4.4
12	60.0	23.1	4.8

Note: 2-min data, fixed time; $\chi^2 = \sum[(\text{observed} - \text{expected})/\text{standard}]^2 = 5353$.

Figure 6. Speed-occupancy data.



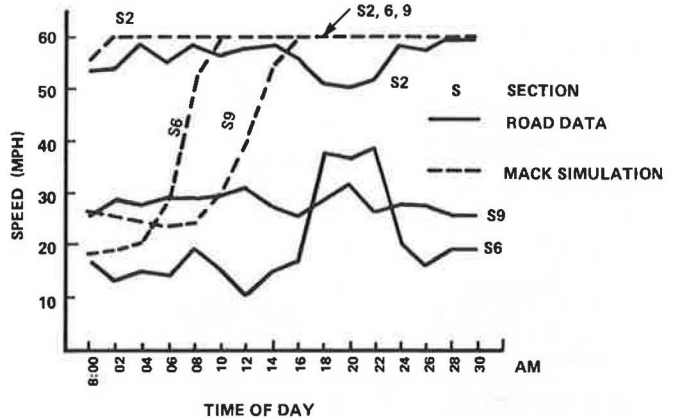
in Table 1. A value of χ^2 in this range has extremely small probability (< 0.005), which can be seen from a standard chi-square table with $f = 12$. The hypothesis is that the values calculated by MACK are the correct section density values. Under the circumstances, a reasonable conclusion is that MACK is not correct.

A visual comparison is possible because MACK output includes a number of graphs and a density plot. Since our procedure generated road data output in the same format, a great deal can be learned by simply comparing plots.

MODIFIED MACK

When difficulties were experienced with MACK, it was

Figure 7. Speeds.



decided to try a modified version prepared by replacing the original finite-difference scheme with one based on Equation 7. The relevant formulas are as follows:

$$\rho_j^{n+1} = \rho_j^n + [(\ell_{j-1} q_j^{n+1} - f_j^{off, n+1} - \ell_j q_{j+1}^{n+1} + f_j^{on, n+1}) \Delta t / \ell_j \Delta x_j] \quad (31)$$

$$q_j^{n+1} = \rho_j^n u_j^n \quad (32)$$

$$u_j^{n+1} = u_j^n - [u_j^n \{ (u_j^n - u_{j-1}^n) / [(\Delta x_j + \Delta x_{j-1}) / 2] \} + \{ \lambda(\rho_j) u_j^n - u_c(\rho_j^n) + (dP/dK)(\rho_j^n) \} (\rho_{j+1}^n - \rho_j^n) / [\rho_j^n (\Delta x_j + \Delta x_{j-1}) / 2] \} \Delta t] \quad (33)$$

where λ , u_c , and (dP/dK) are as defined by Phillips.

RESULTS

We rely on visual comparisons in this presentation because they suffice for indicating the magnitude of the deviation between MACK output and road data. We describe selected output from a typical sequence of runs. The d/o ratio used is 2.0. This value of the d/o ratio has been standardized in all our runs in order to come as close as possible to preserving the continuity relationships: flow = speed times density. It can be seen from Figure 6 that this value of the d/o ratio leads to a speed-density curve that is too far from the origin. This means, of course, that the equilibrium speed at any given density is too high.

Figure 7 shows selected mean speed graphs for real road data. The corresponding output for MACK that uses the default equilibrium speed-density curve gives the same information from MACK output. The key feature to be noted is that the mean speeds increase and the road empties as time progresses in the MACK output, whereas the road data exhibit fairly steady conditions. The same pattern in terms of densities can be seen in Figure 8a and b.

In an attempt to calibrate MACK, the speed-density curve was shifted by scaling the density-occupancy axis. The effects of scaling are shown in Figure 6. A larger scaling constant produces a curve that is closer to the origin, that is, one that yields a lower speed at fixed density. The best fit of speed-density curve to our observed scatter diagrams was obtained with a scaling constant of 1.25 (i.e., density = 2.5 times occupancy). Use of the best-fit speed-density curve did not correct the road-clearing effects exhibited by MACK output (see Figure 8c).

Figure 8. Densities.

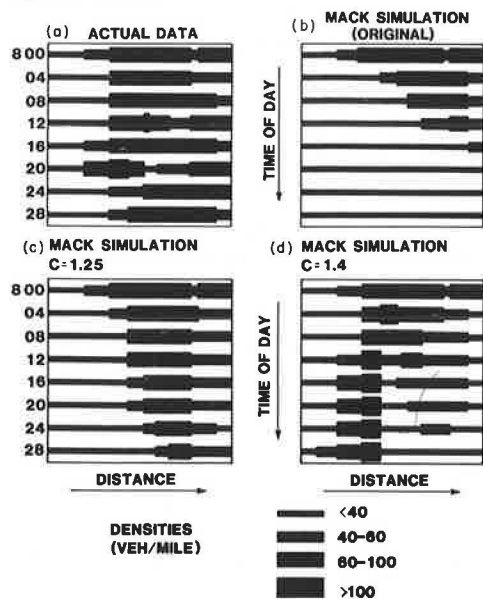
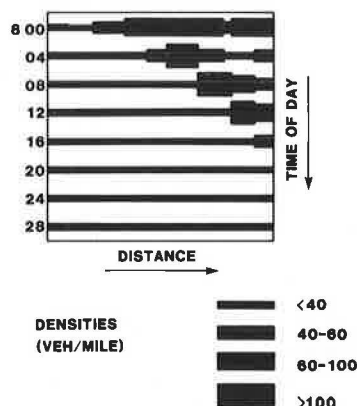


Figure 9. Densities: Phillips simulation.



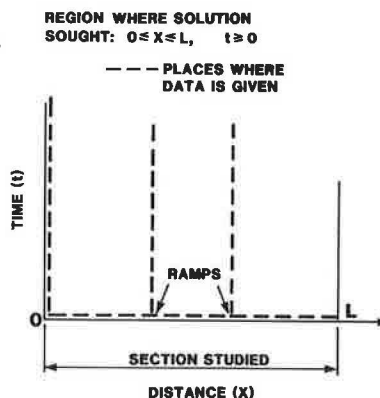
Next, an attempt was made to see whether MACK would track road data correctly for some speed-density curve, even if it did not fit the usual definition of such a curve. Figure 8d shows a density plot for a scaling constant of 1.4 (i.e., density = 2.8 times occupancy). Downstream and upstream clearing is still evident. In addition, however, density concentrations as in a traffic jam have appeared at section 4. When the scaling constant is increased to 1.4, the density concentration continues to coexist with upstream and downstream road clearing except that now an actual shock wave forms that moves upstream.

Finally, the performance of our modification of MACK based on Phillips' differential equation was tested to check whether the differences in terms already discussed could account for the poor MACK tracking properties. Figure 9 shows a density plot for the same run as the one presented in Figure 8. Once more we see that abnormally low and high densities coexist even for the speed-density relation provided by Phillips.

DISCUSSION AND CONCLUSIONS

The behavior we have witnessed in this series of runs is suggestive of an instability in the simula-

Figure 10. Initial-boundary value outline.



tion method. There are several areas where the simulation method, which is actually a solution algorithm for a system of first-order partial-differential equations, needs a more detailed study. The handling of the downstream boundary condition is responsible for the fact that density decrease and associated increases in speed move from the downstream to the upstream end in all cases. It is also conceivable that the handling of the merging process has a significant effect on the relationship between speed and density. This effect does not seem to be large enough to cause the simulation error to concentrate downstream of on ramps. These occur in the present geometry in sections 1, 7, and 12.

It is also conceivable that a careful adjustment of the speed-density curves on a section-by-section basis may improve the fit between the simulation and road data for one particular data set. We must remember, however, that the qualitative character of the simulation has shown itself remarkably sensitive to the simulator speed-density relation. The work simulator is emphasized because the curve that must be used to retard the road-clearing phenomena is far below the picture of speed versus density that one obtains in plots of real road data. Then, within a tiny range of curves, the shock phenomena set in.

Each of these phenomena must be carefully examined with regard to its statistical properties in designing a simulator that will properly track road data. They fall into the realm of fine tuning. We do not believe that they contain the seeds of the explanation and correction of the performance of MACK or for that matter any simulator based on a finite-difference method for solving a system of first-order differential equations governing mean density and speed. We hasten to note that for the MACK modification by using a differential equation based on Phillips' theory, the results were quantitatively worse than those for MACK.

Well-Posedness of Partial-Differential Equation

"Well-posedness" means that the solution must be unique and depend on the given data in a continuous fashion.

We observe that either of the first-order models is a 2-x-2 system of first-order partial-differential equations. The intuitively reasonable initial-boundary value problem (IBVP) is outlined in Figure 10. That is, $u(x,t)$ and $K(x,t)$ are to be found, given initial states on $t=0$, $0 < x < L$, given upstream traffic data (state) for $x=0$, $t > 0$, and given on/off-ramp data for appropriate values of x and $t > 0$.

Strictly speaking, the well-posedness of the IBVP needs proof. The method of characteristics is available to handle such questions for first-order

equations, and although our intuition about traffic flow strongly backs a conclusion that the IBVP is well posed, a study along these lines should be undertaken at some point in the light of the difficulties we have experienced.

Finite-Difference Scheme

In examining this aspect of a simulator, we must remember that we are in fact trying to solve a stochastic problem in partial-differential equations.

Several key questions emerge in looking at a finite-difference solution to the first-order IBVP:

1. Given steady boundary conditions, does there exist a steady-state solution to the IBVP?
2. Does the time-dependent state approach the steady state from any initial value?
3. Do the effects of a momentary perturbation in a boundary condition die out as they would in a real traffic situation? (This is a key stability question.) If they do not, then the statistical fluctuations inherent in real road data make them unusable as initial and boundary data.

The use of MACK or FREFLO to track real road data in fact requires deeper knowledge than the above. We must, in fact, know something about the distribution of state variables generated from distribution of input data. These distributions should reflect stability of the finite-difference scheme as already mentioned and they are needed to devise meaningful tests to determine whether the simulator is tracking faithfully. Techniques are available for tackling these questions, and at this point, it seems that they are well worth trying.

Concluding Remarks

The overall conclusion must be that MACK and FREFLO, by virtue of their identical underlying differential equation, exhibit instabilities in their behavior that make them unsuitable for use as simulators tracking real road data in the sense of solving an IBVP. Furthermore, it appears that the difficulties cannot be corrected by using the first-order differential equation derived by Phillips.

Numerous reasons can be advanced for the observations we have made, but two are dominant in importance. First, road traffic has a very strong and complex stochastic aspect. Fluctuations on a moderate time scale of 5-10 min are very high relative to the mean size of quantities being measured. These fluctuations of necessity find their way into any scheme designed to simulate road behavior and wreak havoc if there are any instabilities present. In fact, for meaningful results we must ask not just for the absence of instabilities--that is, continuous dependence on data and parameters--but for the presence of strong stability properties to cause decay of the influences of earlier fluctuations.

The second reason is that the concepts involved in obtaining a differential equation and an IBVP for the mean speed function lead to a very sensitive dependence on the speed-density relation and possibly other parameters. Once the upstream flow rate and the on/off ramp flow rates are known, a time-averaged flow rate for each road position is a consequence. Suppose that at some point the speed-density relationship is too slow to move a sufficient number of vehicles over the longer time scale of 5-10 min. Then densities will increase upstream of the point in question to conserve vehicles. This effect will depress the road speed even further, ultimately starting the typical shock wave moving upstream. Since such effects are cumulative, it is

clear that they will be observed even if the speed-density relation is mismatched by a tiny amount. Conversely, it is clear that if the speed-density relation used by the simulator is a bit too large, we will find the scenario of less density and more speed, which moves ultimately to clear the road of traffic.

Discussion

Harold J. Payne

This paper contributes to the examination of the FREFLO model necessary to reach a judgment concerning its ability to represent freeway traffic flow and to indicate steps necessary to make use of the model.

In this discussion, three points are addressed:

1. The calibration/validation process necessary before using FREFLO,
2. Applicability and restrictions in the use of FREFLO, and
3. Recent model improvements.

The authors describe an effort to calibrate FREFLO for application to the QEW freeway in Toronto. This entailed collection of data, rectification of measurements to model variables (outputs), and then a cycle of execution of the FREFLO model, assessment of match of model results to data, and adjustments to FREFLO. Similar efforts have been successfully undertaken in connection with NCHRP Project 3-22A, Guidelines for Design and Operation of Ramp Control Systems (12). In that study, FREFLO was calibrated for freeways in Los Angeles and Dallas. The calibrated model subsequently played a major role in that study.

A key element of that calibration, and one that was unfortunately and unnecessarily lacking in the Toronto work, was the involvement of the FREFLO model builders in the calibration effort. At this time, the special requirements for effective use of FREFLO are not widely known or fully documented, so this type of involvement is very important.

In the instance of the work reported in this paper, a critical model restriction was not properly observed. The situation is depicted in Figure 11. The downstream extent of the congestion modeled by FREFLO must be completely interior to the freeway segment modeled; some uncongested zone must exist at the downstream end. (There are adequate techniques to deal with congestion at the upstream end, however.) Such was not the case in the Toronto work. As a consequence, all calibration efforts were doomed to failure.

As a final point, some recent work (13) has revealed that a discontinuous speed-density relationship, of the sort depicted in Figure 12, more

Figure 11. Critical restriction on use of FREFLO.

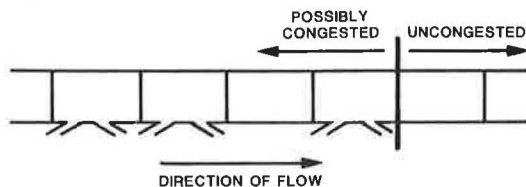
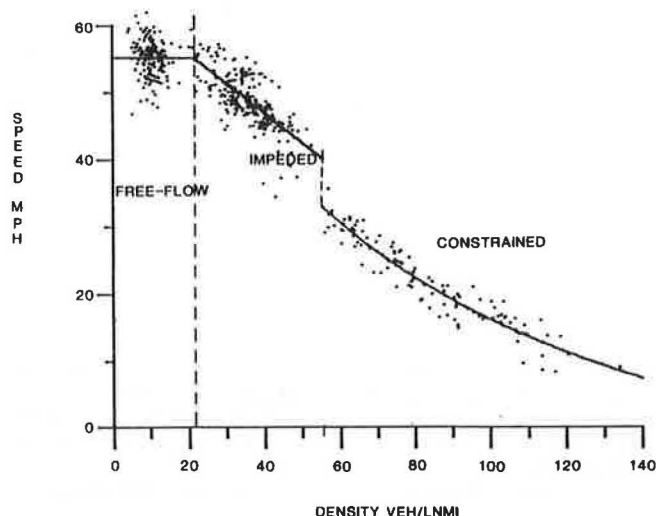


Figure 12. Recent improvements to FREFLO.



accurately reflects the equilibrium behavior of freeway traffic. The discontinuity has significant impacts on the ability of FREFLO to model geometric discontinuities. It may also be very important to the study of ramp metering (13).

Authors' Closure

We would like to thank Payne for his comments.

During the study, and especially in its initial stage, the FREFLO model builder was involved in our work. We had considerable documentation and even the benefit of his personal instructions.

We would like to point out that the behavior exhibited by the model was very similar to that observed in an earlier FREFLO simulation (14) of an idealized freeway with a bottleneck in the middle. The situation with uncongested downstream sections was simulated by using simplified hypothetical test cases, but the results exhibited incorrect patterns similar to those shown by Payne (6). Many different "cures" to the problem were tried, including those suggested by Payne, but to no avail. Finally, it was concluded that the hypothetical speed-density curve used was unrealistic and the decision was made to proceed with an evaluation based on data obtained from an actual system, the subject of this paper.

For a model to be of any practical value, the model builder must strive to develop ones that do

not require any additional involvement by the builder during implementation and application stages.

REFERENCES

1. H.J. Payne. Models of Freeway Traffic and Control. Simulation Council Proc., Mathematical Models of Public Systems, Vol. 1, No. 1, 1971, pp. 51-61.
2. H.J. Payne, W.A. Thompson, and L. Isaksen. Design of a Traffic-Responsive Control System for a Los Angeles Freeway. IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-3, No. 3, May 1973, pp. 213-224.
3. D.N. Goodwin, S.D. Miller, and H.J. Payne. MACK: A Macroscopic Simulation Model of Freeway Traffic. Technology Service Corporation, Santa Monica, CA, July 1974.
4. H.J. Payne. FREFLO: A Macroscopic Simulation Model of Freeway Traffic--Version I: User's Guide. ESSCOR, San Diego, CA, July 1978.
5. H.J. Payne. FREFLO: A Macroscopic Simulation Model of Freeway Traffic--Version I: Program Documentation. ESSCOR, San Diego, CA, July 1978.
6. H.J. Payne. FREFLO: A Macroscopic Simulation Model of Freeway Traffic. TRB, Transportation Research Record 722, 1979, pp. 68-75.
7. E.R. Case and K.M. Williams. Queen Elizabeth Way Freeway Surveillance and Control System Demonstration Project. TRB, Transportation Research Record 682, 1978, pp. 84-93.
8. W.F. Phillips. Kinetic Model for Traffic Flow. U.S. Department of Transportation, Rept. DOT/RSPD/DPB/50-77/17, 1977.
9. W.F. Phillips. A New Continuum Model for Traffic Flow. Utah State Univ., Logan, June 1979.
10. W.F. Phillips. A Kinetic Model for Traffic Flow with Continuum Implications. Transportation Planning and Technology, Vol. 5, 1979, pp. 131-138.
11. I. Prigogine. A Boltzmann-Like Approach to the Statistical Theory of Traffic Flow. In *Theory of Traffic Flow* (R. Herman, ed.), Elsevier, Amsterdam, Netherlands, 1961, pp. 158-164.
12. Guidelines for Selection of Ramp Control Systems. NCHRP, Rept. 232, 1981.
13. H.J. Payne. The Discontinuity in Equilibrium Freeway Traffic Flow: Empirical Findings and Implications for Control. Presented at National Meeting of the Operations Research Society of America, San Diego, CA, 1982.
14. E. Hauer and V.F. Hurdle. Discussion of paper by Payne--FREFLO: A Macroscopic Simulation Model of Freeway Traffic. TRB, Transportation Research Record 722, 1979, pp. 75-76.

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

Passenger-Car Equivalents for Rural Highways

WILEY D. CUNAGIN AND CARROLL J. MESSER

The objective was to determine the passenger-car-equivalent (PCE) value for 14 different vehicle types under varying conditions of traffic and roadway geometry. This was accomplished by analyzing field data collected in several states on both two-lane and four-lane highways. Data included headways, speeds, and travel times by vehicle type, traffic-volume condition, and roadway-section type. An analytical model was developed to estimate PCE values based on speed distributions, traffic volumes, and vehicle types. The calibrated model was used to estimate PCE values for 14 vehicle types under specified typical conditions for two-lane and four-lane rural highways. Composite PCE values for a range of geometric, volume, and percentage-of-truck values are presented and compared with values from other research.

The utility of a rural highway in serving traffic is a function not only of the geometric characteristics of the highway but also of the composition of the traffic. The presence of large and/or low-performance vehicles in the traffic stream reduces the total number of vehicles that can use the highway. The effects of trucks and other low-performance vehicles may be equated to an equivalent number of passenger cars added to the traffic stream if appropriate analytical models of traffic flow, vehicle-operating characteristics, and field validation of data are available. Passenger-car equivalents (PCEs) may be based on a consideration of passing, speed, occupancy, or capacity impacts.

This paper describes the methodology and results of a study done for the U.S. Department of Transportation (1). The results of this study were submitted for use in the Cost-Allocation Study of the Federal Highway Administration (FHWA), which has proposed an adjustment in the allocation of highway costs among the several classes of vehicles. Rural two-lane two-way highways and two-lane one-way (i.e., four-lane) facilities were studied in Texas, North Carolina, Pennsylvania, Colorado, and West Virginia. Flat, moderate, and steep grade geometrics were studied for both types of highways. An attempt was made to collect traffic speed and headway data at all volume levels. PCE values for 14 vehicle types were determined both empirically and theoretically by using an analytical model developed in this research.

BACKGROUND

PCE values have been used primarily in the framework of traffic-capacity procedures. PCE values are employed as a device to convert a traffic stream composed of a mix of vehicle types into an equivalent traffic stream composed exclusively of passenger cars. The availability of such values permits the specification of capacity in terms of PCEs exclusively and provides the basis for development of procedures to express any traffic-stream composition in terms of PCEs. By applying such procedures, an analyst can directly convert an existing (or projected) traffic volume composed of a mix of vehicle types into an equivalent PCE volume, which can then be compared with specified PCE (capacity) service volumes. These methods, whether applied to two-lane two-way highways or two-lane one-way (i.e., four-lane) highways, have generally been those described in the 1965 Highway Capacity Manual (HCM) (2).

The HCM gives both generalized and specific values of PCEs. Average generalized PCEs of trucks are given in the HCM as related to generalized terrain conditions and levels of service. PCEs are also provided for trucks on specific individual

highway grades. The generalized PCEs range in value from 2 to 12. The specific-grade PCEs range from 2 to 108.

The literature review carried out as a part of this research showed consistent agreement that the PCEs presented in the 1965 HCM for specific grades on two-lane highways are too high, especially for large percentages of trucks on steep grades. There was less than unanimous agreement on the issue of whether the incremental negative impact of adding trucks to the traffic stream decreases as the percentage of trucks in the traffic stream increases. The PCEs in the multilane section of the HCM vary greatly and in a complex manner.

The two-lane and four-lane PCE values will be addressed separately in the following sections.

TWO-LANE HIGHWAYS

The basic HCM (2) equation for the conversion of a mixed-traffic-stream volume to an all-passenger-car volume for two-lane highways is as follows:

$$SV_L = 2000 \cdot (v/c)_L \cdot W_L \cdot T_L \cdot B_L \quad (1)$$

where

SV_L = maximum volume for a given level of service (LOS) L (total for both directions),

$(v/c)_L$ = volume-to-capacity ratio at LOS L,

W_L = adjustment factor for lane width and lateral clearance at LOS L,

T_L = truck adjustment factor at LOS L, and

B_L = bus adjustment factor at LOS L.

The truck adjustment factor (T_L) is calculated from the following equation:

$$T_L = 1/[1 + P_T(E_L - 1)] \quad (2)$$

where P_T is the decimal fraction of trucks in the traffic stream and E_L is the PCE for trucks at LOS L.

The PCE values for trucks operating on grades of a given length and percent are calculated in the HCM (2) from separate speed distributions of passenger cars and trucks at a given volume level. The criterion used is the relative number of passings of trucks by passenger cars in relation to the number of passings of passenger cars by passenger cars. The specific method is known as the Walker method. It had been mentioned earlier by both Normann (3) and Wardrop (4) but is named for the man who applied it in the 1965 HCM.

The Walker method is used to calculate PCEs for trucks operating on grades. The separate speed distributions of passenger cars and trucks at a given traffic volume are used to compute the relative number of passings that would have been performed per mile of highway if each vehicle continued at its normal speed for the conditions under consideration. A gradability (speed versus distance up a sustained grade) curve for a weight/horsepower ratio of approximately 325 lb/hp, considered typical of conditions on two-lane highways carrying a variety of trucks [Figure 1 (2)], was used to develop an average speed over grades of varying steepness and length as shown in Figure 2 (2). When this average

Figure 1. Effect of length and steepness of grade on speed of average truck on two-lane highway.

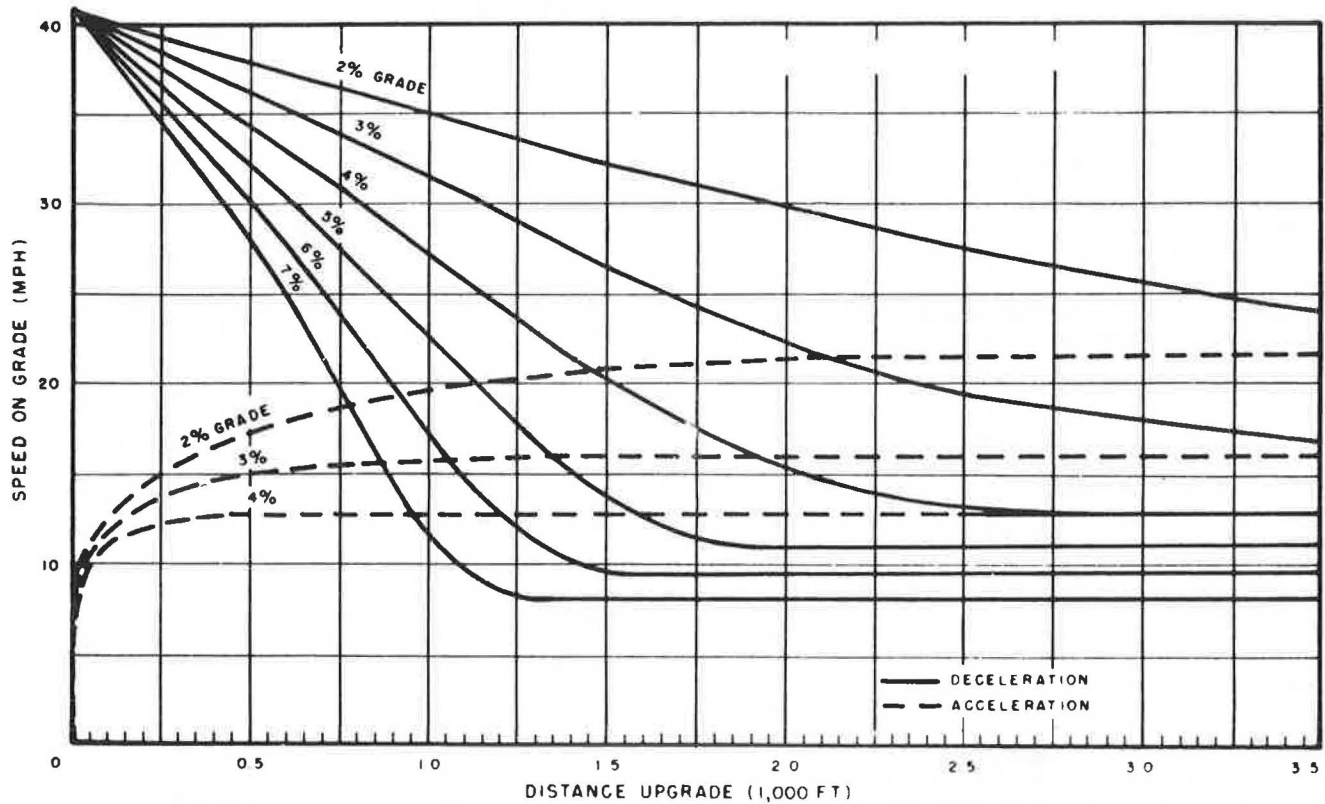
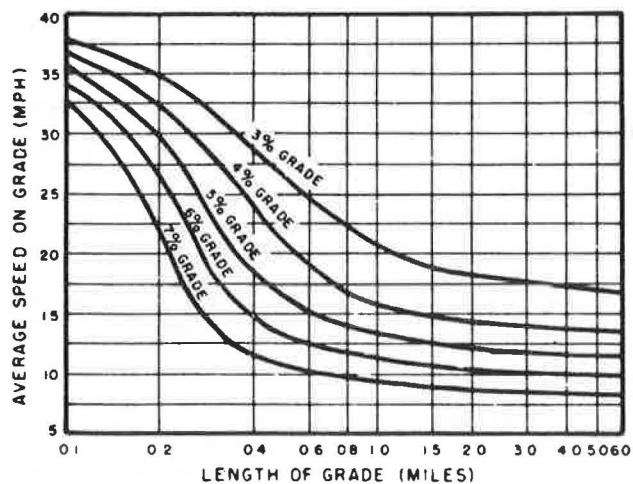


Figure 2. Mean speed of average truck over entire length of grade on two-lane highway.



speed was known, Walker's method was used to calculate PCEs as may be seen in Figure 3 (2). Table 1 contains the HCM PCEs calculated from this methodology for specific terrain conditions. The generalized HCM PCEs are shown below. No PCE values were given for recreational vehicles.

Equivalent E_T (truck)	Level of Service	PCE for		
		Level Terrain	Rolling Terrain	Mountainous Terrain
E_B (bus)	A	3	4	7
	B and C	2.5	5	10
	D and E	2	5	12
E_B (bus)	All	2	4	6

Figure 3. PCEs for various average truck speeds on two-lane highway.

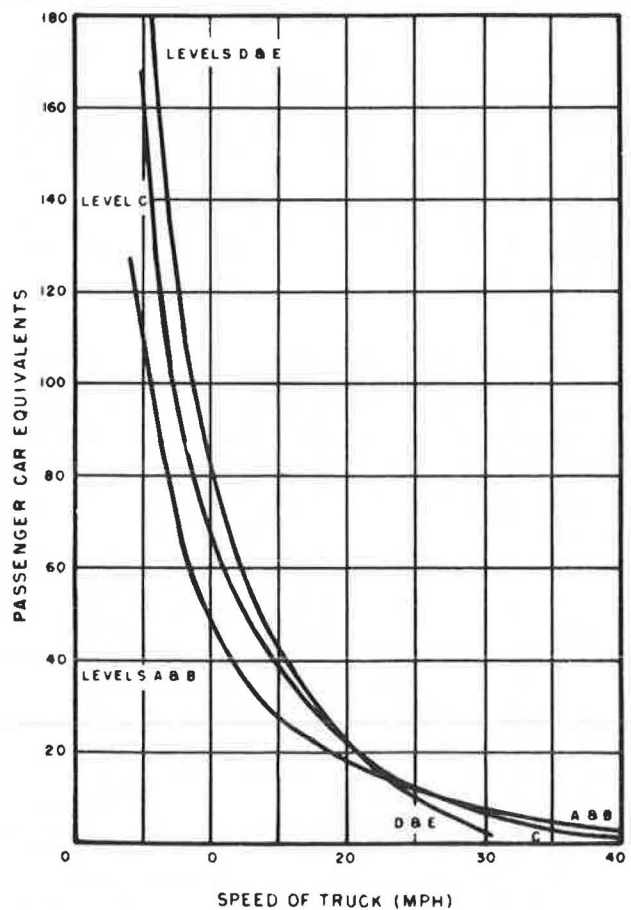


Table 1. PCEs of trucks on two-lane highways on specific individual subsections or grades.

Grade (%)	Length of Grade (miles)	PCE for All Percentages of Trucks (E_T) by Level of Service		
		A and B	C	D and E (Capacity)
0-2	All	2	2	2
3	0.25	5	3	2
	0.50	10	10	7
	0.75	14	16	14
	1	17	21	20
	1.50	19	25	26
	2	21	27	29
4	3	22	29	31
	4	23	31	32
	0.25	7	6	3
	0.50	16	20	20
	0.75	22	30	32
	1	26	35	39
5	1.50	28	39	44
	2	30	42	47
	3	31	44	50
	4	32	46	52
	0.25	10	10	7
	0.50	24	33	37
6	0.75	29	42	47
	1	33	47	54
	1.50	35	51	59
	2	37	54	63
	3	39	56	66
	4	40	57	68
7	0.25	14	17	16
	0.50	33	47	54
	0.75	39	56	65
	1	41	59	70
	1.50	44	62	75
	2	46	65	80
8	3	48	68	84
	4	50	71	87
	0.25	24	32	35
	0.50	44	63	75
	0.75	50	71	84
	1	53	74	90
9	1.50	56	79	95
	2	58	82	100
	3	60	85	104
	4	62	87	108

The spatial-headway method uses the relative amount of space "consumed" by a vehicle to determine its PCE. This method was recently applied by the Institute for Research (IFR) to obtain PCE values on urban freeways (5).

The equivalent-delay method uses the ratio of the delay experienced by a passenger car due to nonpassenger cars to the delay experienced by a passenger car due to other passenger cars. The delay caused to standard passenger cars due to lower-performance passenger cars is explicitly considered.

Craus (6) used the equivalent-delay concept to determine PCE values on two-lane highways based on both the Walker method and the delay to vehicles due to opposing traffic. Craus suggests that the PCE values given in the 1965 HCM are 34 percent too high for 10-mph trucks at LOS A, 40 percent too high at LOS C, and 46 percent too high at LOS E. Craus' model also predicts that PCEs will increase with increasing volume.

FOUR-LANE HIGHWAYS

The basic equation for the conversion of mixed-traffic-stream volume on a multilane rural highway to an all-passenger-car volume is as follows:

$$SV_L = 2000 \cdot N \cdot (v/c)_L \cdot W_L \cdot T_L \quad (3)$$

Table 2. PCEs of trucks on ordinary multilane highways on specific individual subsections or grades.

Grade (%)	Length of Grade (miles)	PCE (E_T)									
		LOS A Through C by Percentage of Trucks					LOS D and E (Capacity) by Percentage of Trucks				
		3	5	10	15	20	3	5	10	15	20
0-1	All	2	2	2	2	2	2	2	2	2	2
2	0.25-0.50	5	4	4	3	3	5	4	4	3	3
	0.75-1	7	5	5	4	4	7	5	5	4	4
	1.50-2	7	6	6	6	6	7	6	6	6	6
	3-4	7	7	8	8	8	7	7	8	8	8
	0.25	10	8	5	4	3	10	8	5	4	3
	0.50	10	8	5	4	4	10	8	5	4	4
3	0.75	10	8	6	5	5	10	8	5	4	5
	1	10	8	6	5	6	10	8	6	5	6
	1.50	10	9	7	7	7	10	9	7	7	7
	2	10	9	8	8	8	10	9	8	8	8
	3	10	10	10	10	10	10	10	10	10	10
	4	10	10	11	11	11	10	10	11	11	11
4	0.25	12	9	5	4	3	13	9	5	4	3
	0.50	12	9	5	5	5	13	9	5	5	5
	0.75	12	9	7	7	7	13	9	7	7	7
	1	12	10	8	8	8	13	10	8	8	8
	1.50	12	11	10	10	10	13	11	10	10	10
	2	12	11	11	11	11	13	12	11	11	11
5	3	12	12	13	13	13	13	13	14	14	14
	4	12	13	15	15	14	13	14	16	16	15
	0.25	13	10	6	4	3	14	10	6	4	3
	0.50	13	11	7	7	7	14	11	7	7	7
	0.75	13	11	9	8	8	14	11	9	8	8
	1	13	12	10	10	10	14	13	10	10	10
6	1.50	13	13	12	12	12	14	14	13	13	13
	2	13	14	14	14	14	14	15	15	15	15
	3	13	15	16	16	15	14	17	17	17	17
	4	15	17	19	19	17	16	19	22	21	19
	0.25	14	10	6	4	3	15	10	6	4	3
	0.50	14	11	8	8	8	15	11	8	8	8
7	0.75	14	12	10	10	10	15	12	10	10	10
	1	14	13	12	12	11	15	14	13	13	11
	1.50	14	14	14	14	13	15	16	15	15	14
	2	14	15	16	16	15	15	18	18	18	16
	3	14	16	18	18	17	15	20	20	20	19
	4	19	19	20	20	20	20	23	23	23	23

in which SV_L is the service volume (mixed vehicles per hour, total for one direction) for LOS L.

The PCE values for multilane highways in the 1965 HCM were based on the relative delay due to trucks, which was determined by the Walker method in conjunction with gradability curves.

The generalized PCEs for multilane highways suggested by the 1965 HCM are shown below. (Separate consideration of E_B (buses) is not warranted in most problems; E_B is used only where bus volumes are significant.)

Level of Service	PCE for		
	Level Terrain	Rolling Terrain	Mountainous Terrain
A	Widely variable; one or more trucks have same total effect, causing other traffic to shift to other lanes; use equivalent for remaining levels in problems		
B through E	2	4	8
	1.6	3	5

The PCEs for specific terrain and traffic-volume conditions are shown in Table 2. The PCE values in Table 2 generally decrease for increasing percentages of trucks.

Researchers at the Polytechnic Institute of New York (PINY) used simulation data from a model developed by the Midwest Research Institute (MRI) (7) to develop truck equivalents for varying percentages of trucks on any severity of sustained grade (8). IFR

recently completed a study to determine PCEs by the spatial-headway method for trucks and other nonpassenger cars on urban freeways (5).

The equivalent-delay method was applied in this research. On four-lane highways, overtaking vehicles are inhibited only by concurrent-flow traffic. Faster vehicles may pass at will except when obstructed by slower vehicles passing still slower vehicles. This concept was used by Newman and Moskowitz in a study for the California Department of Transportation and formed the basis for the PCE values used in the 1965 HCM (2,9). In this research,

$$PCE_{ij} = (D_{ij} - D_{base})/D_{base} \quad (4)$$

where

- PCE_{ij} = PCE of vehicle type i under condition j ,
 D_{ij} = delay to passenger cars due to vehicle type i under condition j , and
 D_{base} = delay to standard passenger cars due to slower passenger cars.

MAJOR FACTORS INFLUENCING PCEs

Determination of PCEs in this study included the following factors: roadway geometrics (up and down roadway grades, the length of grade, and the number of lanes), vehicle performance characteristics (length of vehicle, weight, and horsepower), and traffic flows (volume demand, directional split on two-way two-lane roads, percentage of vehicle types, and vehicle speeds).

Geometric Factors

The range of grades considered extends from 0 to 7 percent. Three levels of grades were defined:

1. Level (0-1 percent),
2. Moderate (2-4 percent), and
3. Steep (5-7 percent).

The 1965 HCM terrain descriptions (flat, rolling, mountainous) were purposely avoided since the terrain descriptions are generalizations of average conditions.

Two grade conditions were studied at each data-collection site. At each site, measurements were taken on a level section of the road and on a moderate or steep grade in the immediate vicinity.

Vehicle Performance Factors

The terminology "vehicle performance" is used to describe the speed capabilities of individual vehicles operating along a road of given geometrics. The term "operations" is used to describe the collective behavior of a mix of vehicle types in the traffic stream. The vehicle performance capabilities of the following 14 vehicle types were considered:

1. Base automobile (standard and compact),
2. Small automobile (subcompact),
3. Motorcycle,
4. Bus (intercity, school, transit),
5. Single-unit truck with two axles and four wheels (pickups, vans, delivery),
6. Single-unit truck with two axles and six wheels (various weights),
7. Single-unit truck with three or more axles (various weights),
8. Three-axle truck combination (2S2, 2-1, various weights),
9. Four-axle truck combination (2S2, various weights),
10. Other four-axle truck combinations (3-1, 2-2, 3S1, 2S1, various weights),
11. Five-axle truck combination (3S2, various weights),
12. Other five-axle truck combinations (2S3, 3-2, 2S1, 2-3, various weights),
13. Truck combinations with six axles or more (3-3, 2S2-2, 3S2, various weights), and
14. Recreational vehicles--car and trailer, motorhome, pickup camper.

- Except for the last category listed, these are the case-study visual categories for the FHWA Highway Performance Monitoring System.

Traffic Flow Factors

Volume levels were chosen to relate approximately to level of service. The following volume ranges correspond to the indicated levels of service for uninterrupted flow on two-lane one-way (four-lane) highways in the 1965 HCM:

Volume (vehicles/h)	Approximate Level of Service	Volume Level
0-600	A	1
601-1000	B	2
1001-1500	C	3
1501-1800	D	4
1801-2000	E	5
Over capacity	F	6

The maximum service volumes for two-lane two-way highways under uninterrupted flow conditions are given by the 1965 HCM in terms of the total traffic in both directions as follows:

Maximum Service Volume (both directions) (vehicles/h)	Approximate Level of Service	Volume Level
400	A	1
900	B	2
1400	C	3
1700	D	4
2000	E	5
Over capacity	F	6

By defining the volume ranges in terms of levels of service, five volume levels (1 through 5) were used.

PCE Matrix

A three-dimensional PCE matrix was constructed. The dimensions of the matrix are the 14 vehicle types listed previously, 6 rural highway section types (one or two lanes per direction times level, moderate, or steep grade), and 5 traffic-volume levels.

The 14 vehicle types, 6 rural highway section types, and 5 volume levels yield a matrix containing 420 cells. As might be expected, sufficient data could not be collected for all of the cells in the matrix. Volume levels above volume level 3 were not observed except at one two-lane one-way site with a moderate grade (Charlotte, North Carolina).

DATA COLLECTION

Data-collection sites were chosen to attempt to collect data that encompass the full range of vehicle types, volume levels, and rural highway section types (both by number of lanes and alignments) listed in the previous section.

Figure 4. Data-collection configuration.

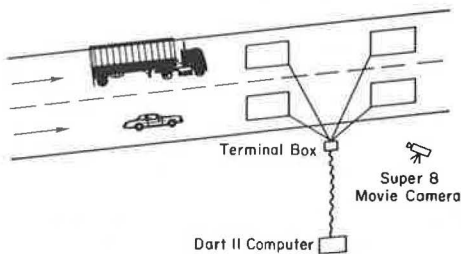
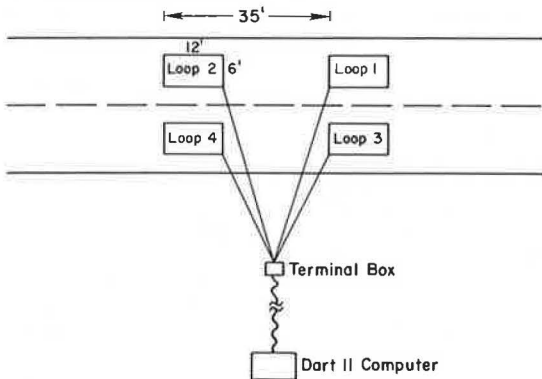


Figure 5. Loop configuration.



Site-Selection Criteria

Four sites were chosen on two-lane two-way highways and four on two-lane one-way highways. Operational data were collected at two locations per site. One location was on a level section of highway and the second was on a moderate or steep section of highway. Truck weight and horsepower data were collected concurrently with the traffic-flow data.

The eight sites (four two-lane and four four-lane) were located in five states, as follows:

1. Colorado: two-lane and four-lane,
2. Texas: two-lane and four-lane,
3. West Virginia: two-lane,
4. Pennsylvania: two-lane and four-lane, and
5. North Carolina: four-lane.

These states included sites where most, if not all, of the combinations of vehicle type, vehicle weight/horsepower ratio, volume level, and rural highway section type are available.

Data-Collection Techniques

Data were collected by using both an automatic data-collection system and a time-lapse camera, as shown in Figure 4. Headways, speeds, and occupancies were determined from data accumulated by the automatic data-collection system. Visual-classification data were collected by using Super-8 Timelapse Corporation cameras. Truck-characteristic data were collected concurrently.

The sensors for the automatic data-collection system were pairs of inductive loops connected to a roadside terminal box, shown in Figure 5. The loops were 6 ft laterally by 12 ft longitudinally in a rectangular shape, spaced 35 ft from leading edge to leading edge; there were two conductors in each loop. The loops were fixed to the pavement surface by using 3-in-wide gray duct tape. A coating of

rubber-based adhesive was then applied to both the tape and approximately 3 in of the adjacent roadway surface to provide a stronger bond. The material used was Miracle Construction Adhesive, and it dried to a firm state in approximately 30 min under dry, warm conditions.

The roadside terminal box contained a remote oscillator for each loop, tuned to different frequencies to minimize interference for transmission to the Radian Corporation DART II computer located approximately 100 ft from the edge of the highway. The Radian DART II is a microprocessor-based data-acquisition system that contained both the vehicle detector circuits and an IBM flexible disk drive for recording data. Traffic-flow data were collected as time of actuation and time of deactuation for each of the four inductive loops in the configuration shown in Figure 5. The date, time, loop identification number, and current loop status (on or off) were recorded for each input event.

Visual vehicle classification data were collected by using Super-8 Timelapse Corporation cameras. The cameras display time of day digitally, to the second, on the film image. The cameras were generally set to 1 frame/s, although in some instances 2 frames/s was used. The time of day for all cameras and computers was set daily prior to starting the study. The time-lapse cameras were aimed at the loops to enable identification of the vehicles actuating the detectors as well as to provide a means of match with the truck-weighing operation data.

Concurrent with the automatic data-collection and visual data-collection effort, trucks were weighed at a point remote from the loop locations. The truck weighing included a side variation in methods, which ranged from the portable "loadometer"-type scales used in Pennsylvania to the directional, double-sided platform scale station in North Carolina, which processed two queues of trucks simultaneously. Truck weight, horsepower, identifying features, and time of day were manually recorded for pairing with observed traffic-flow data. Photographs of the trucks were taken in most cases to confirm site identity.

After extensive communication with the highway departments in Texas, Colorado, Pennsylvania, North Carolina, and West Virginia and following a field inspection trip to each location, the sites were chosen and studied from May through August of 1981.

DATA REDUCTION

The data were stored by the Radian Corporation DART II automatic data-collection system on flexible diskettes. The format of each record was as follows:

1. Date,
2. Time (hour, minute, second, and ticks of a 240-h clock),
3. Detector number, and
4. Detector on or off indicator.

The data on the flexible diskettes were interpreted and transferred to an 800-bit/in magnetic tape in the form of 80 character records in the same general form shown above. These records were then analyzed to determine the actuation trajectory of each vehicle. Simultaneously, the vehicles were identified as one of the 14 types previously described. The resulting vehicle data were then coded in the following format:

1. Site number,
2. Lane number,
3. Vehicle type,

4. Date,
5. Time of first trap detector on,
6. Time of first trap detector off,
7. Time of second trap detector on, and
8. Time of second trap detector off.

Volume-level samples were obtained by expanding 5-min samples to 1-h volumes; 5-min samples of the same volume level were then concatenated. Headways and speeds by vehicle type within each volume level at each unique location were analyzed. Only volume levels 1, 2, and 3 were found at the sites studied, with the exception that the Charlotte, North Carolina, site had some traffic of volume level 5. A total of 13 991 vehicles were observed at the two-lane one-way (four-lane) sites, and 11 213 vehicles were observed at the two-lane two-way sites.

DATA ANALYSIS

The data were analyzed to produce a number of statistics to be applied in the PCE procedures. Spatial-headway and speed mean values were derived by lane, site, vehicle type, and volume level. The truck-weight data obtained concurrently with the traffic-stream data were also processed to provide insight into the effects of weight and horsepower on vehicle performance.

PCE values were determined by using the Walker, spatial-headway, and equivalent-delay methods. These could be computed directly for only those volumes and vehicle types sampled. Since very little data were available at volume levels 3, 4, and 5, models were developed to allow estimation of PCEs on the basis of the Walker method or the equivalent-delay method for varying volume levels and percentages of trucks and recreational vehicles. Vehicle speed estimates were based on the gradability curves developed by MRI (7) and weight/horsepower and traffic-stream data collected during this study.

The model for estimating the Walker-method PCEs used the following equation:

$$PCE_i = (OT_i/VOL_i) / (OT_{LPC}/VOL_{LPC}) \quad (5)$$

where

- PCE_i = PCE of vehicle type i ,
 OT_i = number of overtakings of vehicle type i by passenger cars per mile per hour,
 VOL_i = volume of vehicle type i per hour,
 OT_{LPC} = number of overtakings of lower-performance passenger cars by other passenger cars per mile per hour, and
 VOL_{LPC} = volume of lower-performance passenger cars per hour.

The model for estimating the equivalent-delay PCEs used the following equation:

$$PCE_i = (OT_i/VOL_i) [(1/TSSP) - (1/MPCSP)] / (OT_{LPC}/VOL_{LPC}) [(1/AVCRSP) - (1/MPCSP)] \quad (6)$$

where

- TSSP = mean speed of mixed traffic stream,
MPCSP = mean speed of traffic stream with only higher-performance passenger cars, and
AVCRSP = mean speed of traffic stream when it contains only passenger cars.

Unimpeded vehicle speeds were estimated for typical grade conditions for each of the 14 vehicle types on two-lane and four-lane highways by using typical weight/horsepower values for each vehicle type. Insufficient data were obtained for motor-

cycles to estimate PCEs. The typical grade conditions used for PCE computation were

1. Flat (0 percent grade);
2. Three percent grade, 1 mile long; and
3. Six percent grade, 1 mile long.

The models were calibrated for the grade conditions observed; then PCE values were computed for the above typical conditions.

A fundamental assumption in the Walker method is that faster vehicles are not impeded in passing as they overtake slower vehicles, so queues do not form. Conversely, in the equivalent-delay model, it is assumed that faster vehicles are always impeded by slower vehicles, which results in queues on the analysis section. It was assumed that the Walker method is appropriate for volume level 1, and the equivalent-delay method is appropriate at volume level 5. A linear combination of the Walker and equivalent-delay PCEs was computed for each intermediate volume level.

PCE values for each vehicle type, roadway condition, and volume level were derived. For the two-lane highways, up-grade PCEs were computed. For the four-lane highways, the median and outside lanes were considered separately. Due to space constraints, the individual-vehicle-type PCE values are not presented here. They may be found elsewhere (1).

Composite PCE values for all trucks and recreational vehicles, weighted and pooled by constituent proportion of the truck population, were considered. This mechanism supports continuation of the current percentage-of-truck input to the truck factor with a single truck PCE rather than use of a large number of truck types and PCE values. These composite values were computed and are shown in Table 3 for 5,

Table 3. Composite PCE values for two-lane and four-lane highways.

Roadway Type	Trucks (%)	Volume Level				
		1	2	3	4	5
Two-Lane Highway						
Two-lane flat	5	1.5	1.6	1.6	1.6	1.7
	10	1.5	1.6	1.6	1.7	1.8
	15	1.5	1.6	1.7	1.8	1.8
	20	1.5	1.6	1.8	1.8	1.9
	25	1.5	1.7	1.8	1.9	2.0
Two-lane moderate	5	3.0	3.2	3.5	3.7	4.0
	10	3.0	3.3	3.7	4.0	4.4
	15	3.0	3.4	3.8	4.3	4.8
	20	2.9	3.5	4.0	4.5	5.1
	25	2.9	3.5	4.2	4.8	5.4
Two-lane steep	5	5.5	7.3	9.6	12.0	15.5
	10	5.4	7.6	10.5	13.3	13.3
	15	5.3	8.9	11.1	12.1	12.6
	20	5.3	8.1	10.4	11.7	12.0
	25	5.3	8.2	10.1	11.7	11.1
Four-Lane Highway						
Four-lane flat	5	1.7	1.8	1.8	1.9	2.0
	10	1.7	1.8	1.9	2.0	2.1
	15	1.6	1.8	1.9	2.0	2.1
	20	1.6	1.8	1.9	2.1	2.2
	25	1.6	1.8	1.9	2.1	2.2
Four-lane moderate	5	2.9	3.3	3.7	4.1	4.5
	10	2.9	3.4	3.9	4.4	5.0
	15	2.9	3.5	4.0	4.6	5.3
	20	2.9	3.5	4.2	4.8	5.6
	25	2.9	3.6	4.3	5.1	5.8
Four-lane steep	5	6.8	9.3	14.6	19.8	25.6
	10	6.4	8.2	9.9	12.9	17.0
	15	6.0	7.1	7.1	10.0	13.1
	20	5.6	6.3	6.7	8.0	10.1
	25	5.3	5.6	5.7	6.5	7.8

Table 4. Comparison of PCE values for two-lane highways.

Roadway Type	Volume	PCE Value		
		HCM General	HCM Specific	TTI (25 percent trucks)
Two-lane flat	A	3	2	1.5
	B	2.5	2	1.6
	C	2.5	2	1.6
	D	2	2	1.6
	E	2	2	1.7
Two-lane moderate	A	4	17	2.9
	B	5	17	3.5
	C	5	21	4.2
	D	5	21	4.8
	E	5	20	5.4
Two-lane steep	A	7	41	5.1
	B	10	41	8.2
	C	10	59	10.1
	D	12	59	11.7
	E	12	70	11.1

Table 5. Comparison of PCE values for four-lane highways.

Volume	Trucks (%)	PCE Value				
		HCM General	HCM Specific	IFR	PINY	TTI
Four-Lane Flat						
A-C	5	2	2	1.2		1.8
A-C	10	2	2	1.2		1.8
A-C	15	2	2	1.2		1.8
A-C	20	2	2	1.2		1.8
DE	5	2	2	1.8		2.0
DE	10	2	2	1.8		2.1
DE	15	2	2	1.8		2.1
DE	20	2	2	1.8		2.2
Four-Lane Moderate (3 percent, 1 mile long)						
A-C	5	4	8		8.5	3.3
A-C	10	4	6		7	3.4
A-C	15	4	5		7	3.5
A-C	20	4	6		7	3.6
DE	5	4	8		8.5	4.3
DE	10	4	6		7	4.7
DE	15	4	5		7	5.0
DE	20	4	6		7	5.2
Four-Lane Steep (6 percent, 1 mile long)						
A-C	5	8	13		21.5	10.2
A-C	10	8	12		18	8.2
A-C	15	8	12		18	6.9
A-C	20	8	11		18	6.2
DE	5	8	14		21.5	22.7
DE	10	8	13		18	15.0
DE	15	8	13		18	11.6
DE	20	8	11		18	9.0

10, 15, 20, and 25 percent trucks. These values include all buses and recreational vehicles.

A comparison among PCEs from the 1965 HCM (2), both general and specific; the IFR study (5) on urban freeways; the PINY study (8) from MRI design charts (7); and the values computed in this research at the Texas Transportation Institute (TTI) is shown in Tables 4 and 5. The PCE values found for trucks in this research generally agree with the 1965 HCM general values, as well as with the PCEs found by IFR on flat urban freeway sections. The specific values, however, found in the 1965 HCM for the two-lane moderate grade (3 percent, 1 mile long) and the two-lane steep grade (6 percent, 1 mile long) are seriously divergent. In the light of other recent research, for instance, that of McLean (10) and ours, which concludes that the PCE values for trucks

on long grades are much too high, serious consideration should be given to substantial reductions in the HCM specific-grade PCE values.

The four-lane PCE values for flat terrain are in virtual agreement for all of the values shown in Table 5. For four-lane PCE values for moderate grades, (3 percent, 1 mile long) the HCM specific values and the PINY values are both somewhat higher than the HCM general and the PCEs computed in this study. The PCE values for the four-lane steep grade are all of the same order of magnitude, although a lack of sensitivity to percentage of trucks is apparent in all sources except this research.

CONCLUSIONS

This research provided several significant findings for two-lane and four-lane rural highways, as listed below:

1. The generalized PCE values for two-lane and four-lane rural highways in the 1965 HCM are substantiated by the PCEs determined in this research for the composite stream of nonpassenger car types.
2. The PCE values for specific grades on two-lane rural highways are overly conservative for both moderate and steep grade conditions.
3. It is not necessary to increase the complexity of the truck-factor equation (Equation 2). E_L values may be used for a composite traffic stream consisting of buses, trucks, and recreational vehicles. E_L is shown in this research to be an explicit function of (a) percentage of trucks, (b) volume level (1, 2, 3, 4, 5), and (c) grade condition (flat, moderate, steep, and four-lane or two-lane).

ACKNOWLEDGMENT

The contents of this paper reflect our views, and we are responsible for the facts and accuracy of the data presented herein.

The contents do not necessarily reflect the official view or policies of the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

Discussion

Roger P. Roess

The subject of PCEs has received considerable attention in recent years for two different reasons. The federal government has supported a number of studies for the development of PCE values for various categories of trucks with the purpose of using these values in the allocation of the road-user tax burden among these categories of trucks as well as among other vehicles that use the nation's highways. The research reported by the authors results from one such study. The subject has also received attention in a variety of studies directed at producing methodologies for the third edition of the HCM.

Cunagin and Messer should be complimented on the thoroughness and quality of their work. Their paper points out, however, one of the hazards of working with PCE values: the lack of any general consensus of what PCE values are and how they ought to be calibrated.

Various researchers have used a variety of criteria for calibrating PCE values:

1. Relative numbers of passing maneuvers on two-lane highways,

2. Delay caused by trucks in the traffic stream (equivalent delay),
3. Relative spatial headways of trucks compared with those of passenger cars, and
4. Equivalent volume/capacity ratio.

Cunagin and Messer have used a combination of the first two criteria in calibrating PCE values for their study. This method is an adaptation of the methods used to develop the values of the 1965 HCM.

The authors compare the study results with PCE values calibrated at PINY for Transportation Research Circular 212 (11). In general, the study values are lower than the PINY values in Circular 212.

It should be noted that PINY values were calibrated to produce an equivalent volume in passenger cars per hour that used the same percentage of the roadway's capacity as the actual volume of mixed traffic, i.e., to keep the volume/capacity ratio constant. Thus, since the two studies started with different concepts of PCEs, it is not unusual that the results differ, even significantly.

The use of PCEs is a critical point. Messer has clearly recognized this in his recent draft report on the capacity of two-lane rural highways. In his formulation of new capacity-analysis procedures, he has not used the values resulting from the study reported in this paper, the focus of which was not capacity analysis. For capacity analysis, he suggests the use of factors based on producing an equivalent volume in passenger cars per hour that travels at the same average speed as the actual volume of mixed traffic. This is directly related to the use of average speed as a measure of effectiveness for level of service of two-lane rural highways.

For multilane highways, there is considerable evidence that the PCE values used in Circular 212 are too high, primarily because the standard truck selected for calibration was too heavy to reflect current conditions. These factors are now being revised as part of Project 3-28B of the National Cooperative Highway Research Program, which will produce the third edition of the HCM.

In summary, the paper presents a most comprehensive treatment of PCE calibration and use. The philosophical issue of whether or not PCE values are

an appropriate measure by which to evaluate the allocation of road-user taxation remains but is not one that can be addressed in the context of a study such as that reported here, since it was a given objective of the sponsor.

REFERENCES

1. W.D. Cunagin. Passenger Car Equivalents for Rural Highways. FHWA, Final Rept., July 1982.
2. Highway Capacity Manual. HRB, Special Rept. 87, 1965.
3. O.K. Normann. Results of Highway Capacity Studies. Public Roads, Vol. 23, No. 4, 1939.
4. J.G. Wardrop. Some Theoretical Aspects of Road Traffic Research. Proceedings of the Institute of Civil Engineering, Part 2, Vol. 1, 1952, p. 325.
5. E.D. Seguin and others. Urban Truck Freeway Characteristics. FHWA, draft final rept., 1981.
6. J. Craus and others. A Revised Method for the Determination of Passenger Car Equivalencies. Transportation Research, Vol. 14A, 1980, pp. 241-246.
7. A.D. St. John. Freeway Design and Control Strategies as Affected by Trucks and Traffic Regulations. Midwest Research Institute, Kansas City, MO; FHWA, Rept. FHWA-RD-74-42, April 1975.
8. E.M. Linzer, R.P. Roess, and W.R. McShane. Effect of Trucks, Buses, and Recreational Vehicles on Freeway Capacity and Service Volume. TRB, Transportation Research Record 699, 1979, pp. 17-26.
9. C.J. Messer. Development of Truck Equivalencies on Rural Highways. Texas Transportation Institute, College Station, 1980.
10. J.R. McLean. Two-Lane Traffic Flow and Capacity. Australian Road Research, Australian Road Research Board, Nunawading, Victoria, Australia, 1980.
11. Interim Materials on Highway Capacity. TRB, Transportation Research Circular 212, Jan. 1980.

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

Abridgment

Traffic Data-Collection Systems: Current Problems and Future Promise

RICHARD W. LYLES AND JOHN H. WYMAN

Evaluations were undertaken of off-the-shelf automatic traffic data-collection systems to examine their capabilities in two specific areas: data collection for speed compliance and vehicle classification by type. A prototype system developed by the United Kingdom's Transport and Road Research Laboratory (TRRL) was also briefly tested. System performances in the speed mode were not outstanding although several systems showed promise, whereas in the vehicle-classification mode, performance was not good and most systems experienced serious problems. Significant problems were also encountered in the use of pneumatic tubes as sensors. In the classification mode most systems also suffered from inadequate classification schemes (i.e., number and definition of categories). However, the TRRL system performed quite well; it uses a more sophisticated classification scheme (proposed as an alternative to current schemes) and incorporates inputs from both presence and axle sensors.

Transportation engineers deal continually with large amounts of traffic data generated in a variety of ways and for numerous purposes. Because many states are experiencing financial problems, data-collection activities are being reviewed from several perspectives: whether the data are really needed, what accuracy is required, and what the best ways are to collect them. Institution of the national maximum speed limit (NMSL) and the attention brought by the Federal Highway Administration (FHWA) to developing the Highway Performance Monitoring System (HPMS) have also focused attention on data-collection techniques.

Although several states have investigated and/or invested in a different automatic data-collection system, most data collection in the areas noted has been done manually. Not only is this a labor-intensive operation, but there are other problems that may be introduced, e.g., seasonal bias of data and unreliability of observers. In spite of the apparent need for labor- and cost-saving systems, until recently no organization had undertaken a comprehensive review of currently available systems or their performance. In light of this, FHWA contracted for an extensive evaluation of such systems with regard to how well they could satisfy data needs in two specific areas: collecting speed-compliance data (with regard to the NMSL) and classifying vehicles by type (e.g., on the basis of axle spacing and overall length).

Manufacturers' names will not be used with regard to which system was best or worst; rather, the discussion will be limited to comments about system capabilities in general. [The exception to this will be in the discussion of sensing devices (indirectly) and experimental systems developed in the public domain.] However, findings with regard to specific systems may be found in final reports to FHWA (1,2). Systems that were evaluated were produced by the following manufacturers: Leupold and Stevens, Inc.; Safetran Traffic Systems, Inc.; Streeter-Amet; Redland Automation, Inc. (Sarasota Division); and Golden River Corporation. A prototype unit from the United Kingdom's Transport and Road Research Laboratory (TRRL) was also tested as part of the program but will not be discussed directly.

GENERAL DESCRIPTION

The discussion that follows is directed to available off-the-shelf systems that are marketed for general

purposes and does not necessarily apply to special-purpose systems used in a research context.

In general, systems used for speed monitoring or for vehicle classification have four basic components: sensing devices (sensors), which provide the essential indication, or signal, of a vehicle's presence and movement; detectors, which receive the signals from the sensors and amplify and/or interpret them; and the recorder and processor, which receive the signals from the detectors and calculate the speed or assign a category, record that information, and perform whatever manipulations of the basic data are necessary to present them in final form (e.g., individual vehicle speeds and frequency counts). Although these components essentially serve the functions indicated, most systems do not actually have separable components beyond the sensing device. In addition, all systems are currently limited to using only one type of sensing device in a given installation.

Most systems are capable of producing either speed or classification (by type) data but not both concurrently. However, it should be noted that when a system is operating in the classification mode, a speed calculation is made internally as a prerequisite to classification.

Sensors that are commonly used include inductance loops, pneumatic tubes, coaxial cables, and tape-switches, although the last is typically used only in research situations. Inductive loops are typically imbedded in the pavement and are thus permanent and hence most costly, whereas the tubes and cables are used on the road surface and have shorter lives. Although commonly used, easy to install, and fairly inexpensive, the tubes are most prone to damage: The inherent high visibility of the installation led to one of the most prevalent problems in recent testing in Maine--purposeful damage by truckers (i.e., locking their trailer brakes and skidding over the installation). Hence, life expectancy is unpredictable at best. It should also be noted that air leaks are often hard to find and intermittent operation of systems is possible. The latter is a serious problem since total volumes or populations of classification categories may be incorrectly estimated from data that appear to be good.

The visibility of cables is apparently considerably less than that of tubes (comparative tests showed that the tubes were always seen and damaged by truckers, whereas no attempts were made on cable sensor installations). Easy to install, the cables are longer lasting than tubes and will resist purposeful damage but will eventually succumb to snowplows, studded tires, and sharp dragging objects that breach their protective rubber coating. Variations of the cables (T-shaped cross sections imbedded in pavement) have been developed and used by TRRL with considerable long-term success.

There are several other sensing devices that have at least some potential for collection of vehicle-classification and/or speed data, including the self-powered vehicle detector (3,4), the magnetic-gradient vehicle detector (5,6), optical sensing (7), audio signals (8), and electronic timers (9). Radar can also be used for speed data collection al-

though it is generally conceded that for large-scale collection programs it is cost-inefficient; it has limitations in high-volume situations in being able to record the speeds of specific vehicles in queues and in introducing bias toward lower speeds given the proliferation of citizen-band radios and radar detectors. Although potential use of other such sensors is not barred, their use in the near future in automatic systems seems unlikely.

PROBLEMS WITH EXISTING SYSTEMS

The existing data-collection systems were subjected to a series of tests to evaluate their performance in field situations, i.e., how well they performed in general and whether they did what they were purported to do. The systems were tested on separate occasions in the speed mode and in the vehicle-classification mode.

Results in Speed Mode

Results in the speed mode were as follows:

1. Primary problems were missed vehicles and inaccurate speeds.
2. Inaccurate speeds adversely affected speed-compliance percentages; there was a general high skew that was especially noticeable in the tails of the speed distribution (e.g., percentage over 55, 65).
3. There was variation in accuracy from unit to unit (same manufacturer) or from site to site (same unit).
4. When test system was compared with base system(s) such as radar, fifth wheel, and/or optical timer, average speeds of samples of vehicles were typically within 1 mph.

Results in Vehicle-Classification Mode

The following results were found for the vehicle-classification mode:

1. Classification schemes were based only on either overall length (loop based) or number of axles and spacing (axle sensor based, e.g., pneumatic tubes).
2. Sensitivity of loop detectors caused significant errors.
3. Minimum length error: 5-8 percent of vehicles were not measured within ± 5 ft.
4. Maximum length error: 82 percent of vehicles were not measured within ± 5 ft (loop detector out of tune).
5. Minimum percentage of axle-based classifications: 15-20 percent of vehicles were misclassified.
6. Maximum percentage of axle-based classifications: two-thirds of vehicles were misclassified.

Sensor Problems

The sensor problems included the following:

1. Axle-sensor-based units were limited to the number of axles and axle spacings.
2. Loop-based units were limited to overall length only.
3. Nature of axle sensors was nonpermanent.
4. Undetected intermittent failure of axle sensors (especially tubes) is possible.
5. Tubes are highly visible and often purposefully damaged.

Problems with Other System Components

The following are problems with other system components:

1. System-loop detector adjustments can be critical (and are variable by site).
2. There are numerous minor breakdowns.
3. Missed or misclassified vehicles are not obviously due (in some instances) to sensor failures.
4. Classification schemes are typically simplistic.

Positive Aspects of Systems

The existing systems are not without positive aspects. For example, it is clear that the technology exists to process information from either axle or presence sensors. In addition, some fairly sophisticated differentiations among vehicles were incorporated into the different systems' classification schemes.

CURRENT AND PROPOSED VEHICLE-CLASSIFICATION SCHEMES

As part of the larger question of what data are really needed, there is some debate over what constitutes an adequate vehicle-classification (by type) scheme (10). For currently available systems, the number of available categories varies from four, based entirely on overall length (raw data from loops), to eight, based on the number of axles and axle spacings (raw data from axle sensors). Users of vehicle-classification data, however, appear to desire more detail than is currently being provided. For example, FHWA has examined schemes with from 7 to 32 categories and currently suggests 13 in the Highway Performance Monitoring System (HPMS) program (11). At this point, then, there appears to be a significant gap between what available systems can deliver and what users desire.

The currently proposed schemes themselves also have some problems since they do not necessarily reflect logical differentiations among vehicle types. Some of the problems with currently discussed schemes are summarized below (based on FHWA-supplied definitions and 1977 FHWA truck-weight study data):

1. There is substantial overlap in classifications of automobiles; with the trend in downsizing, current cutoffs (e.g., subcompact versus standard or compact) are inappropriate--for wheelbase, < 100 in; for overall length, < 180 in.
2. There is overlap between vans, light pickups, and standard sedans or station wagons.
3. There is overlap between light trucks and pickup trucks.
4. No system is accurate on motorcycle differentiation.
5. Bus categories overlap with some trucks; buses are seriously overestimated.

In an effort to provide a somewhat more reasonable departure point for further discussion of vehicle classification, the scheme in Table 1 is presented. It is characterized by the following: (a) it is based on information that can be obtained from axle sensors alone, (b) it recognizes some obvious problems with some differentiations and eliminates them, and (c) it considers the frequency of encountering certain types of vehicles (i.e., how important a specific category is). Three comments are pertinent. First, the scheme is based on data from FHWA and assumes that the shortcomings of those data do not seriously compromise the characteristics of different types of vehicles (e.g., some vehicle

Table 1. Vehicle-classification proposal (14 categories).

Vehicle Category	Description	Proposed Rule
E-1	Passenger cars, light trucks, vans	Axles = 2; wheelbase < 10 ft
E-2	Heavy-duty pickups, delivery trucks, 2A6T's	Axles = 2; wheelbase > 10 ft
E-3	Cars and light trucks with one- or two-axle trailers	Axles = 3 or 4; 1,2 spacing < 10 ft; 5,5 ft < 2,3 spacing < 22 ft
E-4	Three-axle single-unit trucks	Axles = 3 and not E-3
E-5	Trucks and semitrailers (2S2)	Axles = 4 and not E-3; 3 ft < 3,4 spacing < 10 ft
E-6	Four-axle single-unit trucks	Axles = 4 and not E-3; 3 ft < 2,3 spacing < 5 ft
E-7	Other four-axle combinations	Axles = 4 and not E-3, E-5, and E-6
E-8	Trucks and semitrailers (3S2)	Axles = 5; 2 ft < 4,5 spacing < 10 ft
E-9	Other five-axle combinations	Axles = 5 and not E-8; 3 ft < 2,3 spacing < 5 ft
E-10	Trucks and semitrailers plus full trailers (2S1-2)	Axles = 5 and not E-8 or E-9
E-11	Trucks and semitrailers plus full trailers (3S1-2)	Axles = 6 and 5,6 spacing > 7 ft
E-12	Trucks and semitrailers (3S3)	Axles = 6 and not E-11; 4,5 spacing < 6 ft
E-13	Other six-axle combinations	Axles = 6 and not E-11 or E-12
E-14	Other seven-or-more-axle combinations	Axles = 7 or more

Note: An optional category would be for 2S1 truck and semitrailer combination.

types are underrepresented although it is assumed that recorded wheelbases, etc., would not differ from a scientifically drawn sample of the type). Second, the scheme does not attempt certain differentiations that could be made if overall length data were also considered (e.g., buses could probably be better differentiated). Third and last, the proposed scheme is presented only as a point of departure for discussion by data users and others and not as the ultimate scheme.

CONCLUDING REMARKS AND RESEARCH SUGGESTIONS

Speed calculation and vehicle classification by length, wheelbase, axle spacings, and number of axles are conceptually straightforward; e.g., the implicit decision rules for given categories are easily stated in terms of overall length and/or the information from axle sensors. While the currently marketed systems tested typically used only a limited number of categories, the more complex classification schemes do not necessarily imply the development of new technology but rather a refinement of that which exists. There are, however, several areas that would benefit from additional work and/or attention. These are outlined below:

1. A permanent, or at least longer-lived, axle sensor should be developed. At this point, it appears that some derivation of the coaxial cable would be the most likely candidate.
2. Prototype systems should be developed to examine the problems, if any, in interfacing coaxial cables with existing systems.
3. A single classification scheme should be developed for use by the states (for their own purposes) and FHWA. Such a scheme should recognize, for example, trends in vehicle sizes, especially in the passenger-car categories.
4. Systems need to be developed that can process inputs from a variety of sensing devices (e.g., from only coaxial cables or from a cable-and-loop combination).
5. System internal logic needs to be versatile

enough to accommodate minor changes in the number and parameters of categories.

6. All systems should have at least the option of printing out hourly data (inherently useful) so that the time of potential breakdowns can be estimated.

7. All systems should have straightforward and fully documented diagnostics, calibrations, and adjustments that can be made by the user.

8. Systems that can simultaneously classify vehicles by type and collect speed data should be investigated.

The Automatic Vehicle Classification System (AVCS) developed by TRRL and tested in Maine (12-16) meets many of the above requirements, and the test results clearly demonstrated that reasonably sophisticated traffic data-collection systems can be developed and used successfully in the field and that the resulting data will have applicability in several areas of concern to engineers, planners, and policymakers.

There are, however, several questions that remain to be addressed by FHWA, the states, and other potential users of such systems and/or the resultant data:

1. Is the demand for such sophisticated data so widespread as to warrant the development of systems capable of delivering them?
2. Can sufficient consensus be achieved among the potential users on the form of required data so that basic parameters required for classification and minimum or maximum system capabilities can be defined?
3. Will the states and other local and nongovernment users pursue purchase of new systems if such data are not required and system acquisition is not supported by the federal government?

The last point to be made concerns the capabilities of systems that use more or less permanent sites (i.e., AVCS or AVCS-type system) versus those that might be truly portable (i.e., that use temporary road-surface sensors). Current sensor technology, and even that only available in prototype form, basically constrains sophisticated equipment to using permanent sensor arrays that require a large initial commitment of time and resources to implement any comprehensive data-collection program; more primitive systems can deliver lower-quality data much more cheaply. In this regard, it is not clear whether a truly portable system (including temporary sensors) can provide the same quality of data as an AVCS-type system in a permanent installation. It is apparent that large quantities of good data can be collected; are they worth the cost of acquisition?

ACKNOWLEDGMENT

Considerably more detail on the work reported here may be found in the final report submitted to FHWA. We are indebted to Robin Moore and Brian Stoneman of TRRL for their assistance in the setting up and testing of the AVCS in Maine. The work was funded by FHWA and undertaken at the Maine Facility Laboratory by personnel from the Maine Department of Transportation and the Social Science Research Institute of the University of Maine and by us. The contributions of the facility staff and others notwithstanding, the responsibility for the material presented here and in the final reports rests with us and does not necessarily represent the opinions of FHWA, Maine Department of Transportation, or any other institution.

REFERENCES

1. R.W. Lyles and J.H. Wyman. Evaluation of Speed Monitoring Systems. FHWA, Final Rept. FHWA/PL/80/006, July 1980.
2. R.W. Lyles and J.H. Wyman. Evaluation of Vehicle Classification Equipment. FHWA, Final Rept., July 1982.
3. J.F. Scarzello and G.W. Usher, Jr. Self-Powered Vehicle Detector: Magnetic Sensor Development. FHWA, Rept. FHWA-RD-76-147, June 1976.
4. J.H. Wyman and R.W. Lyles. Evaluation of the Self-Powered Vehicle Detector. FHWA, June 1981.
5. M.K. Mills. Magnetic Gradient Vehicle Detector. IEEE Transactions on Vehicular Technology, Vol. VT-23, No. 3, Aug. 1974.
6. M. Singleton and J.E. Ward. A Comparative Study of Various Types of Vehicle Detectors. U.S. Department of Transportation, Rept. DT-TSC-OST-77-9, Sept. 1977.
7. T. Takagi. Optical Sensing and Size Discrimination of Moving Vehicles Using Photocell Array and Threshold Devices. IEEE Transactions on Instrumentation and Measurement, Vol. 25, No. 1, March 1976.
8. K. Jakus and D.S. Coe. Speed Measurement Through Analysis of the Doppler Effect in Vehicular Noise. IEEE Transactions on Vehicular Technology, Vol. VT-24, No. 3, Aug. 1975.
9. M.R. Parker, Jr. An Evaluation of the Following Too Closely System. Traffic Engineering, Vol. 46, No. 7, July 1976.
10. John Hamburg and Associates. Improved Methods for Vehicle Counting and Determining Vehicle Miles of Travel--A Review of the Current State-of-the-Art on Vehicle Classification Programs. NCHRP, draft report, Oct. 1979.
11. Highway Performance Monitoring System: Case Study Procedural Manual--Vehicle Classification. Program Management Division, FHWA, Jan. 1980.
12. R.C. Moore. Road Sensors for Traffic Data Collection. Sensors in Highway and Civil Engineering. Institute of Civil Engineers, London, England, Paper 7, 1981.
13. R.C. Moore, P. Davies, and P.R. Salter. An Automatic Method to Count and Classify Road Vehicles. Presented at International Conference of Road Traffic Signalling, Institute of Electrical Engineers, London, England, March-April 1982.
14. Automatic Method to Classify, Count, and Record the Axle Weight of Road Vehicles. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, Leaflet 639, 1979.
15. Road Sensor for Vehicle Data Collection: The Axle Detector. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, Leaflet 905, 1979.
16. Automatic Vehicle Count and Classification: Examples of Data Output. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, Leaflet 884, 1979.

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

Analysis of TRANSYT Platoon-Dispersion Algorithm

NAGUI M. ROUPHAIL

The development of an analytical solution to the recursive platoon-dispersion formula used in TRANSYT models of traffic flow is presented. Flow rates in the predicted platoon measured at the k th interval of the j th simulated cycle are expressed in terms of demand and capacity rates at the source intersection in addition to signal-control and travel-time parameters. It was found that the TRANSYT recursive formula implicitly contains a cycle factor that results in an underestimation of the total flow rate simulated. An estimate of that error has been formulated, which can be applied as a constraint on the required simulation time in TRANSYT. The analytical solution also provided insight into the determination of critical intersection spacings below which signal coordination becomes feasible.

The proliferation of digital computer model applications in the areas of traffic flow and control in the past decade has led to the successful development of several widely used traffic signal operations models, such as Network Simulation Model (NETSIM), Signal Operations Analysis Package (SOAP), Traffic Network Study Tool (TRANSYT), and Traffic Signal Optimization Program (SIGOP) (1-4).

TRANSYT, a program for traffic signal timing and coordination initially developed in the United Kingdom by Robertson (5), has been successfully applied at many intersections in Europe and the United States. The TRANSYT-7F version, for example, has recently been used in the National Signal Timing Optimization Project (6), which encompassed 11 cities and approximately 500 signalized intersections in the United States.

The fundamental principle of traffic representa-

tion in TRANSYT-type models is platoon-dispersion behavior. Simply stated, as a queue of vehicles leaves the stopline on the green indication, its shape is altered along the downstream link in a manner reflective of the desire of individual drivers to maintain comfortable time headways. Thus, although the flow rate at the stopline is equivalent to the saturation rate in the presence of a queue and to the demand rate thereafter (assuming undersaturated operation), the flow patterns measured at an observation point t seconds downstream of the stopline would be considerably different.

Mathematically, platoon-dispersion behavior is expressed by the following recursive relationship:

$$IN(k+t) = F \times OUT(k) + (1-F) \times IN(k+t-1) \quad (1)$$

where

$IN(k+t)$ = flow rate in k th time interval of predicted platoon at observation point;

$OUT(k)$ = flow rate in k th time interval of original platoon at stopline;

t = β times average platoon travel time from stopline to observation point [β is an empirical travel-time factor expressed as ratio between travel time of leading vehicle in platoon

and average travel time of entire platoon; a value of 0.8 has been suggested for use in TRANSYT models (7); and

F = empirically derived platoon-smoothing factor.

F is a function of travel time and geometric conditions on the link. To date, however, it has been expressed in terms of travel time only, as shown below:

$$F = 1/(1 + \alpha t) \quad (2)$$

where α is the platoon-dispersion factor. Estimates of 0.50 and 0.35 were found to give the best fit to observed traffic under moderate travel friction in the United Kingdom and the United States, respectively. However, a recent U.S. study, described in a paper in this Record by McCoy and others, reported a correlation between the value of α and arterial geometries under low-friction traffic flow conditions. Values of $\alpha = 0.21$ and $\alpha = 0.14$ were suggested in the modeling of two-lane two-way and four-lane divided arterials, compared with the 0.25 value recommended in TRANSYT under the same conditions.

Thus the predicted flow rate at any time step is expressed as a linear combination of the original platoon flow rate in the corresponding time step (with a lag of t) and the flow rate of the predicted platoon in the step immediately preceding it.

The platoon-dispersion model formulated in Equation 1 has been successfully validated with field data collected in London and Manchester, England (8).

The objectives of this study thus may be stated as follows:

1. Develop a close-form solution to the platoon-dispersion algorithm in TRANSYT-type models,
2. Investigate the time-dependency impacts of the algorithm on the predicted platoon flow rates, and
3. Explore potential uses of the analytical expressions developed in the study for signal-coordination schemes.

The first step of the analysis was concerned with the determination of the platoon flow rates at the stopline of a signalized, isolated intersection, as discussed below.

FLOW RATES AT ISOLATED INTERSECTIONS

Consider the flow patterns occurring at an isolated signalized intersection (or peripheral intersection in a TRANSYT network), assuming undersaturated operation.

The following variables are defined:

- c = cycle length (s),
- g = effective green time (s),
- r = effective red time (s),
- $\lambda = g/c$,
- g_s = saturated green time (s),
- g_u = unsaturated green time (s),
- n = number of time steps in a cycle as defined in TRANSYT,
- q = average demand rate (vehicles/s),
- s = saturation flow rate (vehicles/s),
- $y = q/s$,
- x = degree of saturation ($= y/\lambda$),
- $IN(k, j)$ = arrival rate in step k of cycle j , and
- $OUT(k, j)$ = departure rate in step k of cycle j .

Based on the flow profiles indicated in Figure 1, it can be shown that

$$g_s = q(r + g_s) \text{ or } g_s = r q / (s - q) = r y / (1 - y)$$

$$\text{Since } r = c - g = c(1 - \lambda),$$

$$g_s = c y (1 - \lambda) / (1 - y) \quad (3)$$

It is assumed that the red interval is made up of k_1 time steps. Therefore,

$$k_1 x (c/n) = r = c(1 - \lambda)$$

or

$$k_1 = n(1 - \lambda) \quad (4)$$

Similarly, k_2 is defined as the last time step in the saturated portion of the green phase, measured from the beginning of the effective-red interval. Thus

$$k_2 x (c/n) = r + g_s$$

or with some manipulation,

$$k_2 = n(1 - \lambda) / (1 - y) \quad (5)$$

Finally, let k_3 be the last time step in the cycle. By definition,

$$k_3 = n \quad (6)$$

Thus the following flow rates are used to describe the original platoon flow in step k of cycle j at the "source" intersection:

$$IN(k, j) = q \quad 1 \leq k \leq n, \text{ for all } j \quad (7)$$

$$OUT(k, j) = 0 \quad 1 \leq k \leq n(1 - \lambda), \text{ for all } j \quad (8)$$

$$OUT(k, j) = s \quad n(1 - \lambda) < k \leq n(1 - \lambda) / (1 - y), \text{ for all } j \quad (9)$$

$$OUT(k, j) = q \quad n(1 - \lambda) / (1 - y) < k \leq n, \text{ for all } j \quad (10)$$

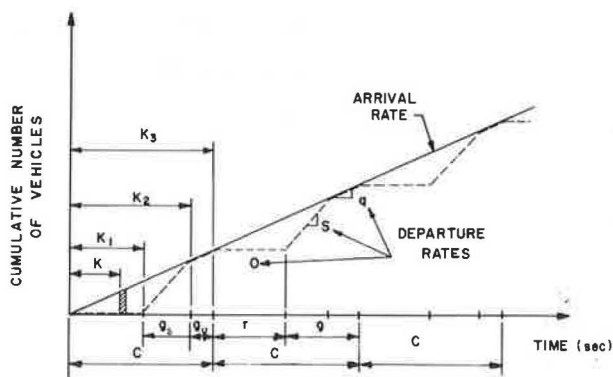
Note that under fully saturated conditions (i.e., $\lambda = y$), the outflow rate described in Equation 10 is eliminated.

ANALYTICAL SOLUTION DEVELOPMENT

Case for Initial Cycle

The inflow rate in step k of the first simulated cycle measured at an observation point t seconds

Figure 1. Flow patterns at isolated intersection.



downstream of the stopline is expressed mathematically by Equation 1, rewritten below:

$$IN(k+t, 1) = F \times OUT(k, 1) + (1-F) \times IN(k+t-1, 1) \quad (1a)$$

From Equation 8, it is evident that no departures occur in the red interval; i.e.,

$$IN(k+t, 1) = 0 \quad 1 \leq k \leq n(1-\lambda) \quad (11)$$

The flow rate corresponding to the saturated portion of the effective-green phase is derived from Equation 9 as

$$IN(k+t, 1) = F \times s + (1-F) \times IN(k+t-1, 1)$$

Let k now be measured from the beginning of the effective-green phase and define the variable u_0 as

$$u_0 = IN[n(1-\lambda) + t, 1]$$

where the subscript zero refers to the last step of the effective-red interval.

Substituting into Equation 1 gives

$$u_{0+1} = F \times s + (1-F)u_0$$

$$u_{1+1} = F \times s + (1-F)u_1$$

and, in general,

$$u_k = F \times s + (1-F)u_{k-1}$$

Therefore,

$$\begin{aligned} u_k &= (1-F)^k u_0 + \sum_{\ell=0}^{k-1} F s (1-F)^\ell \\ &= (1-F)^k u_0 + s[1 - (1-F)^k] \end{aligned}$$

When k is measured from the start of the red interval,

$$u_k = (1-F)^{k-n(1-\lambda)} u_0 + s[1 - (1-F)^{k-n(1-\lambda)}]$$

But from Equation 11, $u_0 = 0$; therefore,

$$IN(k+t, 1) = s[1 - (1-F)^{k-n(1-\lambda)}] \quad n(1-\lambda) < k \leq n(1-\lambda)/(1-y) \quad (12)$$

The flow rate corresponding to the unsaturated portion of the green phase is derived from Equation 10 as

$$IN(k+t, 1) = F \times q + (1-F) \times IN(k+t-1, 1)$$

Let k be measured from the beginning of the unsaturated portion of the green phase and define \dot{u}_0 as

$$\dot{u}_0 = IN\{[n(1-\lambda)/(1-y)] + t, 1\}$$

where the subscript zero refers to the last step of the saturated green time.

From Equation 12, it can be shown that

$$\begin{aligned} \dot{u}_0 &= s\{1 - (1-F)[n(1-\lambda)/(1-y)] - n(1-\lambda)\} \\ &= s\{1 - (1-F)[ny(1-\lambda)/(1-y)]\} \end{aligned} \quad (13)$$

Proceeding in a similar fashion as that above for \dot{u}_1 , \dot{u}_2 , and so on gives

$$\begin{aligned} \dot{u}_k &= (1-F)^k \dot{u}_0 + \sum_{\ell=0}^{k-1} F q (1-F)^\ell \\ &= (1-F)^k \dot{u}_0 + q[1 - (1-F)^k] \end{aligned}$$

Substituting \dot{u}_0 from Equation 13 and letting k be measured from the start of the red interval,

$$\begin{aligned} IN(k+t, 1) &= s[1 - (1-F)^{ny(1-\lambda)/(1-y)}] (1-F)^{k-[n(1-\lambda)/(1-y)]} \\ &\quad + q\{1 - (1-F)^{k-[n(1-\lambda)/(1-y)]}\} [n(1-\lambda)/(1-y)] \\ &\quad < k \leq n \end{aligned} \quad (14)$$

A summary of the flow rates derived in this section is presented in Table 1. Note that the flow rates are valid only in the first simulated cycle in TRANSYT, as the subsequent analysis explains.

Case for Subsequent Cycles

If we refer to Table 1, it can be proved that

$$\sum_{k=1}^n IN(k+t, 1) < \sum_{k=1}^n OUT(k, 1) \text{ for } 0 \leq t \leq \infty$$

This is primarily due to the recursive nature of the platoon-dispersion formula itself, which continuously incorporates a fraction of all previous flow rates in the calculation of the predicted platoon. Thus the difference

$$\sum_{k=1}^n OUT(k, 1) - \sum_{k=1}^n IN(k+t, 1)$$

may be viewed as a residual flow from cycle 1 that will be dispersed in cycles 2, 3, ..., j according to the dispersion formula in Equation 1. The same reasoning may be applied for platoons generated in the second and subsequent cycles throughout a simulation run.

Thus for a particular cycle j , the flow rate in the predicted platoon at time $(k+t)$ may be viewed as the sum of two components:

1. Flow rate due to the original platoon generated in step k of cycle j at the source intersection and

2. Residual flow rate from all previous platoons generated in cycle 1, 2, ..., $j-1$ at the source intersection.

The first component is identical to the flow rate generated in the first simulated cycle, i.e., with no consideration of previous platoons. This concept is illustrated graphically in Figure 2, which depicts the progression of a platoon downstream of an isolated intersection and the associated residual flows generated in each cycle.

The residual flow rate from the i th cycle that occurs in step k of the j th cycle ($j > i$), $R_{i,j,k}$, can be expressed by the following equation:

$$R_{i,j,k} = IN(n+t, 1) \times (1-F)^{k+n(i-1-1)} \quad (15)$$

For example, letting $i = 1$, $j = 2$ gives

$$R_{1,2,k} = IN(n+t, 1) \times (1-F)^k$$

Table 1. Platoon-flow boundary values.

Case	Travel Time	$k_1 \leq k \leq k_2$	Q_{∞}^a
1	0	$1 \leq k \leq n(1-\lambda)$	Zero
2	0	$n(1-\lambda) < k \leq [n(1-\lambda)/(1-y)]$	$qc(1-\lambda)/(1-y)$
3	0	$[n(1-\lambda)/(1-y)] < k \leq n$	$qc(\lambda-y)/(1-y)$
4	∞	Same as 1	$qc(1-\lambda)$
5	∞	Same as 2	$qcy(1-\lambda)/(1-y)$
6	∞	Same as 3	$qc(\lambda-y)/(1-y)$

^a $Q_{\infty} = (c/n) \sum_{k=k_1}^{k_2} IN(k+t, \infty)$.

Figure 2. Platoon-dispersion formula characteristics.

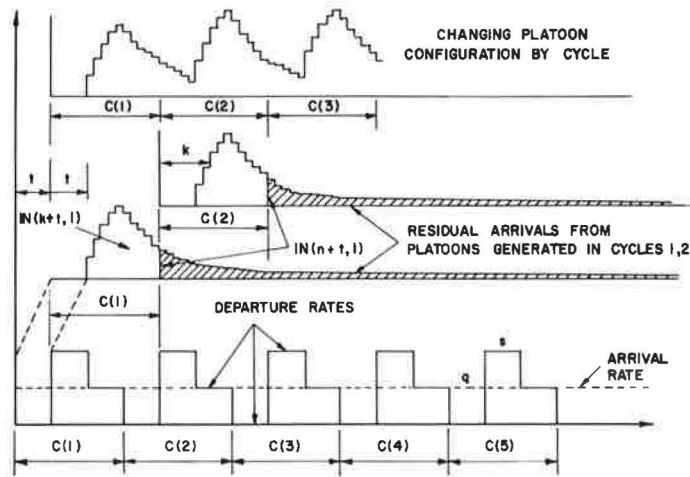


Figure 3. Summary of flow rates for initial cycle.

Region ^a	Boundaries for (k)	OUT(k, 1)	IN(k+t, 1)
1	$1 \leq k \leq n(1-\lambda)$	0	0
2	$n(1-\lambda) < k \leq \frac{n(1-\lambda)}{(1-\gamma)}$	s	$s \left[1 - (1-F)^k - [n(1-\lambda)] \right]$
			$s \left[1 - (1-F)^{ny(1-\lambda)/(1-\gamma)} \right] \times$
3	$\frac{n(1-\lambda)}{(1-\gamma)} < k \leq n$	q	$(1-F)^k - [n(1-\lambda)/(1-\gamma)]$
			$+ q \left[1 - (1-F)^k - [n(1-\lambda)/(1-\gamma)] \right]$

^a1 = Red phase at source intersection

2 = Saturated portion of green phase

3 = Unsaturated portion of green phase

which is consistent with Equation 1 since the OUT(k, 1) term is eliminated when cycle 2 is considered.

Therefore, the total flow rate in the predicted platoon, including residuals from all previous cycles, is expressed as

$$IN(k+t, j) = \left(\sum_{i=1}^{j-1} R_{i,j,k} \right) + IN(k+t, 1) \quad (16)$$

which suggests that the flow rates (and consequently the actual number of vehicles) simulated in TRANSYT appear to be systematically related to the length of a simulation run.

Solving Equation 16 for $j = 2, 3, 4$, respectively, gives

$$IN(k+t, 2) = IN(n+t, 1)(1-F)^k + IN(k+t, 1) \quad (17)$$

$$IN(k+t, 3) = \left(\sum_{i=1}^2 R_{i,3,k} \right) + IN(k+t, 1)$$

But from Equation 15,

$$R_{1,3,k} = IN(n+t, 1) \times (1-F)^{k+n}$$

$$R_{2,3,k} = IN(n+t, 1) \times (1-F)^k$$

Therefore, it can be shown that

$$IN(k+t, 3) = IN(n+t, 1)(1-F)^k [1 + (1-F)^n] + IN(k+t, 1) \quad (18)$$

A similar derivation for $j = 4$ gives

$$IN(k+t, 4) = IN(n+t, 1)(1-F)^k [1 + (1-F)^n + (1-F)^{2n}] + IN(k+t, 1) \quad (19)$$

which leads to a general expression for cycle j , $j > 1$:

$$IN(k+t, j) = \left[IN(n+t, 1)(1-F)^k \sum_{m=0}^{j-2} (1-F)^{mn} \right] + IN(k+t, 1)$$

$$= IN(n+t, 1)(1-F)^k \left\{ \frac{[1 - (1-F)^{n(j-1)}]}{[1 - (1-F)^n]} \right\} + IN(k+t, 1) \quad (20)$$

Equation 20 is a general platoon-dispersion formula that predicts the flow rate in the k th step ($1 \leq k \leq n$) of the j th simulated cycle in TRANSYT. Because the term containing the cycle designation j vanishes at $j = 1$, the expression is considered valid for any cycle, including the initial one.

By substituting the values of $IN(n+t, 1)$, $IN(k+t, 1)$ from Figure 3, the set of general expressions for flow rates is realized. A summary of these is presented in Figure 4.

CYCLE FACTOR AND ERROR ESTIMATION

The dependency of platoon flow rate on cycle designation implies that each time an IN histogram is

Figure 4. General expressions for flow rates.

Cycle (j)	Region ^a	$k_1 \leq k \leq k_2$	OUT(k, j)	IN(k+t, j)
	1*	$1 \leq k \leq n(1-\lambda)$	0	0
1	2*	$n(1-\lambda) < k \leq \frac{n(1-\lambda)}{(1-y)}$	s	$s[1 - (1-F)^{k-n(1-\lambda)}]$
	3*	$\frac{n(1-\lambda)}{(1-y)} < k \leq n$	q	$s[1 - (1-F)^{ny(1-\lambda)/(1-y)}](1-F)^{k-[n(1-\lambda)/(1-y)]}$ + $q[1 - (1-F)^{k-[n(1-\lambda)/(1-y)]}]$
	1	$1 \leq k \leq n(1-\lambda)$	0	$(1-F)^k \left[\frac{1 - (1-F)^{n(j-1)}}{1 - (1-F)^n} \right] \{ q[1 - (1-F)^{n(\lambda-y)/(1-y)}]$ + $s(1-F)^{n(\lambda-y)/(1-y)} \times [1 - (1-F)^{ny(1-\lambda)/(1-y)}] \}$
> 1	2	$n(1-\lambda) < k \leq \frac{n(1-\lambda)}{(1-y)}$	s	$s[1 - (1-F)^{k-n(1-\lambda)}] + (1-F)^k \left[\frac{1 - (1-F)^{n(j-1)}}{1 - (1-F)^n} \right]$ $\times \{ q[1 - (1-F)^{n(\lambda-y)/(1-y)}] + s(1-F)^{n(\lambda-y)/(1-y)}$ $\times [1 - (1-F)^{ny(1-\lambda)/(1-y)}] \}$
	3	$\frac{n(1-\lambda)}{(1-y)} < k \leq n$	q	$s[1 - (1-F)^{ny(1-\lambda)/(1-y)}](1-F)^{k-[n(1-\lambda)/(1-y)]}$ + $q[1 - (1-F)^{k-[n(1-\lambda)/(1-y)]}] + (1-F)^k \left[\frac{1 - (1-F)^{n(j-1)}}{1 - (1-F)^n} \right]$ $\times \{ q[1 - (1-F)^{n(\lambda-y)/(1-y)}] + s(1-F)^{n(\lambda-y)/(1-y)}$ $\times [1 - (1-F)^{ny(1-\lambda)/(1-y)}] \}$

^a1 = Red signal indication at source intersection. 2 = Queue released on green indication (g_s in eq. 3).

3 = Vehicles Released at arrival rate on green indication.

constructed in TRANSYT from essentially the same OUT patterns, the resulting flow rates will be different. This time dependency is reflected in Equation 20 by the term $[1 - (1-F)^{n(j-1)}]$, hereafter designated the cycle factor.

To demonstrate the impact of the cycle factor on the number of vehicles generated in a TRANSYT run, consider a simulation period lasting m cycles. Since the total number of vehicles leaving the stopline each cycle is qc , the total number of vehicles simulated is mqc . At an observation point t seconds downstream of the stopline, the corresponding number of vehicles arriving in cycle j can be calculated as follows:

$$Q_j = (c/n) \sum_{k=1}^n \text{IN}(k+t, j) \quad (21)$$

and the total number of vehicles simulated in m cycles is

$$Q = \sum_{j=1}^m Q_j \quad (22)$$

The relative cycle error based on m simulated cycles (E_m) is now defined as

$$E_m = [qc - (Q/m)] / qc \times 100 \text{ percent} \quad (23)$$

where Q/m is the average number of vehicles generated per cycle in the predicted platoon, based on a simulation run lasting m cycles.

The variable Q_j in Equation 21 has been calculated for the three flow patterns considered at the source intersection. The results are tabulated in Figure 5.

In order to ascertain the validity of the total flow expressions of Q , two special cases were considered:

1. Travel time is zero ($F = 1$): In this case, the IN flow value in Equation 21 should duplicate the OUT flow values at the source intersection.

2. Travel time is infinity ($F = 0$): In this case, the IN flow values in Equation 21 should duplicate the IN flow values that occur at an isolated intersection.

In both cases, cycle dependency was considered negligible, since j was set to infinity. Formulas for the two cases are summarized in Table 1.

Proof: Based on the OUT patterns derived in Equations 8 through 10, the total flow generated at the source intersection can be expressed as follows:

$$Q_1 = \sum_{k=1}^{n(1-\lambda)} 0 \times (c/n) \quad 1 \leq k \leq n(1-\lambda) \quad (24)$$

$$Q_2 = \sum_{k=[n(1-\lambda)/(1-y)]+1}^{n(1-\lambda)/(1-y)} s \times (c/n) \quad n(1-\lambda) < k \leq [n(1-\lambda)/(1-y)] \quad (25)$$

$$Q_3 = \sum_{[n(1-\lambda)/(1-y)]+1}^n q \times (c/n) \quad [n(1-\lambda)/(1-y)] < k \leq n \quad (26)$$

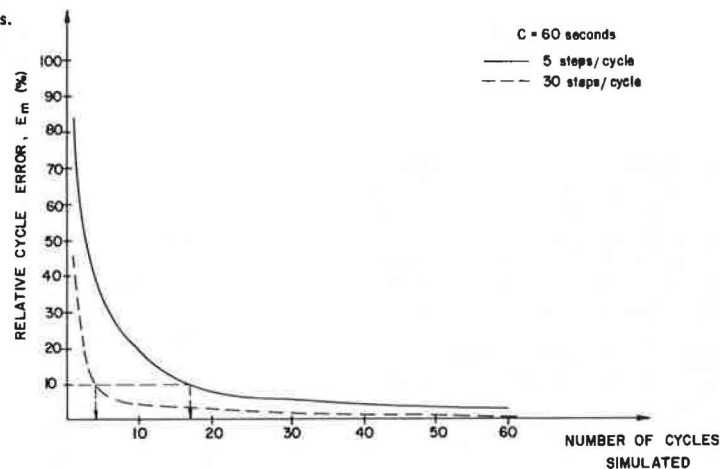
It can be readily shown that the values of Q_1 , Q_2 , and Q_3 correspond precisely to expressions 1, 2, and 3 given in Table 1.

Proceeding in a similar fashion for $t = \infty$, it is observed that the q multipliers in expressions 4, 5, and 6 are actually the durations of the red, saturated-green, and the unsaturated-green intervals, respectively, at the source (and in fact isolated) intersection. In other words, the flow rate in each of the three regions has a fixed value q throughout the cycle, a distinct feature of the isolated-intersection arrival-flow pattern.

A computer program was written to calculate the relative cycle error (E_m) for a variety of travel-

Figure 5. Number of vehicles generated in predicted platoons.

Region	$k_1 \leq k \leq k_2$	$Q_j = \frac{c}{n} \sum_{k_1}^{k_2} IN(k+t, j)$
1	$1 \leq k \leq n(1-\lambda)$	$\frac{c}{n} \left\{ \frac{1 - (1-F)^{n(j-1)}}{1 - (1-F)^n} \times \frac{(1-F)}{F} \times [1 - (1-F)^{n(1-\lambda)}] \right.$ $\times \left\{ q \left[1 - (1-F)^{n(\lambda-y)/(1-y)} \right] + s(1-F)^{n(\lambda-y)/(1-y)} \right.$ $\times \left. \left[1 - (1-F)^{ny(1-\lambda)/(1-y)} \right] \right\} \left. \right\}$
2	$n(1-\lambda) < k \leq \frac{n(1-\lambda)}{(1-y)}$	$\frac{c}{n} \left\{ \frac{qn(1-\lambda)}{(1-y)} + \left\{ \frac{1 - (1-F)^{n(j-1)}}{1 - (1-F)^n} \right. \right.$ $\times q \left[1 - (1-F)^{n(\lambda-y)/(1-y)} \right] + s(1-F)^{n(\lambda-y)/(1-y)}$ $\times \left. \left[1 - (1-F)^{ny(1-\lambda)/(1-y)} \right] \right\} - \frac{s}{(1-F)^{n(1-\lambda)}} \left. \right\}$ $\times \left\{ \frac{(1-F)^{n(1-\lambda)+1}}{F} \times \left[1 - (1-F)^{ny(1-\lambda)/(1-y)} \right] \right\}$
3	$\frac{n(1-\lambda)}{(1-y)} < k \leq n$	$\frac{c}{n} \left\{ \frac{qn(\lambda-y)}{(1-y)} + \left\{ \frac{s[1 - (1-F)^{ny(1-\lambda)/(1-y)}]}{(1-F)^{n(1-\lambda)/(1-y)}} \right. \right.$ $+ \left\{ \frac{1 - (1-F)^{n(j-1)}}{1 - (1-F)^n} \times \left\{ q \left[1 - (1-F)^{n(\lambda-y)/(1-y)} \right] \right. \right.$ $+ s(1-F)^{n(\lambda-y)/(1-y)} \times \left. \left[1 - (1-F)^{ny(1-\lambda)/(1-y)} \right] \right\} \left. \right\}$ $\times \left\{ \frac{(1-F)^{n(1-\lambda)+1}}{F} \times \left[1 - (1-F)^{n(\lambda-y)/(1-y)} \right] \right\}$

Figure 6. Relative cycle error for $t = 20$ s.

time, signal-control, and simulation-run parameters. The results are depicted graphically in Figures 6 and 7. These graphs are useful in the preliminary determination of the required duration of a simulation run, so as to maintain the relative cycle error below a prespecified threshold. For example, for a pair of intersections that are 20 s apart (under free-flow conditions), a relative cycle error of 10 percent will not be exceeded if a simulation period is selected that is (a) 4 cycles long, each made up of 30 steps, or (b) 17 cycles long, each made up of 5 steps. The corresponding values for a travel time of 60 s are 11 steps and 49 cycles.

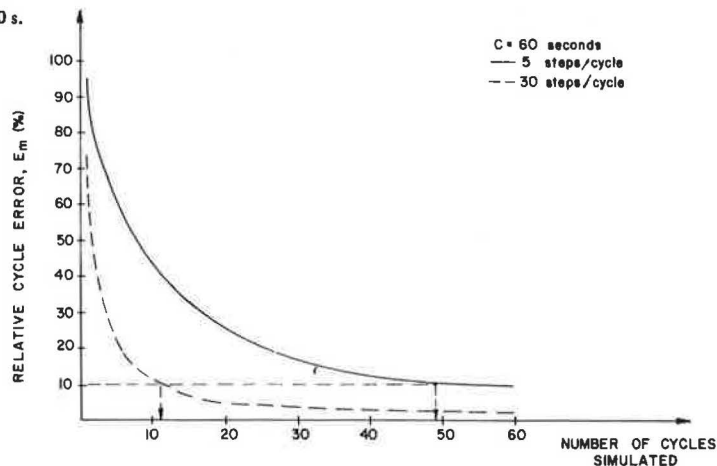
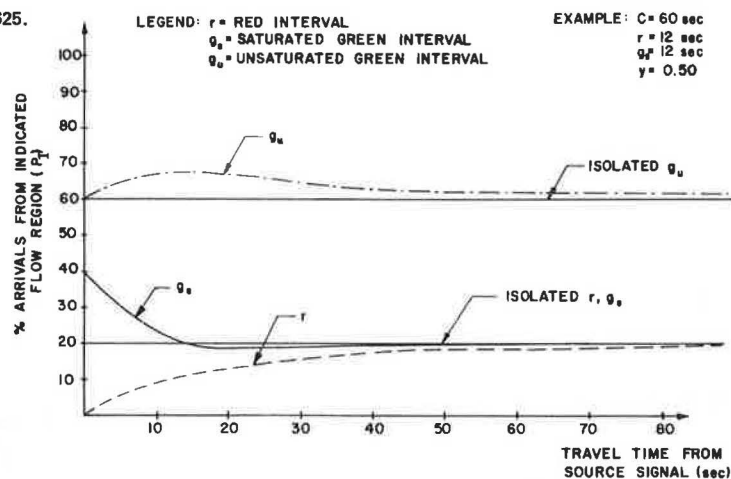
It was also noted that the cycle error is virtually independent of the degree of saturation at the source intersection, at least within the degree of saturation range tested in this study ($0.55 \leq x$

≤ 1.00). Thus in a simulation of a TRANSYT network, the controlling factor in the selection of required simulation time and interval duration is in fact the highest link travel time. Once an acceptable cycle error is attained for that critical link, all other links in the network will automatically satisfy that same requirement.

Of course, if the analytical expressions derived in this study were to substitute for the recursive relationship in TRANSYT, under steady-state cycle conditions (i.e., $j = \infty$), then the cycle error would be totally eliminated.

SIGNAL COORDINATION FEASIBILITY LIMITS

The derivation of analytical solutions to the platoon recursive formula makes it possible to study

Figure 7. Relative cycle error for $t = 60$ s.Figure 8. P_i versus travel-time functions for $x = 0.625$.

the potential value of providing signal coordination between the source intersection and a hypothetical destination intersection located within t seconds of travel. This is accomplished by comparing the actual flow rates arriving at the destination intersection with the predicted flow rates had the destination intersection been considered isolated in nature.

Consider a time interval of length r that corresponds to the red signal indication at the source intersection. For an isolated intersection, the proportion of vehicles arriving in r (P_r) is simply the proportional duration of interval r to the cycle length (c). In other words,

$$P_r = c(1 - \lambda)/c = 1 - \lambda \quad (27)$$

Now if we consider the actual proportion of arrivals in interval r based on platoon dispersion after time t ($P_{r,t}$), it can be shown that

$$P_{r,t} = Q_{\infty,r} / \sum_{i=1}^3 Q_{\infty,i} \times 100 \text{ percent} \quad (28)$$

where $Q_{\infty,i}$ is the steady-state ($j = \infty$) number of vehicle arrivals at the destination intersection that originated in interval i ($i = 1, 2, 3$) at the source intersection. Q -values are depicted in Figure 5.

Similar expressions are developed for vehicle arrivals corresponding to the saturated (g_s) and

unsaturated (g_u) portions of the green interval, as follows:

$$P_{gs} = cy(1 - \lambda)/c(1 - y) = y(1 - \lambda)/(1 - y) \quad (29)$$

$$P_{gs,t} = (Q_{\infty,gs} / \sum_{i=1}^3 Q_{\infty,i}) \times 100 \text{ percent} \quad (30)$$

$$P_{gu} = c(\lambda - y)/c(1 - y) = (\lambda - y)/(1 - y) \quad (31)$$

$$P_{gu,t} = (Q_{\infty,gu} / \sum_{i=1}^3 Q_{\infty,i}) \times 100 \text{ percent} \quad (32)$$

The values of P_r , where I denotes a general interval, are plotted against free-flow travel time between source and destination intersections in Figures 8 and 9 by using three different demand/capacity ratios.

The graphs clearly demonstrate that the source platoon rapidly degenerates (with time) into a uniform flow pattern, although theoretically the two patterns coincide only at $t = \infty$. It is also observed that the rate of platoon degeneration is dependent on the degree of saturation at the source intersection; in general, a platoon degenerates more rapidly at higher degrees of saturation.

To the extent that the platoon-dispersion formula in TRANSYT is valid, Figures 8 and 9 may be used to investigate intersection spacing thresholds for signal coordination. There is no attempt in this paper to develop any such guidelines, in part because of the narrow scope of the numerical examples depicted in the figures. However, as additional data are successfully tested, a mathematical model with

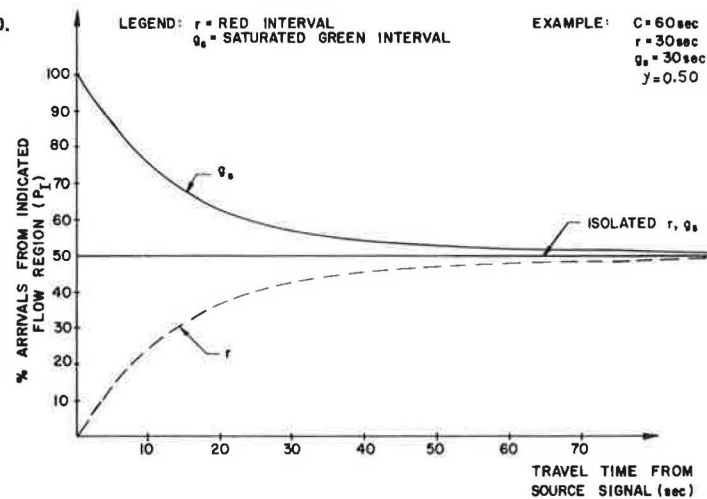
Figure 9. P_I versus travel-time functions for $x = 1.00$.

Table 2. Cycle dependency of queue lengths and delays: numerical example.

Variable	Cycle			
	1st	3rd	10th	50th
Average flow rate in predicted platoon ^a (vehicles/s)	0.032	0.210	0.247	0.249
Average flow rate in green interval ^b (vehicles/s)	0.064	0.221	0.253	0.255
Average queue length ^c (vehicles)	0	2.58	3.55	3.63
Uniform vehicle delay [vehicles/(s-cycle)]	0	154.8	213.0	217.8

^a Average at source intersection = 0.25 vehicle/s.

^b Assuming zero green offset between intersections and green interval of 30 s.

^c Queue length is zero in first cycle since all arrivals occur in the green interval at the destination intersection.

travel time and degree of saturation as independent variables and a measure of the absolute difference between P_I and $P_{I,t}$ as the dependent variable could be used directly to estimate the value of providing signal coordination between a pair of intersections.

ILLUSTRATIVE EXAMPLE

The following numerical example demonstrates the impact of cycle error on the magnitude of uniform vehicle delay, which is a component of the performance index in TRANSYT. Mathematically, delay is expressed as follows:

$$d_u = (c/n) \sum_{k=1}^n m_k \quad (33)$$

where d_u is the uniform delay in vehicles per second per cycle and m_k is the queue length during step k in vehicles c , n , k as defined earlier. m_k is calculated as follows:

$$m_k = \max \{m_k - 1 + (c/n) [IN(k) - OUT(k)], 0\} \quad (34)$$

Thus, for a pair of intersections operating under a simultaneous-progression pattern (i.e., offset = 0) and identical g/c ratios, it is assumed that $c = 60$ s, $g = 30$ s, $t = 60$ s, $n = 10$, $\alpha = 0.35$, $q = 900$ vehicles/h (0.25 vehicle/s), and $s = 1800$ vehicles/h (0.50 vehicle/s).

By using the flow equations shown in Figure 4, the predicted platoon-arrival rates at the destination intersection were calculated. When these IN flow rates were overlaid on the OUT patterns (that

is, zero offset, $g/c = 0.50$ at the destination intersection), queue length and delay estimates were obtained by using Equations 33 and 34.

This procedure was carried out for platoons generated in the 1st, 10th, and 50th cycles of a simulated TRANSYT run. The results are summarized in Table 2.

The average flow rates, queue lengths, and uniform delays exhibited virtually identical variations with cycle number. All were underestimated in cycles 1, 2, and 3. Between the 3rd and the 10th cycles, all performance measures were within 5 percent of their terminal values reached at the later cycles. The location of a knee (point of substantial slope change) in the delay-versus-cycle relationship may be viewed as a desirable upper limit on the simulation time beyond which little will be gained in terms of computational accuracy of the performance measures.

Further work is planned regarding an investigation of cycle-factor impact on the final TRANSYT settings, which was beyond the scope of this paper.

CONCLUSIONS

An analytical solution to the recursive platoon-dispersion formula used in TRANSYT-type models has been developed in this paper.

The findings of the study have potential significance in several aspects of macroscopic traffic simulation modeling:

1. The recursive relationship in TRANSYT contains a time-dependent factor, which in turn underestimates the total flow generated in a simulation run and consequently all performance measures (such as delays and stops) associated with it. The analytical expressions developed in the study give a precise measurement of that factor.

2. Time dependency of flow rates generated in TRANSYT may be reduced by increasing the simulation period, the number of steps per cycle, or both when the recursive relationship is used. As an option, however, it is possible to use the analytical expressions derived in this study under steady-state cycle conditions (i.e., assuming an infinite number of runs), which thus eliminates the time-dependency effect.

3. A number of expressions developed in this study describe the process by which a platoon of vehicles degenerates into a uniform-flow pattern. The rate of degeneration was found to increase with the degree of saturation at the source intersec-

tion. Mathematical models are suggested to evaluate signal coordination feasibility limits between a pair of intersections.

ACKNOWLEDGMENT

This study was supported in part by a University of Illinois at Chicago Research Board Grant.

REFERENCES

1. E. Lieberman and others. NETSIM Model--User's Guide. FHWA, Rept. FHWA-RD-770-44, 1977.
2. Signal Operations Analysis Package--User's Manual. FHWA, 1979.
3. D.I. Robertson. TRANSYT: A Traffic Network Study Tool. British Road Research Laboratory, Crowthorne, Berkshire, England, Rept. BRRL 25S, 1969.
4. Peat, Marwick, Livingstone and Co. SIGOP: Traffic Signal Optimization Program--User's Manual. U.S. Bureau of Public Roads, 1968. NTIS: PB 182835.
5. D.I. Robertson. Contribution to Discussion on Paper 11. Presented at Symposium on Area Control of Road Traffic, Institution of Civil Engineers, London, 1967.
6. National Signal Timing Optimization Project: Executive Summary. ITE Journal, Vol. 52, No. 10, Oct. 1982.
7. TRANSYT-7F User's Manual. FHWA, Feb. 1981.
8. P.A. Sneddon. The Prediction of Platoon Dispersion in the Combination Methods of Linking Traffic Signals. Transportation Research, Vol. 6, 1972, pp. 125-130.

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

Optimization Model for Isolated Signalized Traffic Intersections

W.B. CRONJE

The existing methods for the optimization of isolated fixed-time signalized traffic intersections are applicable either to undersaturated stationary conditions or to oversaturated conditions. As far as is known, no model exists that is applicable to all conditions. A model is developed for the optimization of fixed-time signalized intersections that is applicable to undersaturated as well as to oversaturated conditions. In the model, the macroscopic approach to traffic flow is used. Although it is not so accurate as the microscopic approach, values are obtained for delay and number of stops that are accurate enough for practical purposes and that use much less computer time. Macroscopic simulation is then approximated by the geometric probability distribution. In this case also, values for delay and number of stops are obtained that are accurate enough for practical purposes and that use much less computer time. Consequently, the geometric probability distribution model is recommended for the optimization of fixed-time signalized traffic intersections.

The purpose of this paper is the development of a model for the optimization of fixed-time signalized intersections.

Most of the research in the field of signalized intersections has been done for undersaturated conditions. In this paper, however, we shall not refer to specific shortcomings, but as a result of these shortcomings, it has been decided to develop an accurate model for practical application to undersaturated and oversaturated conditions.

First, microscopic and macroscopic simulation are compared in the stationary zone with reference to average delay and number of stops. The difference is found to be negligible for practical purposes, and macroscopic simulation is used in the further development of the model because it uses much less computer time.

Second, average delay and number of stops are determined by macroscopic simulation in the nonstationary zone. Good agreement is found between the values obtained at the end of the nonstationary zone and those in the stationary zone. Macroscopic simulation in the nonstationary zone can therefore be deemed correct (see Figure 1).

Last, macroscopic simulation is approximated by the geometric probability distribution to further

reduce computer time. Good agreement is found for all practical purposes, and the geometric model is therefore recommended for the optimization of fixed-time signalized intersections.

COMPARISON BETWEEN MICROSCOPIC AND MACROSCOPIC SIMULATION

Macroscopic traffic flow at a signalized intersection is indicated in Figure 2, which shows average arrivals per unit time interval (q), overflow of vehicles at the end of the previous cycle (Q_B), overflow of vehicles at the end of the cycle (Q_E), cycle length (c) in seconds, effective green time (g) in seconds, effective red time (r) in seconds, and saturated flow (s) in vehicles per second. The total delay per cycle (D) is the area under the queue-length diagram:

$$D = [(2Q_B + q)r]/2 + [(qr + Q_B + Q_E)g]/2 \quad (1)$$

The number of stops per cycle (N) is the number of vehicles that arrive while there is a queue plus the overflow at the start of the cycle (Q_B):

$$N = c \cdot q + Q_B \quad (2)$$

Microscopic traffic flow is indicated in Figure 3.

In the macroscopic case, arrival of vehicles per cycle is obtained by generating random numbers. In the microscopic case, gaps between vehicles are obtained similarly.

By working from a zero origin, the times of arrival and departure are obtained; thus the delay is experienced. By summation of the delay for all vehicles, the total delay (D) is obtained. The average delay (d) is then the total delay divided by the sum of all the vehicles arriving during the period considered.

The number of stops is obtained as follows.

Figure 1. Transition from nonstationary to stationary zone as flow increases.

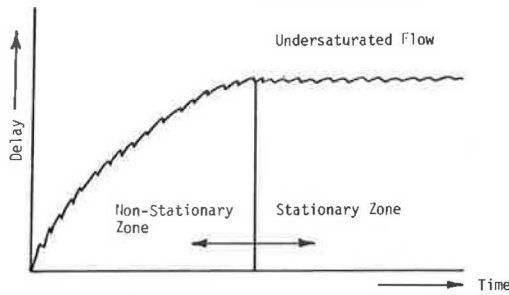


Figure 2. Queue-length diagram with overflow for macroscopic traffic flow.

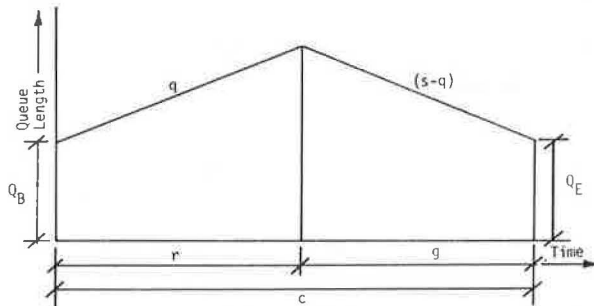
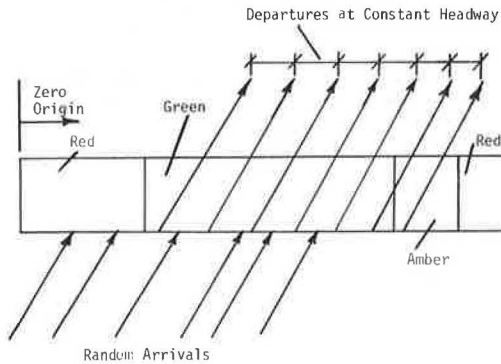


Figure 3. Arrivals and departures at signalized intersection.



If a vehicle is delayed, it is counted as a stop and if a vehicle does not depart during the cycle in which it arrives, an extra stop is counted. The average number of stops is obtained by dividing the total number of stops (N) by the total number of arrivals during the period considered.

The values for average delay and average number of stops as obtained for microscopic and macroscopic simulation are indicated in Table 1, from which it is clear that the difference between microscopic and macroscopic simulation is, for practical purposes, negligible. Because it uses less computer time, macroscopic simulation is used for further analysis.

MACROSCOPIC SIMULATION MODEL

Analysis

The following equations are used:

$$Q_E = Q_B + q \cdot c - s \cdot g \quad (3)$$

$$P(Q_E) = P(Q_B) \cdot P(q \cdot c) \cdot P(s \cdot g) \quad (4)$$

Table 1. Average delay and average number of stops for microscopic and macroscopic simulation.

c	g	s	x	q	Microscopic Simulation		Macroscopic Simulation	
					d	n	d	n
40	12	0.5	0.83	450	25.92	1.23	25.67	1.27
	24	0.5	0.74	800	8.17	0.73	8.20	0.85
60	18	0.5	0.83	450	30.59	1.09	30.77	1.14
	36	0.5	0.74	800	10.74	0.71	10.62	0.81
80	24	0.5	0.83	450	36.23	1.04	36.15	1.07
	48	0.5	0.74	800	13.46	0.71	13.24	0.79
100	32	0.5	0.78	450	35.95	0.94	35.76	0.97
	64	0.5	0.69	800	12.94	0.63	12.68	0.69
120	38	0.5	0.79	450	42.30	0.94	42.24	0.97
	76	0.5	0.70	800	15.66	0.64	15.52	0.70

Notes: Average delay per vehicle in seconds: $d = D/(q \cdot c)$.
 Average number of stops per vehicle: $n = N/(q \cdot c)$.
 Ratio of average number of arrivals per cycle to the maximum number of departures per cycle: $x = (q \cdot c)/(s \cdot g)$.
 q = average number of arrivals per hour.

$$P(D) = P(Q_B) \cdot P(q \cdot c) \cdot P(s \cdot g) \quad (5)$$

$$P(N) = P(Q_B) \cdot P(q \cdot c) \cdot P(s \cdot g) \quad (6)$$

$$E(D) = \sum_i D_i \cdot P(D_i) \quad (7)$$

$$E(N) = \sum_i N_i \cdot P(N_i) \quad (8)$$

where

$q \cdot c$ = average number of arrivals per cycle,

$s \cdot g$ = number of departures per cycle,

$P(Q_E)$ = probability of overflow Q_E ,

$P(D)$ = probability of total delay,

$P(N)$ = probability of total number of stops,

$E(D)$ = expected value of total delay, and

$E(N)$ = expected value of total number of stops.

$P(q \cdot c)$ is obtained from the arrival distribution. $P(s \cdot g)$ is obtained from the departure rate, and $P(Q_B)$ is obtained as follows.

Assume zero flow initially. With zero flow, there can be no overflow at the end of the cycle; thus $Q_E = 0$. But Q_E becomes Q_B for the next cycle; therefore, $Q_B = 0$ for the first cycle and it is clear that $P(Q_B = 0) = 1$ and that $P(Q_B = 1, 2, 3, \dots) = 0$. The probability diagram is therefore as indicated in Figure 4.

By varying the number of arrivals per cycle ($q \cdot c$) between zero and j so that

$$\sum_{i=0}^j P(q \cdot c = i) > 0.9999 \quad (9)$$

and by substituting the possible values of Q_B and $q \cdot c$ in Equation 3, different values for Q_E are obtained.

The procedure will be illustrated with an example. Assume that $s \cdot g$ is an integer such that $P(s \cdot g) = 1$.

Example 1

If we use cycle length (c) of 40 s, saturated flow (s) of 0.5 vehicle/s, flow (q) of 800 vehicles/h, Poisson arrivals, and effective green time (g) of 16 s,

$$s \cdot g = 0.5 \cdot 16 = 8 \text{ vehicles departing per cycle,}$$

$$m = q \cdot c = (800 \cdot 40)/3600 = 8.888 \text{ 888 9 arrivals per cycle.}$$

It is found that j in Equation 9 equals 22. Therefore, $q \cdot c$ is varied between zero and 22. For the first cycle, $Q_B = 0$.

Put $Q_B = 0$ and $s \cdot g = 8$ in Equation 3; then

$$Q_E = q \cdot c - 8 \quad (10)$$

Figure 4. Probability distribution diagram for Q_B in first cycle.

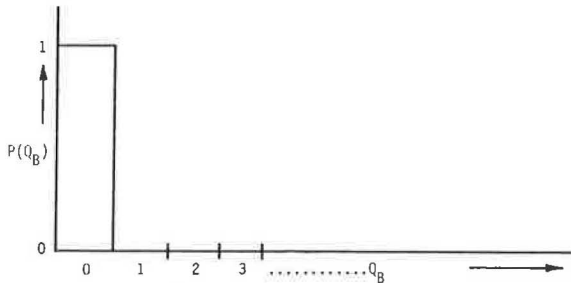


Table 2. Probability of Q_E .

$q \cdot c$	Q_E	$P(Q_E)$	$q \cdot c$	Q_E	$P(Q_E)$
0	-8	0.000 137 9	12	4	0.070 054 9
1	-7	0.001 225 9	13	5	0.047 900 8
2	-6	0.005 448 4	14	6	0.030 413 2
3	-5	0.016 143 5	15	7	0.018 022 6
4	-4	0.035 874 4	16	8	0.010 012 6
5	-3	0.063 776 6	17	9	0.005 235 3
6	-2	0.094 483 9	18	10	0.002 585 4
7	-1	0.119 979 6	19	11	0.001 209 5
8	0	0.133 310 6	20	12	0.000 537 6
9	1	0.131 664 8	21	13	0.000 227 5
10	2	0.117 035 4	22	14	0.000 091 9
11	3	0.094 574 1			

Figure 5. Probability distribution diagram for Q_E .

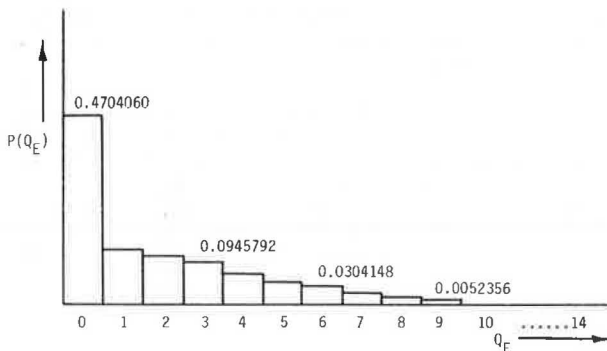


Table 3. Average delay and number of stops by macroscopic simulation.

c	g	s	x	q	Stationary Condition		Nonstationary Condition	
					d	n	d	n
40	12	0.5	0.83	450	25.67	1.27	25.68	1.27
	24	0.5	0.74	800	8.20	0.85	8.21	0.86
60	18	0.5	0.83	450	30.77	1.14	30.82	1.14
	36	0.5	0.74	800	10.62	0.81	10.64	0.81
80	24	0.5	0.83	450	36.15	1.07	36.27	1.08
	48	0.5	0.74	800	13.24	0.79	13.23	0.79
100	32	0.5	0.78	450	35.76	0.97	35.79	0.97
	64	0.5	0.69	800	12.68	0.69	12.67	0.69
120	38	0.5	0.79	450	42.24	0.97	42.25	0.97
	76	0.5	0.70	800	15.52	0.70	15.51	0.70

By substituting $q \cdot c$ in Equation 10 and $P(Q_B = 0) = 1$, $P(s \cdot g) = 1$, and $P(q \cdot c)$ from the Poisson model in Equation 4, the values of $P(Q_E)$ in Table 2 are obtained.

Q_E cannot be negative; thus

$$P(Q_E = 0) = \sum_{i=-8}^0 P(Q_E = i) = 0.470 380 8$$

By adjusting the values in Table 2 so that $\sum P(Q_E) = 1$, the probability diagram in Figure 5 is obtained.

For the next cycle, Q_E becomes Q_B . Values of Q_B from zero to 14 are therefore available. For each value of Q_B , the series of values of $q \cdot c$ from zero to 22 is substituted in Equation 3 and a probability distribution diagram for Q_E for each value of Q_B is obtained.

The probabilities of Q_E on these diagrams are then summed over all the diagrams, that is, over Q_B , according to the following equation:

$$P(Q_E) = \sum_{Q_B} P(Q_B) \cdot P(q \cdot c) \cdot P(s \cdot g) \quad (11)$$

The probabilities obtained are again adjusted to sum to 1, Q_E again becomes Q_B , and the probability distribution for the next cycle is determined.

This procedure is repeated until average delay becomes constant in the unsaturated case or until the increase in average delay from cycle to cycle becomes constant in the oversaturated case.

If this constant average delay in the nonstationary undersaturated case is equal to the average delay as obtained for stationary conditions, then the method whereby average delay is obtained for nonstationary conditions by macroscopic simulation can be deemed to be correct.

The values obtained for average delay and average number of stops for stationary and nonstationary conditions are indicated in Table 3.

From Table 3 it is clear that the differences are negligibly small. Thus the nonstationary analysis can be deemed correct.

APPROXIMATING MACROSCOPIC SIMULATION BY GEOMETRIC PROBABILITY DISTRIBUTION MODEL

The following form of the geometric distribution is suggested as an approximation model:

$$P(K) = (1 - f)f^K \quad (12)$$

where $P(K)$ is the probability of a queue of length K at the start of the cycle and f is the probability of a queue at the start of the cycle. Let

$$f = E(Q_B) / [1 + E(Q_B)] \quad (13)$$

In another paper in this Record, I have shown that for the geometric approximation model the expected overflow at the end of the cycle is

$$E(Q_E) = E(Q_B) + E(q \cdot c) - E(s \cdot g) - \sum_{s \cdot g} P(s \cdot g) \sum_{q \cdot c=0}^{s \cdot g-1} P(q \cdot c) [E(Q_B)(1 - f^{s \cdot g - q \cdot c}) + q \cdot c - s \cdot g] \quad (14)$$

The expected number of stops is

$$E(N) = E(Q_B) + E(q \cdot c) + \sum_{s \cdot g} P(s \cdot g) \sum_{q \cdot c=0}^{s \cdot g-1} P(q \cdot c) \left\{ (q \cdot c)/c / [(s \cdot g)/g] - [(q \cdot c)/c] \right\} [E(Q_B)(1 - f^{s \cdot g - q \cdot c}) + q \cdot c - s \cdot g] \quad (15)$$

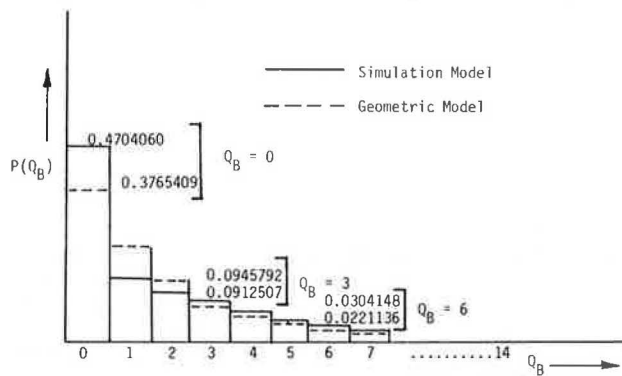
Figure 6. Probability distribution diagram for Q_B .

Table 4. Macroscopic simulation and geometric model at undersaturation.

c	g	s	q	x	Simulation		Geometric Model	
					d	n	d	n
40	12	0.5	450	0.83	25.68	1.27	22.26	1.19
	24	0.5	800	0.74	8.21	0.86	7.88	0.85
60	18	0.5	450	0.83	30.82	1.14	27.93	1.09
	36	0.5	800	0.74	10.64	0.81	10.45	0.81
80	24	0.5	450	0.83	36.27	1.08	33.82	1.05
	48	0.5	800	0.74	13.23	0.79	13.12	0.79
100	32	0.5	450	0.78	35.79	0.97	35.02	0.97
	64	0.5	800	0.69	12.67	0.69	12.66	0.69
120	38	0.5	450	0.79	42.25	0.97	41.53	0.96
	76	0.5	800	0.70	15.51	0.70	15.50	0.70

and the expected total delay is

$$E(D) = E(Q_B) \cdot c + 0.5 [E(q \cdot c) - E(s \cdot g)] + \sum_{s \cdot g} P(s \cdot g) \times \sum_{q \cdot c=0}^{s \cdot g-1} P(q \cdot c) \left(\frac{1}{2} \left\{ \frac{(s \cdot g)}{g} - \frac{(q \cdot c)}{c} \right\} \right) \times \{ E(Q_B) [2 - f^{s \cdot g - q \cdot c} (1 + f)] + (q \cdot c - s \cdot g)^2 - (q \cdot c - s \cdot g - 1)^2 f \} \times [1 / (1 - f)] \quad (16)$$

By applying Equations 12 through 16 to example 1, the probability distribution below is obtained:

Q_B	$P(Q_B)$	Q_B	$P(Q_B)$
0	0.376 540 9	8	0.008 595 6
1	0.234 757 9	9	0.005 359 0
2	0.146 361 9	10	0.003 341 1
3	0.091 250 7	11	0.002 083 1
4	0.056 891 1	12	0.001 298 7
5	0.035 469 2	13	0.000 809 7
6	0.022 113 6	14	0.000 504 8
7	0.013 786 9		

The probability distributions in Table 2 and the tabulation above are indicated in Figure 6.

In Table 4, macroscopic simulation is compared with the geometric model at undersaturation and in Table 5 at oversaturation.

From Figure 6 and Tables 4 and 5, it is clear that there is close agreement between macroscopic simulation and the geometric model.

An application to a two-phase intersection is illustrated in example 2. The intersection data are given below (T_1 , T_2 , and T_3 are consecutive time periods in seconds):

Phase	s	g	q_1	T_1	q_2	T_2	q_3	T_3
1	0.5	16	400	1800	600	2400	800	1200
2	0.5	18	500	1800	750	2400	1000	1200

Table 5. Macroscopic simulation and geometric model at oversaturation.

c	g	s	x	q	Simulation		Geometric Model	
					d	n	d	n
40	24	0.5	1.00	1080	36.46	1.72	34.91	1.68
	24	0.5	1.10	1188	69.38	2.60	70.68	2.63
60	36	0.5	1.00	1080	46.86	1.58	44.70	1.55
	36	0.5	1.10	1188	99.03	2.51	100.74	2.54
80	48	0.5	1.00	1080	56.25	1.50	53.58	1.47
	48	0.5	1.10	1188	128.54	2.47	130.54	2.50
100	64	0.5	1.00	1152	61.59	1.43	58.52	1.40
	64	0.5	1.10	1267	155.38	2.44	157.47	2.46
120	76	0.5	1.00	1140	69.99	1.40	66.51	1.37
	76	0.5	1.10	1254	185.20	2.42	187.41	2.44

Rates of $3.1 \cdot 10^{-2}$ rand/stop and $1.74 \cdot 10^{-4}$ rand/s for total delay, as obtained from research by the National Institute for Transportation and Road Research in Pretoria, Republic of South Africa, are used. The rand is the unit of currency in the Republic of South Africa [1 rand = \$0.78 (1983 U.S.)].

The results obtained for a range of cycle lengths are indicated below, from which it is seen that the minimum cost occurs at the same underlined cycle length:

c	Cost (rands)	
	Simulation Model	Geometric Model
40	122.54	115.24
50	95.13	91.24
60	82.56	79.77
70	74.60	72.29
80	70.90	68.82
90	65.70	63.95
100	65.45	63.83
110	59.08	57.65
120	62.79	61.34

CONCLUSIONS

It is clear from the results indicated in the graphs and tables that the differences between macroscopic simulation and the geometric model are for all practical purposes negligibly small. The fact is substantiated by the results obtained for example 2 and indicated at the end of the previous section. The geometric approximation model for the cost optimization of fixed-time signalized traffic intersections is therefore recommended because it uses much less computer time.

Only the Poisson probability distribution model was used in the research. I have shown (1) that, irrespective of which probability model of the Poisson, binomial, and negative binomial is used for the arrival of vehicles at a signalized intersection, the minimum cost occurs at the same cycle length. The Poisson distribution, being simpler, was therefore used in this research.

REFERENCE

1. W.B. Cronjé. A Model for Use in the Optimization of Fixed-Time Signalized Intersections. Transactions of the Annual Transportation Convention, Vol. 4, Pretoria, Republic of South Africa, 1982.

Comparison of SOAP and NETSIM: Pretimed and Actuated Signal Controls

ZOLTAN A. NEMETH AND JAMES R. MEKEMSON

Delay and fuel-consumption rates estimated by the relatively easy-to-use, deterministic Signal Operations Analysis Package (SOAP) were compared with results generated by the microscopic and stochastic Network Simulation Model (NETSIM). The study involved three cases of isolated signalized intersections: two-phase pretimed controller, two-phase fully actuated controller, and multi-phase pretimed controller. More than 80 combinations of left-turning and through traffic volumes were investigated in each case. Whereas SOAP estimates excess fuel consumption at intersections, NETSIM generates total fuel consumption. The difference between the two was found to be fairly uniform and corresponded to a realistic 18-mile/gal fuel efficiency under uninterrupted 30-mph flow conditions. In terms of delay prediction, SOAP and NETSIM are found to be entirely compatible after the differences in delay definitions, SOAP's more conservative left-turn saturation-flow-rate relationship, and NETSIM's delay sensitivity to unit extensions for actuated signal controllers were taken into account. In addition, the volume/capacity ratio at which SOAP begins to overestimate delay due to the use of Webster's delay equation may be lower than now assumed. Last, the difference between SOAP and NETSIM average delays can probably be reduced by a more studied coordination between SOAP and NETSIM input parameters. Evidence is offered to the operating engineer that the easy-to-use SOAP produced results supported by the sophisticated NETSIM.

Poorly timed traffic signals result in the inefficient use of intersection capacity and contribute to delay and fuel waste. The considerable amount of research effort that has been directed in the past at the problem of efficient signal timing has resulted in a variety of tools that range from relatively easily applied computer programs to sophisticated and complex digital simulation models.

The Network Simulation Model (NETSIM) is an example of a complex digital simulation model. It was developed for the Federal Highway Administration (FHWA) (1). Peat, Marwick, Mitchell and Company and General Applied Science Laboratories developed the UTCS-1, the earlier version of the model. Although it is basically a network simulation model, it is also applicable to the analysis of a single signalized intersection. The Signal Operations Analysis Package (SOAP) is one example of a relatively easy-to-use tool. It offers a practical method of signal timing and intersection performance evaluation in the form of a computer program. This program was developed for the Florida Department of Transportation and FHWA by the University of Florida. The implementation package has been widely distributed (2).

Although SOAP and NETSIM are very different in their computational base, they are generally assumed to produce realistic results. Whereas SOAP is a deterministic, macroscopic model based on a set of simple equations, NETSIM is a stochastic, microscopic, digital simulation model that handles each vehicle separately. NETSIM is based on car-following and lane-changing rules; it considers different vehicle types and also recognizes conflicts between left turns and oncoming traffic as well as the impact of traffic that is backed up from the preceding intersection.

SOAP is a relatively simple method to use, whereas in comparison NETSIM is very complex. The difference raises a very intriguing question: Can SOAP and NETSIM produce compatible results under similar traffic conditions? A positive answer would of course reflect favorably on both NETSIM and SOAP.

OBJECTIVE AND SCOPE OF STUDY

The objective of this study was to apply both SOAP

and NETSIM in the analysis of a signalized intersection and compare generated delays and fuel consumption for consistency. Three cases were investigated.

Case 1 involved the intersection of two two-lane roadways. Left-turn lanes were added on all approaches. The intersection was controlled by a pretimed two-phase signal.

Case 2 involved the same intersection layout but the signal control was changed. A fully actuated two-phase traffic signal was specified in this case.

Case 3 involved the intersection of two four-lane roadways. Left-turn bays were added on each approach. The intersection was controlled by a pretimed multiphase signal. Left-turn phases were provided for all left-turning movements.

In each case, the east-west roadway was considered the minor street. Approach volumes and left-turn percentages were held constant on this roadway in each case. Seventy percent of the major-street volume was northbound and 30 percent was southbound. At least 80 combinations of intersection volumes and left-turn percentages were investigated in each case.

DEFINITION OF SELECTED PERFORMANCE MEASURES

Delay and fuel consumption were selected as performance measures.

Average Delay

SOAP uses the widely known Webster delay formula to estimate delay:

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

where

- d = delay per vehicle (s) on particular movement of intersection approach,
- c = cycle length (s),
- λ = proportion of effective green time (g) given to movement (i.e., g/c),
- q = approach flow (vehicles/s),
- x = degree of saturation (i.e., $q/\lambda s$), and
- s = saturation flow (vehicles/s).

From these average delays, total delays per approach and, by summation, total intersection delays are calculated. From total intersection delays the average delay to all vehicles passing through the intersection is determined. (Further references to average delay in this paper will be to this average delay.)

Delay is defined in SOAP as the difference in average travel time through the intersection and the travel time for a vehicle that is not stopped or slowed down by a signal.

The definition of delay in NETSIM appears to be identical: Total delay time is computed as the difference between the total travel time and idealized travel time for each link based on a designated target speed. However, a significant difference is introduced by the microscopic nature of NETSIM, in that each vehicle is assigned an individual target

speed, which ranges from 75 to 127 percent of the link target speed. The travel time of a vehicle is thus influenced not only by the traffic signal, but also by friction among individual vehicles within the traffic stream. Since the total length of the upstream and downstream links simulated in NETSIM for this study amounts to 4000 ft, a significant proportion of the total delay may be unrelated to the traffic signal itself.

Delays generated by NETSIM, therefore, could be expected to be higher than delays calculated by Webster's delay equation, since the latter is based on estimated time spent in queue.

Total Delay

Total delay is defined by both NETSIM and SOAP as the product of average delay and total intersection volume.

Fuel Consumption

The definitions of fuel consumption are clearly different in the two methods.

NETSIM generates the total gallons of fuel consumed by all vehicles. The computation is based on an assumed proportion of vehicle types and corresponding fuel-consumption rates by each type during idling, accelerating, and traveling at a given speed.

SOAP, on the other hand, computes only the excess fuel consumption due to idling delays and accelerations from stopped positions. Two equations are used to calculate these two components.

If the two methods are compatible, NETSIM total fuel consumption is expected to be consistently higher than SOAP excess fuel consumption by a fairly uniform amount.

STUDY PROCEDURES

SOAP was first run to compute the optimal cycle lengths and splits for each 30-min simulation period corresponding to the different volume combinations. The signal timing selected by SOAP was then specified for NETSIM as input.

In general, inputs were specified for both NETSIM and SOAP with care in order to achieve maximum compatibility. No grades, parking, or pedestrian interference were assumed. Desired free-flow speed was specified as 30 mph.

Delays and fuel-consumption levels were then generated by both SOAP and NETSIM.

Scatter plots and regression equations were developed as a first step to establish that the patterns of delays and fuel consumption generated by the two methods under the different conditions were consistently similar and that the differences in actual values did not conflict with what is expected due to the differences in definitions, as explained above.

The regression analysis presented in Table 1 indicates that the differences were very consistent, and, as expected, NETSIM produced higher average delay and higher fuel consumption.

ANALYSIS OF DIFFERENCES

Average Delay

Case 1: Pretimed Two-Phase

In Figures 1, 2, 3, and 4, average delays predicted by SOAP and NETSIM are presented related to total intersection volumes and left-turn percentages on the major roadway.

As stated earlier, directional distribution on

Table 1. Correlation between NETSIM and SOAP outputs.

Case	Sample Size	Regression Equation	R ²	SE
1	80	ADNET = 4.432 + 1.333ADSO	0.911	1.898
		FCNET = 2.903 + 5.203FCSO	0.991	0.743
2	88	ADNET = 7.832 + 1.85ADSO	0.863	1.643
		FCNET = 4.463 + 4.92FCSO	0.985	0.951
3	85	ADNET = -6.174 + 1.294ADSO	0.936	1.602
		FCNET = 19.673 + 2.800FCSO	0.991	0.689

Notes: These equations were developed for the sole purpose of testing the level of correlation. ADNET = NETSIM average delay (s); ADSO = SOAP average delay (s); FCNET = NETSIM total fuel consumption (gal); FCSO = SOAP excess fuel consumption (gal).

Figure 1. Average delay profiles: SBLT, 30 percent (case 1).

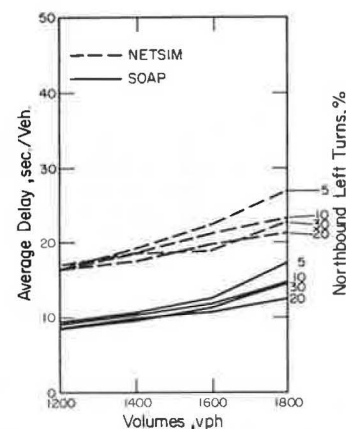


Figure 2. Average delay profiles: SBLT, 5 percent (case 1).

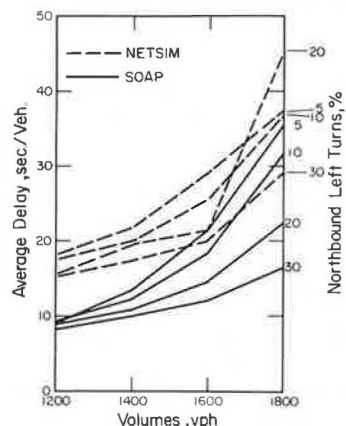


Figure 3. Average delay profiles: NBLT, 5 percent (case 1).

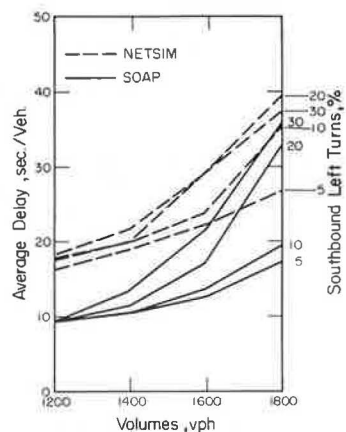
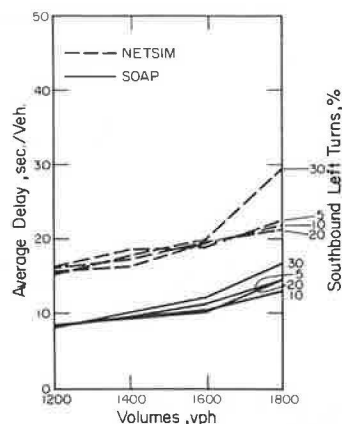


Figure 4. Average delay profiles: NBLT, 30 percent (case 1).



the major street was 70 percent northbound and 30 percent southbound. In Figures 1 and 2, southbound left-turn (SBLT) percentages were held constant, and northbound left-turn (NBLT) percentages and intersection volumes were varied. In Figures 3 and 4, the NBLT percentages were held constant.

The following observations can be made:

1. There is a fairly uniform 7.5- or 8-s basic difference in average delays between SOAP and NETSIM in the lower volume range. At the specified 30-mph free-flow (or target) speed, this difference corresponds to an approximate 2.4-mph drop in average speeds within NETSIM. It is not unreasonable to assume that the simulated internal friction, as explained in the definition of NETSIM average delay, could realistically account for that much speed difference.

2. The second observation is that the patterns of delays predicted by SOAP and NETSIM as volumes and left-turn percentages were varied are similar in all four figures. The only major exception to this second observation is the high NETSIM delay estimate seen in Figure 2 for an intersection volume of 1800 with 20 percent NBLT. This particular data point demonstrates the highly stochastic nature of NETSIM and therefore the occasional random appearance of a measure of performance outside the general pattern of results.

3. The difference in delays between NETSIM and SOAP is observed to be less uniform at the higher intersection volume levels in Figures 2 and 3. A possible explanation for the nonuniform delay differences may be the more conservative left-turn saturation-flow-rate relationship within SOAP as compared with NETSIM. This would result in higher degrees of saturation and therefore higher delay estimates in SOAP.

In general, average delays increase as intersection volumes increase. Delays, however, increase especially rapidly with increased intersection volumes when

1. SBLT percentages are high (compare Figure 1 with Figure 2 and note that SBLT percentages are 5 and 30 percent, respectively) and

2. NBLT percentages are low (compare Figure 3 with Figure 4 and note that NBLT percentages are 5 and 30 percent, respectively).

A review of the approach-by-approach distribution of total delays at an extreme combination of northbound (5 percent) and southbound (20 percent) left-turn percentages will help to understand the above observations (intersection volume, 1800 vehicles/h;

cycle length, 90 s; major/minor green split, 82.6/17.4 percent):

Type of Delay	Hours per 30-Min Period
Total	7.973
Northbound	
Through lane	0.902
Left lane	0.049
Southbound	
Through lane	0.132
Left lane	1.300
Minor	
Through lanes (each)	2.378
Left lanes (each)	0.417

At any given approach volume, low NBLT percentages correspond to high northbound through percentages and also to high conflicts between northbound through and SBLTs. The high delay on the SBLT lane (corresponding to low NBLT) thus becomes understandable.

However, the major source of high average intersection delay is the delay on the minor street. Apparently, the long cycle time (90 s) and short minor green phase (17.4 percent) created a nearly saturated condition on the minor street.

In conclusion, the pattern of average delays under various volume and left-turning percentages as calculated by the Webster delay equation in SOAP is similar to that generated by the stochastic NETSIM model for case 1. At least some of the differences in average delays (NETSIM delays are higher than SOAP delays) can be related to the travel-time delay simulated in NETSIM over the 2000-ft approach link and 2000-ft-long departure link. Some of the non-uniform delay differences might be attributed to different left-turn saturation-flow-rate relationships in the two models.

Case 2: Actuated Two-Phase

Results are presented in Figures 5, 6, and 7. The following observations can be made:

1. The basic difference in delay estimates increased sharply as compared with the two-phase pretimed case. (Compare Figure 4 with Figure 7.) NETSIM estimated delays 14-18 s higher than SOAP as compared with a difference of 7-8 s for the pretimed two-phase case. Closer examination reveals that SOAP delay estimates for actuated control are about 2 s lower than the pretimed case and are therefore acceptable. However, NETSIM delay estimates are 5-8 s higher than those for the pretimed case. A review of NETSIM data input parameters showed that the unit extension used in the simulation was chosen to be 4 s. Studies have shown that an actuated controller with 4-s unit extensions will result in delays much higher than those with an optimally timed pretimed controller. A 3-s or lower unit extension would have generated much lower NETSIM delay estimates. This sensitivity of delay to unit extension is clearly shown in Figure 8 (3).

2. The second observation is that SOAP greatly overestimated delay for three data points, as shown in Figure 5 at 1600 and 1800 vehicles/h. These high delay estimates are probably due to conservative left-turn saturation-flow rates, which in turn result in near-saturated conditions where Webster's equations are known to overestimate delays.

Case 3: Pretimed Multiphase

Results of the multiphase pretimed-signal case are presented in Figures 9, 10, and 11. The first ob-

servation is that NETSIM and SOAP results are very close. NETSIM delays tend to be higher by a few seconds only, except at the highest intersection volume, at which differences in delay increase. The small delay differences at the lower volume levels are a result of reduced friction between vehicles of varying target speeds in NETSIM. This reduction in friction is due to a segregation of vehicles with respect to individual target speeds between the two lanes on each approach and exit link. In general, as volume increases, segregation of vehicles with respect to target speeds declines due to fewer lane-changing opportunities, and hence NETSIM-simulated delay increases. The patterns of delay correspond-

ing to volume and left-turn percentage changes are identical. Observe the delay pattern at 3500 vehicles/h:

1. At 15 percent NBLT, delays increase rapidly in the higher volume range as SBLT percentages increase, as shown in Figure 9.
2. At 30 percent NBLT, delay is still highest when SBLT percentage is the highest, but delays overlap at lower SBLT percentages (Figure 10).
3. At 35 percent NBLT (Figure 11), the relationship between average delays and SBLT percentages

Figure 5. Average delay profiles: NBLT, 5 percent (case 2).

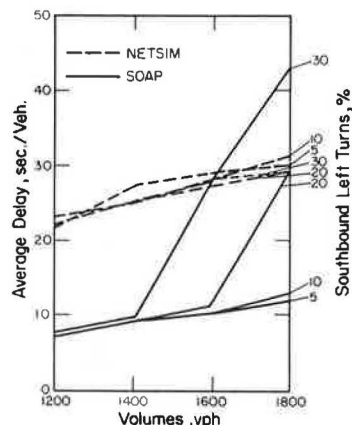


Figure 6. Average delay profiles: NBLT, 20 percent (case 2).

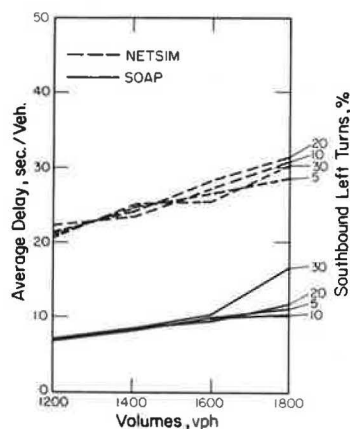


Figure 7. Average delay profiles: NBLT, 30 percent (case 2).

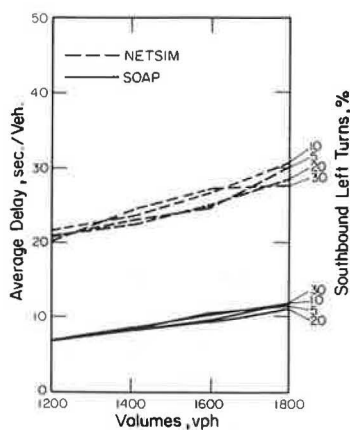


Figure 8. Relationship between unit extension and delay.

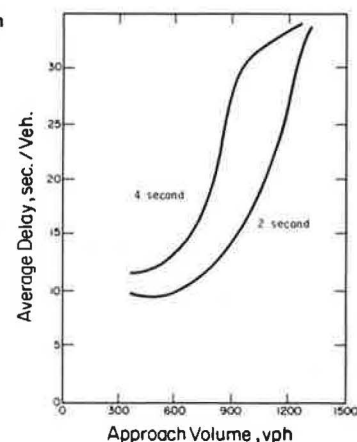


Figure 9. Average delay profiles: NBLT, 15 percent (case 3).

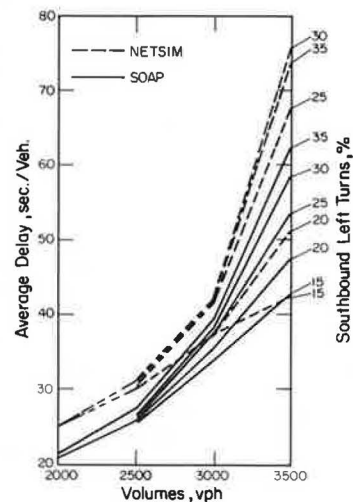


Figure 10. Average delay profiles: NBLT, 30 percent (case 3).

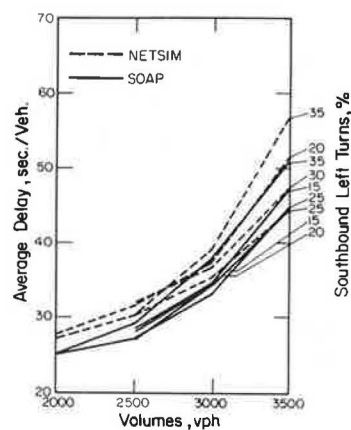
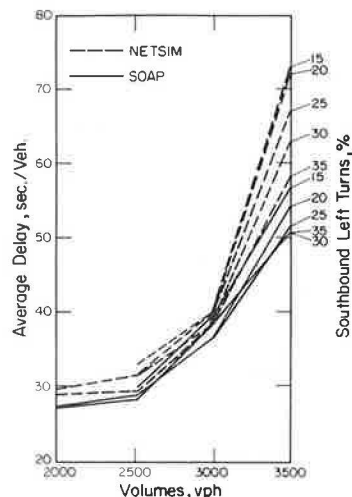


Figure 11. Average delay profiles: NBLT, 35 percent (case 3).



reverses. Average delays decrease as SBLT percentages increase both in SOAP and in NETSIM.

In conclusion, NETSIM and SOAP produce delays in case 3 that are almost identical except at the higher through and left-turn conflicts of the highest intersection volumes.

Fuel Consumption

Samples of SOAP excess fuel consumption and NETSIM fuel consumption are presented in Table 2. As stated earlier, NETSIM calculates total fuel consumed by traffic over a 4000-ft length, whereas SOAP estimates only excess fuel consumption caused by the traffic signal. If both methods are correct, then the difference between the two should be consistent.

The last column of Table 2 presents this difference, expressed with an accuracy of 0.001 gal.

This difference in case 1 (pretimed two-phase signal) is between 0.041 and 0.042 gal/vehicle. Over the 4000-ft section simulated by NETSIM, this difference corresponds to 18.0-18.5 miles/gal.

In case 2 (two-phase fully actuated signal) the difference is between 0.042 and 0.043 gal/vehicle, only slightly higher than in case 1. In case 3 the difference is slightly lower in general than in case 1 or case 2.

In summary, the difference between NETSIM fuel consumption and SOAP excess fuel consumption is very consistent and corresponds to approximately 18-mile/gal fuel efficiency under uninterrupted flow conditions.

COMPARISON OF LEFT-TURN SATURATION-FLOW RATES

Comparison of the NETSIM and SOAP delay estimates indicated that SOAP overestimates delay for some high-volume and left-turn combinations. It was suggested earlier that SOAP's left-turn saturation-flow rate is conservative as compared with that of NETSIM. As part of a larger research project, a graph of left-turn saturation-flow rate versus opposing volume was developed by using NETSIM. As can be seen in Figure 12, SOAP's left-turn saturation-flow rate is indeed conservative as compared with that of NETSIM.

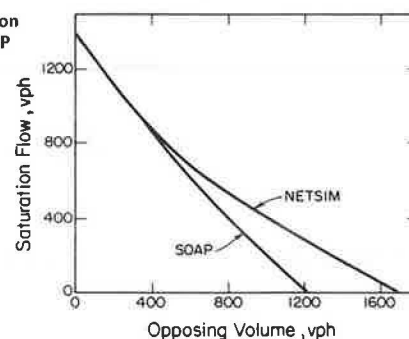
Incorporating the NETSIM developed left-turn saturation-flow-rate relationship into SOAP and again running the experiments from Figure 5 resulted in much lower delay estimates for the three high-delay points of Figure 5, as shown in Figure 13. Only one point, 30 percent SBLT, is still higher than de-

Table 2. Correlations between SOAP excess fuel consumption and NETSIM total fuel consumption.

Major-Street Left Turns		Inter- section Volume (vehicles/h)	Fuel Consumption (gal/30 min)		Difference ^a (gal/vehicle)
NBLT (%)	SBLT (%)		SOAP	NETSIM	
Case 1					
10	10	1800	8.32	46.28	0.042
10	10	1600	7.08	40.12	0.041
10	10	1400	6.19	35.08	0.041
10	10	1200	5.19	29.77	0.041
10	20	1800	9.15	49.01	0.042
10	20	1600	7.15	40.67	0.042
10	20	1400	6.09	35.06	0.041
10	20	1200	5.19	29.94	0.041
10	30	1800	12.02	49.08	0.041
10	30	1600	7.51	40.93	0.042
10	30	1400	6.15	35.23	0.042
10	30	1200	5.18	29.96	0.041
Case 2					
10	10	1800	8.63	47.45	0.043
10	10	1600	7.75	41.41	0.042
10	10	1400	6.51	35.93	0.042
10	10	1200	5.20	30.68	0.042
10	20	1800	8.22	47.36	0.043
10	20	1600	7.44	41.77	0.043
10	20	1400	6.53	36.08	0.042
10	20	1200	5.21	30.49	0.042
10	30	1800	8.33	47.39	0.043
10	30	1600	7.26	41.41	0.043
10	30	1400	6.42	36.28	0.043
10	30	1200	5.23	30.35	0.042

^aSample calculation, first row: difference in gallon consumption = $46.28 - 8.32 = 37.96$ gal; 30-min volume = $1800/2 = 900$ vehicles; difference = $37.96/900 = 0.0422$ gal/vehicle.

Figure 12. Comparison of NETSIM and SOAP left-turn saturation-flow rates.



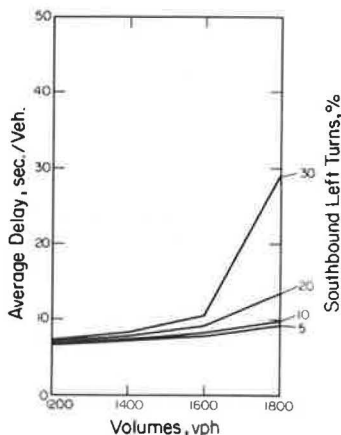
sired. This overestimate is probably due to the fact that Webster's delay equation is very sensitive to high degrees of saturation.

SUMMARY AND CONCLUSION

Delays and fuel consumption estimated by the deterministic SOAP based largely on Webster's delay equation were compared with results generated by the microscopic NETSIM.

More than 80 combinations of 30-min left-turn and

Figure 13. SOAP average delay profiles with NETSIM derived left-turn saturation-flow rates.



through traffic volumes were studied in three different signal-control variations.

Results were almost identical when a multiphase pretimed traffic signal was simulated. In the case of a two-phase pretimed signal, NETSIM delays were somewhat higher, as expected, and the relative changes in delays corresponding to relative changes in volumes and left turns were similar.

In the case of a two-phase fully actuated signal, the difference between NETSIM and SOAP average delay was higher than in the other two cases but can be explained by too long a unit extension specified in NETSIM. SOAP appeared to overestimate delays at a few points, which corresponded to conditions of high

volume/capacity ratios. The overestimated delays for these points were due to conservative SOAP estimates of left-turn saturation-flow rates.

The patterns of NETSIM and SOAP delays were similar enough to indicate that with additional research the correlation could be further improved. In this study no attempt was made to change any of the first set of inputs (unit extension time, minimum green, maximum green, lost time, etc.) in order to increase the correlation between NETSIM and SOAP.

After differences in definitions had been accounted for, NETSIM and SOAP fuel-consumption estimates were found to be identical for all three cases.

ACKNOWLEDGMENT

The data that provided the foundation for this paper were generated by Ching-Sheng Yang, Mohsen Zarean, and Lei Wu, graduate students in the Department of Civil Engineering, Ohio State University, under our direction.

REFERENCES

1. P. Ross and D. Gibson. Review of Road Traffic Network Simulation Models. TRB, Transportation Research Record 644, 1977, pp. 36-41.
2. Signal Operations Analysis Package. FHWA, July 1979.
3. Selecting Traffic Signal Control at Individual Intersections. NCHRP, Rept. 233, June 1981.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Analysis of Existing Formulas for Delay, Overflow, and Stops

W.B. CRONJÉ

An analysis is made of existing formulas for average delay, average overflow, and average number of stops for undersaturated conditions. The examination of these formulas covers a large variation in flows and cycle lengths, so recommendations are based on a thorough examination. The formulas examined are those developed by Webster, Miller, and Newell. It is concluded that the Newell formulas give the most accurate results.

The delay formulas that are predominant in practice are those developed by Webster (1), Miller (2,3), and Newell (4). Hutchinson (5) examined these formulas for accuracy. The standard of comparison is, however, a derived formula. Furthermore, Hutchinson (5) covered only average delay. In this paper, however, the standard of comparison is computer simulation, and in addition to average delay, average overflow and average number of stops are also examined. The reason for this is that in the optimization of fixed-time signalized intersections, delay as well as number of stops should be used in the optimization process.

Throughout the comparison the value of I , the variance-to-mean ratio of flow per cycle, is taken as 1 because it has been shown (6) that for the optimization of fixed-time signalized traffic intersections it is immaterial which probability distribution

is used for the arriving traffic at a signal. The Poisson distribution, because of its simplicity, is therefore used.

ANALYSIS OF AVERAGE DELAY AND OVERFLOW FORMULAS

The Webster (1) equations are as follows:

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

$$Q_0 = q[d - 0.5 \cdot c(1-\lambda)] \quad Q_0 \leq 0 \quad (2)$$

The Miller 1 equations (2) are as follows:

$$d = [(1-\lambda)/2(1-\lambda \cdot x)] \{ c(1-\lambda) + [(2 \cdot x - 1)/q(1-x)] + [(1+\lambda \cdot x - 1)/s] \} \quad (3)$$

$$Q_0 = 0 \text{ for } x \leq 0.5 \\ = 1(2 \cdot x - 1)/2(1-x) \text{ for } x > 0.5 \quad (4)$$

The Miller 2 equations (3) are as follows:

$$d = [(1-\lambda)/2(1-\lambda \cdot x)] \{ c(1-\lambda) + \{ \exp[-(4/3)] [(\lambda \cdot c \cdot s)^{0.5} (1-x)/x] \} \div q(1-x) \} \quad (5)$$

$$Q_0 = \exp[-(4/3)] (\lambda \cdot c \cdot s)^{0.5} [(1-x)/x] / 2(1-x) \quad (6)$$

The Newell 1 equations (4) are as follows:

$$d = [c(1-\lambda)^2/2(1-\lambda \cdot x)] + [I \cdot H(\mu)x/2 \cdot q(1-x)] + [I(1-\lambda)/2 \cdot s(1-\lambda \cdot x)^2] \quad (7)$$

$$Q_0 = I \cdot H(\mu)x/2(1-x) \quad (8)$$

A suggested modification to Newell 1 gives Newell 2:

$$d = [c(1-\lambda)^2/2(1-\lambda \cdot x)] + [I \cdot H(\mu)x/2 \cdot q(1-x)] \quad (9)$$

where

d = average delay (s/vehicle),

Q_0 = average overflow at end of cycle,

$y = q/s$ = ratio of average arrival rate to saturation flow,

$\lambda = g/c$ = proportion of cycle that is effectively green,

g = effective green time (s),

I = variance-to-mean ratio of flow per cycle,

x = ratio of average number of arrivals per cycle to maximum number of departures per cycle,

c = cycle length (s),

q = average number of arrivals per unit time, and

s = saturation flow (vehicles/s).

It can be shown that

$$H(\mu) = \exp[-\mu - (\mu^2/2)]$$

where

$$\mu = (1-x) \cdot (s \cdot g)^{0.5}$$

The values obtained for average delay and average overflow and the simulation values are indicated in Tables 1 and 2.

If we take the simulation values as the standard and calculate the standard deviation for the values in Tables 1 and 2, the values given below are obtained:

Variable	Standard Formula				
	Webster	Miller 1	Miller 2	Newell 1	Newell 2
Avg de- lay	2.061	3.820	2.122	1.445	1.471
Avg over- flow	0.293	1.063	0.193	0.204	

ANALYSIS OF AVERAGE NUMBER OF STOPS

Figure 1 is a queue-length diagram for a signal cycle with overflow at the end of the cycle in which the overflow of vehicles at the end of previous cycle (Q_0), the overflow of vehicles at the end of

Table 1. Average delay for macroscopic simulation and standard formulas.

					Avg Delay (d)					
c	g	s	x	q	Simulation	Standard Formula				
						Webster	Miller 1	Miller 2	Newell 1	Newell 2
40	12	0.5	0.50	270	13.07	13.76	11.65	11.95	13.42	12.45
			0.70	378	17.12	17.32	18.22	15.89	17.59	16.47
			0.90	486	38.42	37.60	42.10	38.15	40.06	38.75
			0.95	513	78.90	70.25	75.82	71.58	73.62	72.25
40	24	0.5	0.50	540	5.20	5.99	4.74	4.61	5.52	4.70
			0.70	756	7.34	8.35	8.00	6.28	7.85	6.66
			0.90	972	18.37	18.82	20.31	16.61	19.95	18.06
			0.95	1026	36.68	35.14	37.35	33.05	37.22	35.06
60	18	0.5	0.50	270	18.47	19.37	17.42	17.50	18.75	17.78
			0.70	378	22.32	23.11	24.42	21.15	22.74	21.62
			0.90	486	43.71	43.47	48.81	42.93	45.06	43.74
			0.95	513	81.17	76.12	82.67	76.25	78.67	77.30
60	36	0.5	0.50	540	7.40	8.24	7.03	6.87	7.72	6.90
			0.70	756	9.76	10.96	10.75	8.76	10.16	8.97
			0.90	972	20.76	21.85	23.79	19.04	22.29	20.40
			0.95	1026	37.63	38.29	41.07	35.42	39.69	37.52
80	24	0.5	0.50	270	24.08	25.00	23.18	23.17	24.29	23.32
			0.70	378	27.92	28.99	30.62	26.76	28.22	27.10
			0.90	486	49.22	49.51	55.52	48.14	50.36	49.05
			0.95	513	85.59	82.19	89.52	81.32	84.01	82.64
80	48	0.5	0.50	540	9.67	10.50	9.31	9.15	9.97	9.16
			0.70	756	12.32	13.59	13.51	11.37	12.66	11.47
			0.90	972	23.46	24.96	27.27	21.72	24.86	22.97
			0.95	1026	40.36	41.56	44.79	38.05	42.37	40.20
100	32	0.5	0.50	288	28.41	29.31	27.65	27.57	28.60	27.64
			0.70	403	32.23	33.49	35.19	31.12	32.44	31.31
			0.90	518	51.73	52.89	59.05	50.62	52.93	51.59
			0.95	547	84.23	83.38	90.93	81.37	84.34	82.94
100	64	0.5	0.50	576	9.98	10.81	9.70	9.53	10.31	9.53
			0.70	806	12.79	14.10	13.96	11.93	13.14	11.96
			0.90	1037	23.56	25.38	27.60	21.70	24.88	22.87
			0.95	1094	38.70	40.90	44.05	36.65	41.28	38.94
120	38	0.5	0.50	285	34.14	35.00	33.42	33.32	34.32	33.35
			0.70	399	38.18	39.51	41.47	37.10	38.33	37.21
			0.90	513	57.82	59.57	66.28	56.79	59.09	57.76
			0.95	542	90.74	91.78	99.99	89.08	92.20	90.81
120	76	0.5	0.50	570	12.26	13.08	11.97	11.81	12.59	11.81
			0.70	798	15.48	16.79	16.76	14.64	15.82	14.64
			0.90	1026	26.60	28.69	31.21	24.78	27.79	25.81
			0.95	1083	42.21	44.79	48.34	40.13	44.69	42.38

the cycle (Q_B), and the effective red time in seconds (r) are shown. The total delay per cycle is the area under the queue-length curve such that

$$D = [(2 \cdot Q_B + q \cdot r)/2] + [(q \cdot r + Q_B + Q_E)g/2] \quad (10)$$

where D is the total delay per cycle in seconds.

The number of stops per cycle is the number of arrivals while there is a queue plus the overflow at the beginning of the cycle such that

$$N = c \cdot q + Q_B \quad (11)$$

where N is the number of stops per cycle.

Figure 2 is a queue-length diagram for a signal

cycle without overflow at the end of the cycle. This case also has to be considered.

$$D = [(2 \cdot Q_B + q \cdot r)/2] + [(q \cdot r + Q_B)^2/2(s - q)] \quad (12)$$

$$N = \{r + [(q + Q_B)q/(s - q)]\} + Q_B \quad (13)$$

The average delay is

$$d = D/(q \cdot c) \quad (14)$$

and the average number of stops is

$$n = N/(q \cdot c) \quad (15)$$

Table 2. Average overflow for macroscopic simulation and standard formulas.

c	g	s	x	q	Avg Overflow (Q_0)				
					Simulation	Standard Formula			
						Webster	Miller 1	Miller 2	Newell 1
40	12	0.5	0.50	270	0.06	0.00	0.00	0.04	0.07
			0.70	378	0.43	0.35	0.67	0.41	0.43
			0.90	486	3.34	3.19	4.00	3.48	3.42
			0.95	513	9.28	8.02	9.00	8.42	8.34
40	24	0.5	0.50	540	0.02	0.00	0.00	0.01	0.02
			0.70	756	0.25	0.07	0.67	0.23	0.24
			0.90	972	2.97	2.92	4.00	3.00	3.00
			0.95	1026	8.27	7.74	9.00	7.85	7.87
60	18	0.5	0.50	270	0.03	0.00	0.00	0.02	0.04
			0.70	378	0.32	0.22	0.67	0.30	0.32
			0.90	486	3.14	3.03	4.00	3.21	3.19
			0.95	513	8.62	7.85	9.00	8.11	8.09
60	36	0.5	0.50	540	0.01	0.00	0.00	0.00	0.01
			0.70	756	0.16	0.00	0.67	0.15	0.15
			0.90	972	2.64	2.66	4.00	2.67	2.69
			0.95	1026	7.45	7.49	9.00	7.43	7.51
80	24	0.5	0.50	270	0.02	0.00	0.00	0.01	0.02
			0.70	378	0.25	0.10	0.67	0.23	0.24
			0.90	486	2.97	2.90	4.00	3.00	3.00
			0.95	513	8.27	7.72	9.00	7.85	7.87
80	48	0.5	0.50	540	0.00	0.00	0.00	0.00	0.00
			0.70	756	0.10	0.00	0.67	0.10	0.09
			0.90	972	2.41	2.42	4.00	2.42	2.45
			0.95	1026	7.15	7.28	9.00	7.10	7.22
100	32	0.5	0.50	288	0.01	0.00	0.00	0.00	0.01
			0.70	403	0.18	0.00	0.66	0.17	0.17
			0.90	518	2.71	2.72	3.97	2.74	2.75
			0.95	547	7.72	7.50	8.93	7.49	7.56
100	64	0.5	0.50	576	0.00	0.00	0.00	0.00	0.00
			0.70	806	0.06	0.00	0.66	0.07	0.05
			0.90	1037	2.17	2.13	4.01	2.17	2.19
			0.95	1094	6.60	6.96	8.93	6.66	6.81
120	38	0.5	0.50	285	0.00	0.00	0.00	0.00	0.01
			0.70	399	0.15	0.00	0.67	0.14	0.13
			0.90	513	2.59	2.65	4.00	2.63	2.65
			0.95	542	7.59	7.65	9.18	7.54	7.64
120	76	0.5	0.50	570	0.00	0.00	0.00	0.00	0.00
			0.70	798	0.04	0.00	0.67	0.05	0.03
			0.90	1026	2.00	1.91	4.00	2.01	2.01
			0.95	1083	6.45	6.85	9.00	6.50	6.66

Figure 1. Queue-length diagram for cycle with overflow at end of cycle.

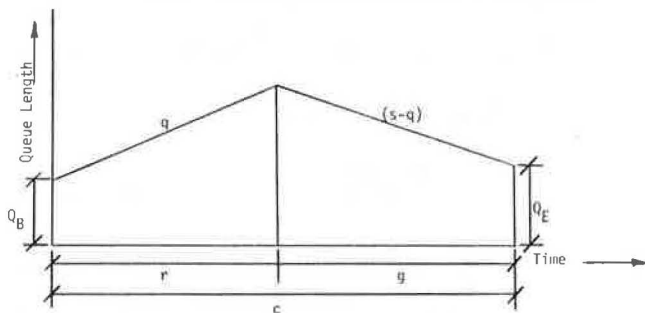


Figure 2. Queue-length diagram for cycle without overflow at end of cycle.

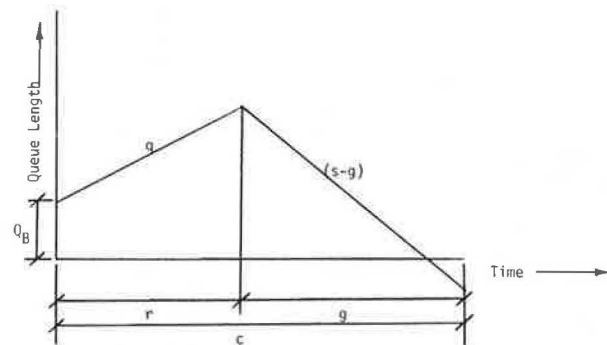


Table 3. Average number of stops for macroscopic simulation and standard formulas.

c	g	s	x	q	Avg No. of Stops (n)				
					Simulation	Standard Formula			
						Webster	Miller 1	Miller 2	Newell 1
40	12	0.5	0.50	270	0.90	0.82	0.82	0.84	0.85
			0.70	378	1.04	0.99	1.09	1.01	1.01
			0.90	486	1.60	1.59	1.74	1.64	1.63
			0.95	513	2.62	2.41	2.58	2.48	2.46
40	24	0.5	0.50	540	0.64	0.57	0.57	0.57	0.58
			0.70	756	0.81	0.70	0.83	0.74	0.74
			0.90	972	1.21	1.27	1.37	1.28	1.28
			0.95	1026	1.69	1.68	1.79	1.69	1.69
60	18	0.5	0.50	270	0.87	0.82	0.82	0.83	0.83
			0.70	378	0.97	0.93	1.02	0.95	0.95
			0.90	486	1.36	1.37	1.49	1.40	1.39
			0.95	513	2.00	1.92	2.05	1.95	1.95
60	36	0.5	0.50	540	0.62	0.57	0.57	0.57	0.57
			0.70	756	0.77	0.69	0.78	0.71	0.71
			0.90	972	1.08	1.16	1.25	1.16	1.17
			0.95	1026	1.40	1.44	1.53	1.43	1.44
80	24	0.5	0.50	270	0.86	0.82	0.82	0.83	0.83
			0.70	378	0.94	0.90	0.99	0.92	0.92
			0.90	486	1.25	1.27	1.37	1.28	1.28
			0.95	513	1.71	1.68	1.79	1.69	1.69
80	48	0.5	0.50	540	0.60	0.57	0.57	0.57	0.57
			0.70	756	0.75	0.69	0.76	0.70	0.70
			0.90	972	1.03	1.11	1.19	1.11	1.11
			0.95	1026	1.27	1.32	1.39	1.31	1.32
100	32	0.5	0.50	288	0.83	0.81	0.81	0.81	0.81
			0.70	403	0.92	0.88	0.95	0.90	0.90
			0.90	518	1.16	1.19	1.28	1.19	1.19
			0.95	547	1.49	1.49	1.59	1.49	1.50
100	64	0.5	0.50	576	0.55	0.53	0.53	0.53	0.53
			0.70	806	0.70	0.65	0.71	0.66	0.66
			0.90	1037	0.97	1.02	1.14	1.03	1.03
			0.95	1094	1.17	1.23	1.29	1.22	1.22
120	38	0.5	0.50	285	0.83	0.81	0.81	0.81	0.81
			0.70	399	0.91	0.88	0.94	0.89	0.89
			0.90	513	1.12	1.15	1.23	1.15	1.15
			0.95	542	1.41	1.42	1.51	1.42	1.42
120	76	0.5	0.50	570	0.56	0.54	0.54	0.54	0.54
			0.70	798	0.70	0.66	0.70	0.66	0.66
			0.90	1026	0.95	0.98	1.12	0.99	0.99
			0.95	1083	1.13	1.19	1.25	1.18	1.18

By substituting the average-overflow equations (Equations 2, 4, 6, and 8) into Equations 10 through 15, the values for average number of stops are obtained, which are indicated in Table 3.

If we take the simulation values as the standard and calculate the standard deviation for the values in Table 3, the values below are obtained:

Variable	Standard Formula			
	Webster	Miller 1	Miller 2	Newell 1
Avg no. of stops	0.060	0.094	0.049	0.050

CONCLUSIONS AND RECOMMENDATIONS

From an inspection of the values for average delay given above, it is clear that the Newell formulas for average delay approximate the simulated values much more closely than the other formulas. Newell 1 gives a slightly better approximation than Newell 2. For practical purposes, this difference is negligible and the fact that Newell 2 contains one term less than Newell 1 makes the former equation, namely, Equation 9, preferable for calculating average delay at all signalized intersections for undersaturated stationary conditions.

From an inspection of the values for average overflow, it is seen that Miller 2 approximates the simulated values more closely than the other formulas. However, since Newell is, for practical purposes, as accurate and since Equation 9, which has already been suggested for calculating delay, also contains $H(\mu)$, the Newell equation, Equation 8, is

recommended for calculating average overflow at all signalized intersections for undersaturated stationary conditions.

From an inspection of the values for average number of stops, it follows that for the same reasons as in the case of average overflow, the Newell equation for average overflow is suggested for calculating the average number of stops for undersaturated stationary conditions.

Hutchinson (5) indicates that within the practical limits in which flow and saturation flows can be measured, any formula can be used for calculating average delay. The Newell 2 delay equation, because of its simplicity and accuracy, however, is recommended for the calculation of average delay at fixed-time signalized traffic intersections.

REFERENCES

1. F.V. Webster. Traffic Signal Settings. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, TRRL Tech. Paper 39, 1958.
2. A.J. Miller. Settings for Fixed-Cycle Traffic Signals. Operational Research Quarterly, Vol. 14, 1963.
3. A.J. Miller. The Capacity of Signalized Intersections in Australia. Australian Road Research Board, ARRB Bull. 3, 1968.
4. G.F. Newell. Approximation Methods for Queues with Application to the Fixed-Cycle Traffic Light. SIAM Review, Vol. 7, 1965.
5. T.P. Hutchinson. Delay at a Fixed-Time Traffic

Signal--II: Numerical Comparisons of Some Theoretical Expressions. Transportation Science, Vol. 6, 1972.

6. W.B. Cronjé. A Model for Use in the Optimization of Fixed-Time Signalized Intersections. Transac-

tions of the Annual Transportation Convention, Vol. 4, Pretoria, Republic of South Africa, 1982.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Derivation of Equations for Queue Length, Stops, and Delay for Fixed-Time Traffic Signals

W.B. CRONJÉ

The existing methods for the calculation of queue length, number of stops, and delay for isolated traffic intersections are applicable either to undersaturated stationary conditions or to oversaturated conditions. As far as is known, no model exists that is applicable to all conditions. Equations are derived for the calculation of queue length, number of stops, and delay for isolated fixed-time signalized intersections that are applicable to undersaturated as well as to oversaturated conditions. In the derivation the macroscopic approach to traffic flow is used. This approach has been shown to be sufficiently accurate for practical purposes. Traffic flow at a signalized intersection is considered a Markov process. Equations are derived for expected queue lengths, expected number of stops, and expected total delay. These equations can also be used for the optimization of isolated fixed-time signalized intersections.

Traffic flow at a signalized intersection is a Markov process. The states being considered are the queue lengths at the beginning and the end of the signal cycle, and the time interval over which changes in these states take place is the length of the signal cycle.

The equation governing the states is as follows:

$$Q_E = Q_B + q \cdot c - s \cdot g \quad Q_E \geq 0 \quad (1)$$

where

Q_E = overflow of vehicles at end of cycle,
 Q_B = overflow of vehicles from previous cycle,
 q = average arrivals per unit time interval,
 c = cycle length (s),
 g = effective green time (s),
 s = saturated flow (vehicles/s),
 $q \cdot c$ = number of arriving vehicles per cycle, and
 $s \cdot g$ = maximum number of departing vehicles per cycle.

Equation 1 can be represented by the transition probability matrix shown in Figure 1, in which $P(Q_B, Q_E)$ is the probability of transition from state Q_B to state Q_E , $P(s \cdot g)$ is the probability distribution of departing vehicles, and $P(q \cdot c)$ is the probability distribution of arriving vehicles.

Equation 1 is illustrated by Figure 2, in which r is effective red time in seconds.

DERIVATION

Consider one approach to an intersection controlled by a fixed-time signal. Consider one cycle on the approach in which vehicles are expected to arrive according to a distribution $P(q \cdot c)$. The saturation-flow vehicles are distributed according to $P(s \cdot g)$.

Let $P(Q_B, Q_E)$ be the probability of an overflow of Q_E vehicles at the end of the cycle, given

an overflow Q_B at the start of the cycle. The form of this probability is shown as a matrix in Figure 3.

The expected overflow at the end of the cycle is given by

$$\begin{aligned} E(Q_E) &= \sum_{Q_E=0}^{\infty} Q_E \cdot P(Q_E) \\ &= \sum_{s \cdot g} P(s \cdot g) \left[\sum_{q \cdot c=0}^{\infty} P(q \cdot c) \sum_{Q_B=0}^{\infty} (Q_B + q \cdot c - s \cdot g) P(Q_B) \right. \\ &\quad \left. - \sum_{q \cdot c=0}^{s \cdot g-1} P(q \cdot c) \sum_{Q_B=0}^{s \cdot g-q \cdot c-1} (Q_B + q \cdot c - s \cdot g) P(Q_B) \right] \\ &= E(Q_B) + E(q \cdot c) - E(s \cdot g) \\ &= \sum_{s \cdot g} P(s \cdot g) \sum_{q \cdot c=0}^{s \cdot g-1} P(q \cdot c) \sum_{Q_B=0}^{s \cdot g-q \cdot c-1} (Q_B + q \cdot c - s \cdot g) P(Q_B) \quad (2) \end{aligned}$$

A queue-length diagram with overflow at the end of the cycle is indicated in Figure 4.

The total number of stops per cycle is the number of vehicle arrivals while there is a queue. Overflow vehicles stop twice.

If there is overflow at the end of the cycle, the total number of stops per cycle is as follows:

$$N = Q_B + q \cdot c \quad (3)$$

A queue-length diagram without overflow at the end of the cycle is indicated in Figure 5.

If there is no overflow at the end of the cycle, the total number of stops per cycle is as follows:

$$N = Q_B + r \cdot q + [(Q_B + r \cdot q) / (s - q)] q \quad (4)$$

Figure 1. Transition probability matrix for Equation 1.

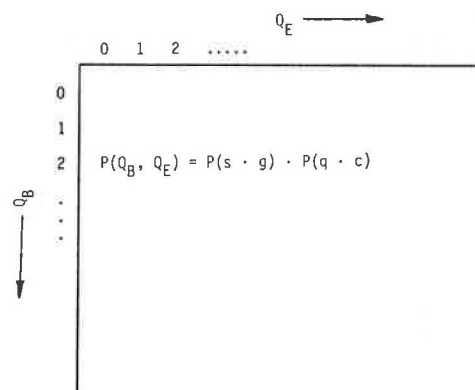


Figure 2. Queue-length diagram illustrating Equation 1.

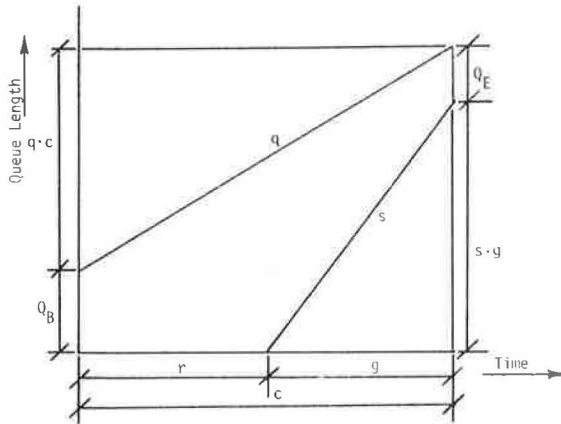
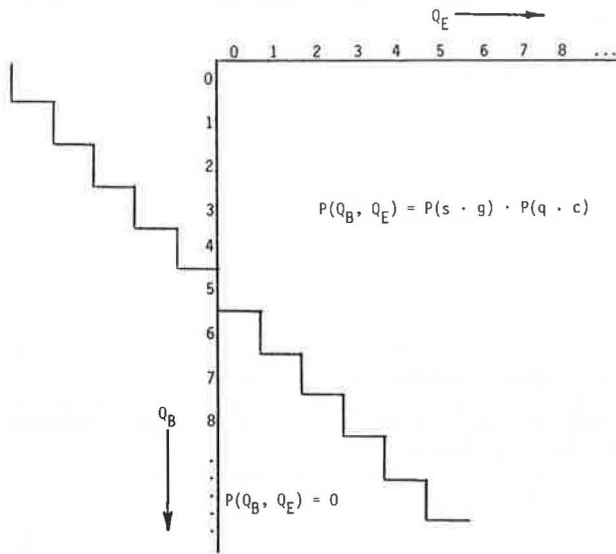


Figure 3. Transition probability matrix.



Replacing r in Equation 4 by $(c - g)$ gives

$$N = Q_B + c \cdot q + [(q \cdot c)/c] \left\{ (Q_B + q \cdot c - s \cdot g) / \left\{ [(s \cdot g)/g] - [(q \cdot c)/c] \right\} \right\} \quad (5)$$

Except for the third term of Equation 5, Equations 3 and 5 are identical. The third term of Equation 5 clearly covers the case of no overflow at the end of the cycle and is applicable to the zone to the left of the origin in Figure 3.

The expected number of stops per cycle is therefore given by

$$\begin{aligned} E(N) &= \sum_0^{\infty} N \cdot P(N) \\ &= \sum_{s \cdot g} P(s \cdot g) \left\{ \sum_{q \cdot c=0}^{\infty} P(q \cdot c) \sum_{Q_B=0}^{\infty} \left[Q_B + q \cdot c + [(q \cdot c)/c] \right. \right. \\ &\quad \left. \left. \cdot (Q_B + q \cdot c - s \cdot g) / \left\{ [(s \cdot g)/g] - [(q \cdot c)/c] \right\} \right] \cdot P(Q_B) \right\} \\ &= E(Q_B) + E(q \cdot c) + \sum_{s \cdot g} P(s \cdot g) \sum_{q \cdot c=0}^{s \cdot g-1} P(q \cdot c) \sum_{Q_B=0}^{s \cdot g-q \cdot c-1} P(Q_B) \cdot [(q \cdot c)/c] \\ &\quad \times (Q_B + q \cdot c - s \cdot g) / \left\{ [(s \cdot g)/g] - [(q \cdot c)/c] \right\} \quad (6) \end{aligned}$$

Figure 4. Queue-length diagram with overflow at end of cycle.

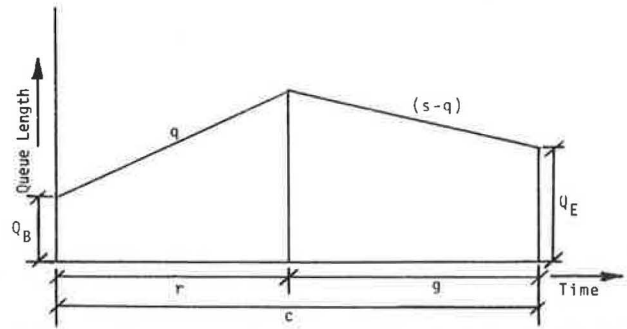
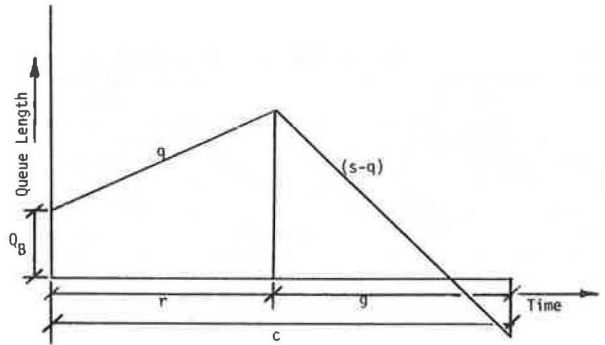


Figure 5. Queue-length diagram without overflow at end of cycle.



A queue-length diagram with overflow at the end of the cycle can also be represented by Figure 2.

Total delay is the area under the queue-length curve.

If there is overflow at the end of the cycle, total delay is given by (see Figure 2)

$$D = Q_B \cdot c + 0.5(q \cdot c - s \cdot g) \quad (7)$$

If there is no overflow at the end of the cycle, total delay is given by (see Figure 5)

$$D = Q_B \cdot r + q \cdot r \cdot (r/2) + [(Q_B + q \cdot r)/(s - q)] \cdot [(Q_B + q \cdot r)/2] \quad (8)$$

Replacing r in Equation 8 by $(c - g)$ gives

$$D = Q_B \cdot c + 0.5(q \cdot c - s \cdot g) + 0.5 \left\{ (Q_B + q \cdot c - s \cdot g) \right. \\ \left. \div \left\{ [(s \cdot g)/g] - [(q \cdot c)/c] \right\} \right\} \quad (9)$$

If the same reasoning is applied to Equations 7 and 9 as in the case of the number of stops, the expected total delay is given by

$$\begin{aligned} E(D) &= c \cdot E(Q_B) + 0.5[c \cdot E(q \cdot c) - g \cdot E(s \cdot g)] + \sum_{s \cdot g} P(s \cdot g) \sum_{q \cdot c=0}^{s \cdot g-1} P(q \cdot c) \\ &\quad \times \sum_{Q_B=0}^{s \cdot g-q \cdot c-1} P(Q_B) \left\{ (Q_B + q \cdot c - s \cdot g)^2 / \left\{ [(s \cdot g)/g] - [(q \cdot c)/c] \right\} \right\} \quad (10) \end{aligned}$$

Assume that the overflow at the start of the cycle is distributed according to the following geometric distribution:

$$P(Q_B) = (1 - f) f^{Q_B} \quad (11)$$

with

$$f = E(Q_B) / [1 + E(Q_B)] \quad (12)$$

Some properties of the geometric probability distribution are

$$\sum_{i=0}^n (1-f)^i = 1 - f^{n+1} \quad (13)$$

$$\sum_{i=0}^n (i-n)(1-f)^i = [f/(1-f)](1-f^n) - n \quad (14)$$

$$\sum_{i=0}^n (i-n)^2(1-f)^i = [1/(1-f)] \{ [f/(1-f)] [2-f^n(1+f)] + n^2 - f(n+1)^2 \} \quad (15)$$

The last part of Equation 2 is as follows:

$$\sum_{Q_B=0}^{s \cdot g - q \cdot c - 1} (Q_B + q \cdot c - s \cdot g) P(Q_B) \quad (16)$$

In expression 16, put $n = (s \cdot g - q \cdot c)$ and substitute $P(Q_B)$ from Equation 11. Then expression 16 becomes

$$\begin{aligned} \sum_{Q_B=0}^{n-1} (Q_B - n)(1-f)^{Q_B} &= \sum_{Q_B=0}^n (Q_B - n)(1-f)^{Q_B} - (n-n)(1-f)^n f^n \\ &= \sum_{Q_B=0}^n (Q_B - n)(1-f)^{Q_B} \end{aligned} \quad (17)$$

If Equation 14 is applied, Equation 17 becomes

$$\begin{aligned} [f/(1-f)](1-f^n) - n &= \{ E(Q_B)/[1 + E(Q_B)] \} / \{ 1 - \{ E(Q_B) \\ &\quad \div [1 + E(Q_B)] \} \} (1-f^n) - n = E(Q_B)(1-f^n) - n \\ &= E(Q_B)(1-f^{s \cdot g - q \cdot c}) - s \cdot g + q \cdot c \end{aligned}$$

Equation 2 therefore becomes

$$\begin{aligned} E(Q_E) &= E(Q_B) + E(q \cdot c) - E(s \cdot g) - \sum_{s \cdot g} P(s \cdot g) \sum_{q \cdot c=0}^{s \cdot g-1} P(q \cdot c) \\ &\quad \times [E(Q_B)(1-f^{s \cdot g - q \cdot c}) + q \cdot c - s \cdot g] \end{aligned} \quad (18)$$

The transformation of Equation 6 is identical, which gives

$$\begin{aligned} E(N) &= E(Q_B) + E(q \cdot c) + \sum_{s \cdot g} P(s \cdot g) \sum_{q \cdot c=0}^{s \cdot g-1} P(q \cdot c) \{ [(q \cdot c)/c] / \{ [(s \cdot g)/g] \\ &\quad - [(q \cdot c)/c] \} \} [E(Q_B)(1-f^{s \cdot g - q \cdot c}) + q \cdot c - s \cdot g] \end{aligned} \quad (19)$$

The numerator of the last part of Equation 10 is

$$\sum_{Q_B=0}^{s \cdot g - q \cdot c - 1} P(Q_B)(Q_B + q \cdot c - s \cdot g)^2 \quad (20)$$

If the same transformation is applied, expression 20 becomes

$$\sum_{Q_B=0}^n (Q_B - n)^2(1-f)^{Q_B} - (n-n)(1-f)^n f^n = \sum_{Q_B=0}^n (Q_B - n)^2 \times (1-f)^{Q_B} \quad (21)$$

If Equation 15 is applied, Equation 21 becomes

$$\begin{aligned} [1/(1-f)] \{ [f/(1-f)] [2-f^n(1+f)] + n^2 - (n+1)^2 f \} \\ = [1/(1-f)] \{ E(Q_B)[2-f^{s \cdot g - q \cdot c}(1+f)] + (s \cdot g - q \cdot c)^2 \\ - (s \cdot g - q \cdot c + 1)^2 f \} \end{aligned}$$

Equation 10 therefore becomes

$$\begin{aligned} E(D) &= c \cdot E(Q_B) + 0.5[c \cdot E(q \cdot c) - g \cdot E(s \cdot g)] + \sum_{s \cdot g} P(s \cdot g) \sum_{q \cdot c=0}^{s \cdot g-1} P(q \cdot c) \\ &\quad \times \{ 1/2 \{ [(s \cdot g)/g] - [(q \cdot c)/c] \} \} \{ E(Q_B)[2-f^{s \cdot g - q \cdot c}(1+f)] \\ &\quad + (q \cdot c - s \cdot g)^2 - (s \cdot g - q \cdot c + 1)^2 f \} [1/(1-f)] \end{aligned} \quad (22)$$

where Equation 18 gives $E(Q_B)$, the expected overflow at the end of the cycle; Equation 19 gives $E(N)$, the expected number of stops per cycle; and Equation 22 gives $E(D)$, the expected total delay per cycle.

CONCLUSIONS AND RECOMMENDATIONS

I have shown in another paper in this Record that the equations developed in this paper for expected queue length, expected number of stops, and expected total delay show very close agreement with simulation.

They are also applicable to undersaturated as well as to oversaturated conditions. By assigning monetary rates to number of stops and delay as calculated by Equations 19 and 22 and varying the green times for the various phases, the optimum cycle length at the minimum cost can be obtained.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Developmental Study of Implementation Guidelines for Left-Turn Treatments

HAN-JEI LIN AND RANDY B. MACHEMEHL

At signalized intersections, the common treatment for improving left-turn performance is to increase left-turn capacity by installing a left-turn bay or a separate left-turn phase. However, under given traffic conditions and geometric configurations, there have been no universally accepted guidelines for ascertaining the need for a left-turn treatment. In this research, the TEXAS traffic simulation model is employed to study the capacity and performance of left-turn movements at signalized intersections in order to devise warrants for left-turn treatments. Since left-turn performance is germane to left-turn capacity, existing methods for estimating left-turn capacity are thoroughly reviewed and a new method that can yield reasonable estimates for left-turn capacity under general conditions of left-turn movements is proposed. Furthermore, different measures of effectiveness are used to evaluate the performance of left-turn movements under various traffic conditions. With a set of delay criteria, critical conditions of left-turn movements are identified. Finally, a new capacity-based warrant is derived from the relationship between the critical left-turn volume and left-turn capacity.

Left-turn maneuvers at signalized at-grade intersections have been recognized as highly problematic. Numerous guidelines have been used to indicate the need for separate left-turn lanes and signal phases, yet none seems to have achieved general acceptance. This paper represents a summary of some significant findings of a three-year research effort directed toward development of guidelines for implementation of left-turn treatments. The study was sponsored by the Texas Department of Highways and Public Transportation in cooperation with the Federal Highway Administration, U.S. Department of Transportation.

REVIEW OF WARRANT CONCEPTS

At signalized intersections, the common treatment for improving left-turn performance is to increase left-turn capacity by adding a bay or a separate left-turn phase. However, under given traffic conditions and geometric configurations, there have been no universal guidelines for traffic engineers to determine whether a bay or a separate left-turn phase is justified. The variations in existing guidelines stem from different methodologies and criteria adopted for evaluating left-turn performance. The methodologies could be analytical models, simulation models, or field observations, whereas the criteria may be a certain level of delay, conflict, or accidents. The resulting guidelines usually will fall into five categories of warrants: delay, volume, capacity, conflict, and accident. Although conflict and accident warrants are useful for the trade-off analysis of a left-turn treatment, study of them by analytical or simulation analysis is very difficult. Thus, only the first three types of warrants will be discussed here. Existing left-turn warrants will be reviewed, and by applying a set of delay criteria to left-turn performance curves (1), critical conditions of left-turn operations can be defined. Efforts will be devoted to developing a general form of left-turn warrant that can identify the need for a left-turn treatment under various traffic conditions and geometric configurations.

SEPARATE LEFT-TURN PHASE

Agent and Deen (2) conducted a survey of warrants currently being used by state highway agencies for

installing a separate left-turn phase and found that numerous discrepancies exist:

Type of Warrant	Left-Turn Warrant
Delay	Left-turn delay in excess of two cycles One left-turner in 1 h being delayed more than one cycle
Volume	Product of left-turn and opposing volumes < 50 000 Product of left-turn and opposing volumes > 100 000 More than two vehicles per approach per cycle during peak hour 50 or more left-turn vehicles in 1 h on one approach and average speed of through traffic > 45 mph > 100 left-turn vehicles during peak hour Left-turn volume > 90 vph Left-turn ADT > 500 for two-lane roadway 100-150 left-turn vehicles during peak hour (small cities) 150-200 left-turn vehicles during peak hour (large cities) 120 left-turn vehicles in design hour 90-120 left-turn vehicles in design hour > 100 turns per hour
Accident	Five or more left-turn accidents within 12-month period

It has also been observed (3,4) that a left-turn phase, when not required, will cause more delay to drivers during other phases and even to left-turners. Therefore, it is very important to have clear and effective guidelines for implementing a separate left-turn phase.

In order to develop warrants, a set of criteria must be chosen. If criteria on delay are employed, left-turn warrants in terms of delay, volume, and capacity can be obtained. A volume warrant may be a minimum left-turn volume level or a product of the left-turn and opposing volumes. The latter is also called the volume-product warrant. From the tabulation above, it can be seen that a minimum left-turn volume level is the most popular type of left-turn warrant. However, this type of warrant does not include the interactive effect of opposing traffic volume and the number of opposing lanes. It also makes no distinction between the left-turn and the opposing volumes. For example, if a left-turn phase is justified when the product of the left-turn and the opposing volumes is greater than 50 000, it does not matter whether there are 500 vph and 100 vph of opposing and left-turn volumes, respectively, or the other way around. Moreover, for a single opposing flow of 100 vph, according to the volume-product warrants shown below, the warranted left-turn volumes would be higher than the left-turn capacity

estimated by any of the commonly used estimation methods discussed elsewhere (1):

No. of Opposing Lanes	Product of Opposing and Left-Turn Peak-Hour Volumes		
	Agent and Deen (2)	SSITE	Texas Transportation Institute (5)
1	50 000	45 000	50 000
2	100 000	90 000	90 000
3		135 000	110 000

A recent report (5) presented a capacity warrant in which a separate left-turn phase is recommended if the ratio of left-turn demand to capacity is greater than 0.7. This capacity warrant can be misleading, as pointed out by Lin (1), because two traffic conditions with the same degree of left-turn saturation may not be equally severe for left-turn operations.

Left-turn operations evaluated with different performance measures by using the TEXAS traffic simulation model have been studied (1). For the purpose of developing warrants, the following left-turn delay criteria are used to define critical conditions for left-turn operations:

1. The average left-turn delay reaches 35 s,
2. The 90th-percentile left-turn delay reaches 73 s,
3. Five percent of left-turners are delayed more than two cycles, and
4. Four left-turners in 1 h are delayed more than two cycles.

Performance curves were developed that related each of these criteria to left-turn and opposing flow volumes by using the TEXAS model. Examples of the curves are presented here as Figures 1 and 2, and they illustrate the relationships for 90th-percentile left-turn delay when opposing flows consist of two and three lanes, respectively. Each plotted point of each performance curve represents the arithmetic mean of eight repetitions of 45 min of simulated observation time.

By applying each of these criteria to its corresponding left-turn performance curve (1), critical left-turn volumes can be determined as shown in Table 1. It can be seen that the criteria of 35 s for the average left-turn delay and 73 s for the 90th-percentile left-turn delay will usually generate the lowest critical left-turn volumes. On the other hand, the criteria of 5 percent of left-

turners delayed more than two cycles and four left-turners in 1 h delayed more than two cycles generally will lead to the highest critical left-turn volumes. Traffic engineers may choose any level between the highest and lowest critical left-turn volumes as the warranted left-turn volume depending on which criterion they regard as more important. The decision regarding a separate left-turn phase can be made as follows: A separate left-turn phase is required if all four delay criteria are met; no separate left-turn phase is needed if none of the four criteria is satisfied. When some but not all of the four delay criteria are satisfied, a judgment is required by the traffic engineer. A typical decision chart is shown in Figure 3.

Figure 2. The 90th-percentile left-turn delay under various traffic conditions at six-by-six signalized intersections with adequate length of bay.

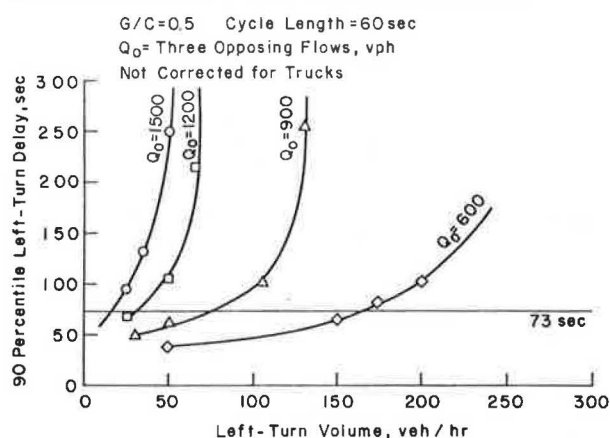


Table 1. Critical left-turn volumes based on different criteria for three types of signalized intersections with adequate length of bay.

Criterion	Opposing Traffic Volume (vph)			
	200	300	400	500
Two-by-Two Signalized Intersection				
Average left-turn delay, 35 s	255	170	90	50
90th-percentile left-turn delay, 73 s	255	170	90	50
Five percent of left-turners delayed more than two cycles	255	195	120	70
Four left-turners in 1 h delayed more than two cycles	260	180	110	70
Ratio of left-turn demand and capacity, 0.7	222	176	128	85
Product of left-turn and opposing volume, 50 000	250	167	125	100
Four-by-Four Signalized Intersection				
Average left-turn delay, 35 s	275	200	155	110
90th-percentile left-turn delay, 73 s	275	195	155	110
Five percent of left-turners delayed more than two cycles	290	220	170	130
Four left-turners in 1 h delayed more than two cycles	275	195	160	120
Ratio of left-turn demand and capacity, 0.7	217	179	153	122
Product of left-turn and opposing volumes, 90 000	300	225	180	150
Six-by-Six Signalized Intersection				
Average left-turn delay, 35 s	165	65	25	15
90th-percentile left-turn delay, 73 s	165	75	30	15
Five percent of left-turners delayed more than two cycles	195	90	40	30
Four left-turners in 1 h delayed more than two cycles	175	75	55	35
Ratio of left-turn demand and capacity, 0.7	147	93	68	45
Product of left-turn and opposing volumes, 110 000	183	122	92	73

Notes: Green per cycle (G/C) = 0.5; C = 60 s. Not corrected for trucks and buses.

Figure 1. The 90th-percentile left-turn delay under various traffic conditions at four-by-four signalized intersections with adequate length of bay.

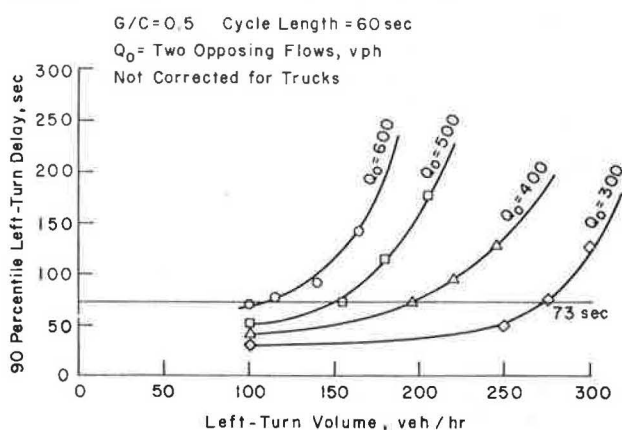
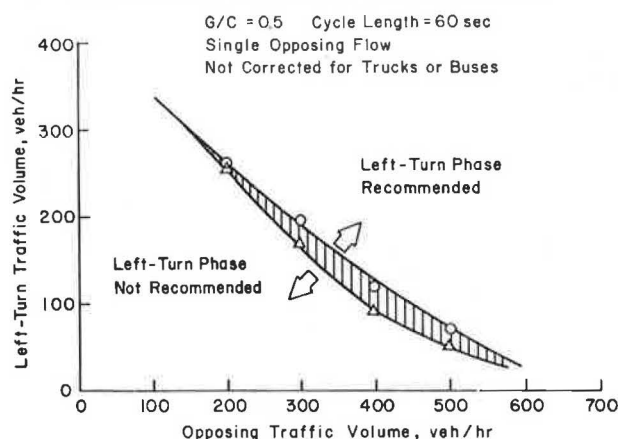


Figure 3. Typical decision chart for implementing left-turn treatment.



With these critical left-turn volumes, Tables 2 and 3 show that neither the volume products nor the volume-to-capacity ratios remain constant over opposing volumes. This is not surprising, since these two types of warrants were found inadequate in the previous discussions. The question is what kind of left-turn warrant could appropriately describe the results shown in Table 1; in other words, what type of left-turn warrant might apply if it is not a volume-capacity ratio or a cross product of volumes. The answer might be revealed by examining the relation between the left-turn capacity and the opposing volume from a different viewpoint.

It was found that the left-turn capacity Q_L can in general be obtained from a linear equation as follows (1):

$$Q_L = Q_C (G/C) - e_0 Q_0 \quad (1)$$

where Q_C and e_0 assume different values over different ranges of opposing volume. Equation 1 can also be written as follows:

$$Q_L + e_0 Q_0 = Q_C (G/C) \quad (2)$$

The physical meaning of Equation 2 can be explained as follows. The coefficient e_0 (1) is the equivalence factor for converting opposing to left-turn vehicles. Thus, the left-hand side of Equation 2 is the sum of total conflicting flows in terms of left-turn vehicles produced by converting opposing to left-turn vehicles by using the equivalence factor e_0 . In this sense, the right-hand side of Equation 2 is the maximum volume of total conflicting flows that can be processed through the signalized intersection or can be regarded as the capacity of the conflict area. It follows that Q_C will be the maximum volume of conflicting flows that can be processed in 1 h of green time, or it can be called the effective capacity of the conflict area. When the capacity of the conflict area is used, opposing vehicles not only have priority over left-turn vehicles but also have a weight less than that of left-turn vehicles.

Note that if Equation 2 is divided by e_0 , it will become

$$Q_L/e_0 + Q_0 = (Q_C/e_0)(G/C) \quad (3)$$

Let $e_L = 1/e_0$ and $Q'_C = Q_C/e_0$. Then Equation 3 will become

$$e_L Q_L + Q_0 = Q'_C (G/C) \quad (4)$$

Table 2. Ratios of critical left-turn volumes to left-turn capacities under different levels of opposing volumes and number of opposing lanes.

No. of Opposing Lanes	Opposing Volume (vph)	Criteria for Determining Critical Left-Turn Volumes			
		Avg Left-Turn Delay, 35 s	90th-Percentile Left-Turn Delay, 73 s	5 Percent of Left-Turners Delayed > Two Cycles	Four Left-Turners in 1 h Delayed > Two Cycles
Single	200	0.80	0.80	0.80	0.82
	300	0.67	0.67	0.77	0.71
	400	0.49	0.49	0.65	0.60
	500	0.41	0.41	0.58	0.58
Two	300	0.87	0.87	0.92	0.87
	400	0.78	0.76	0.86	0.76
	500	0.71	0.71	0.78	0.74
	600	0.69	0.63	0.74	0.69
Three	600	0.79	0.79	0.93	0.83
	900	0.49	0.56	0.68	0.56
	1200	0.26	0.31	0.41	0.57
	1500	0.23	0.23	0.47	0.54

Notes: $G/C = 0.5$; $C = 60$ s. Not corrected for trucks and buses.

Table 3. Cross products of critical left-turn volumes and opposing volumes under different levels of opposing volumes and number of opposing lanes.

No. of Opposing Lanes	Opposing Volume (vph)	Criteria for Determining Critical Left-Turn Volumes			
		Avg Left-Turn Delay, 35 s	90th-Percentile Left-Turn Delay, 73 s	5 Percent of Left-Turners Delayed > Two Cycles	Four Left-Turners in 1 h Delayed > Two Cycles
Single	200	51 000	51 000	51 000	52 000
	300	51 000	51 000	58 500	54 000
	400	36 000	36 000	48 000	44 000
	500	25 000	25 000	35 000	35 000
Two	300	82 500	82 500	87 000	82 500
	400	80 000	78 000	88 000	78 000
	500	77 500	77 500	85 000	80 000
	600	72 000	66 000	78 000	72 000
Three	600	99 000	99 000	117 000	105 000
	900	58 500	67 500	81 000	67 000
	1200	30 000	36 000	48 000	66 000
	1500	22 500	22 500	45 000	52 500

Notes: $G/C = 0.5$; $C = 60$ s. Not corrected for trucks and buses.

Equation 4 has a physical meaning similar to that of Equation 2 except that the total conflicting flows are represented in terms of opposing vehicles by converting left-turn to opposing vehicles with the left-turn equivalence factor e_L . A left-turn equivalence factor of 1.6 has been used in the literature and found suitable for single opposing flow less than 1000 vph in the TEXAS model. However, the left-turn equivalence factor e_L , as will be shown later, is not a constant value for all opposing volumes and geometric configurations.

In order to preclude critical conditions of left-turn operations, left-turn demand or the total conflicting flows should not be near capacity. Let Q_w be a critical left-turn volume at signalized intersections that have adequate length of bay without a separate left-turn phase. Let f_c be the allowable utilization factor of the conflict area, defined as follows:

$$f_c = (Q_w + e_0 Q_0) / Q_C (G/C) \quad (5)$$

Hence, for any critical left-turn volume $Q_w < Q_L$, there exists an allowable utilization factor

of the conflict area $f_c < 1.0$ such that the following equations hold:

$$Q_w + e_0 Q_0 = f_c Q_c (G/C) \quad (6)$$

or

$$Q_w = f_c Q_c (G/C) - e_0 Q_0 \quad (7)$$

As Q_w approaches Q_L , f_c will approach 1.0. In this case, Equation 6 is reduced to Equation 2. If values of e_0 , f_c , and Q_c under various traffic conditions and geometric configurations are known, the critical left-turn volume Q_w can be determined from Equation 7. Therefore, Equation 7 can serve as a left-turn warrant. Typical values of e_L , e_0 , Q_c , and f_c are shown in Table 4. To assist traffic engineers in using their judgement, f_c -values for predicting the lowest and highest critical left-turn volumes are provided. From Table 4, the following conclusions can be drawn:

1. For a given intersection geometry and cycle split, the left-turn equivalence factor [$e_L (= 1/e_0)$], the effective capacity of the conflict area (Q_c), and the allowable utilization factor of the conflict area (f_c) have different values for different ranges of opposing volume.

2. The left-turn equivalence factor varies from 1.6 for low volumes of single opposing flow to 8.9 for high volumes of three opposing flows. Generally, the fewer the number of acceptable gaps, the larger the left-turn equivalence factor will be.

3. The effective capacity of the conflict area varies from 465 to 930 vehicles/green hour. For the same opposing volume, the effective capacity increases with the number of opposing lanes.

4. The allowable utilization factor of the conflict area varies from 0.79 to 0.96. For a given

intersection geometry, the allowable utilization factor of the conflict area decreases as the opposing volume increases.

From Equation 7, the relation between the critical left-turn volume and the left-turn capacity can be obtained as follows:

$$\begin{aligned} Q_w &= f_c Q_c (G/C) - e_0 Q_0 \\ &= [Q_c (G/C) - e_0 Q_0] - [Q_c (G/C) - f_c Q_c (G/C)] \\ &= Q_L - (1 - f_c) Q_c (G/C) \end{aligned} \quad (8)$$

Let

$$M = (1 - f_c) Q_c (G/C) \quad (9)$$

Then

$$Q_w = Q_L - M \quad (10)$$

Equation 10 reveals that the critical left-turn volume is M vehicles less than the left-turn capacity. This implies that there exists a threshold located at M vehicles lower than the left-turn capacity and that once the left-turn demand reaches this threshold, the left-turn operations will become critical. The value of M depends on the geometric configuration, the signal-timing scheme, and the level of the opposing volume. Left-turn warrants for a separate left-turn phase under various traffic conditions and geometric configurations can be obtained from Table 5 by using Table 4. Decision charts for a separate left-turn phase are provided in Figures 4 through 6. If a left-turn demand is greater than the warranted left-turn volume obtained from Table 5 or Figures 4 through 6, the four left-turn delay criteria are all satisfied. Thus, a separate left-turn phase is required.

Compared with simulation results from the TEXAS model, the recommended left-turn warrants in Table 5 predict the highest critical left-turn volume within about 10 vehicles for the case of a 0.5-cycle split and a 60-s cycle length. The volume-product warrant, the volume-capacity-ratio warrant, and the recommended warrant are compared in Figures 7 through 9.

LEFT-TURN BAY

An adequate length of bay has been assumed in studying warrants for a separate left-turn phase. Should a left-turn bay not be adequately long or not be provided at all, left-turn and through vehicles will incur more delay due to interactions among them. Moreover, through vehicles impeded by the left-turn queue may attempt hazardous lane changes. A left-turn bay is always desired; however, the construction of a left-turn bay usually involves redesigning the intersection and thus is costly. Therefore, it is important to know when a left-turn bay is required and how long the bay should be. This section will concentrate on developing warrants for a left-turn bay, and the bay length will be left for discussion in the next section.

For unsignalized intersections, Failmezger (6) and Harmelink (7) proposed a relative warrant and volume warrants, respectively, for the construction of a left-turn bay. The relative warrant is based on an index of hazards, construction costs, and past traffic accident data. If the numerical value of the indicator parameters of the relative warrant is greater than 1, a left-turn bay is recommended. The volume warrants developed by Harmelink are based on queuing-theory analysis and field studies of traffic behavior. If the opposing and left-turn volumes are known, the bay length required can be determined

Table 4. Values of e_L , e_0 , f_c , and Q_c for different opposing volumes and number of opposing lanes.

No. of Opposing Lanes	Opposing Volume Q_0 (vph)	Equivalence Factor		Effective Capacity of Conflict Area Q_c (vehicles/green hour)	Allowable Utilization Factor f_c
		e_L	e_0		
Single	$0 < Q_0 C/G < 1000$	1.6	0.634	879	0.84-0.87
	$1000 < Q_0 C/G < 1350$	2.9	0.348	590	0.79-0.82
Two	$0 < Q_0 C/G < 1000$	2.0	0.500	930	0.86-0.92
	$1000 < Q_0 C/G < 1350$	2.8	0.353	780	0.82-0.87
	$1350 < Q_0 C/G < 2000$	6.0	0.167	465	0.79-0.84
Three	$0 < Q_0 C/G < 1000$	2.2	0.448	930	0.91-0.96
	$1000 < Q_0 C/G < 1350$	3.4	0.297	780	0.88-0.94
	$1350 < Q_0 C/G < 2400$	8.9	0.112	465	0.72-0.84

Table 5. Recommended left-turn warrants for separate left-turn phase under different levels of opposing volumes and number of opposing lanes.

No. of Opposing Lanes	Opposing Volume Q_0 (vph)	Critical Left-Turn Volume Q_w (vph)
Single	$0 < Q_0 C/G < 1000$	$770(G/C) - 0.634Q_0$
	$1000 < Q_0 C/G < 1350$	$480(G/C) - 0.348Q_0$
Two	$0 < Q_0 C/G < 1000$	$855(G/C) - 0.500Q_0$
	$1000 < Q_0 C/G < 1350$	$680(G/C) - 0.353Q_0$
	$1350 < Q_0 C/G < 2000$	$390(G/C) - 0.167Q_0$
Three	$0 < Q_0 C/G < 1000$	$900(G/C) - 0.448Q_0$
	$1000 < Q_0 C/G < 1350$	$735(G/C) - 0.297Q_0$
	$1350 < Q_0 C/G < 2400$	$390(G/C) - 0.112Q_0$

Figure 4. Decision chart for implementing separate left-turn phase at signalized intersections for which $G/C = 0.4$ and $C = 60$ s.

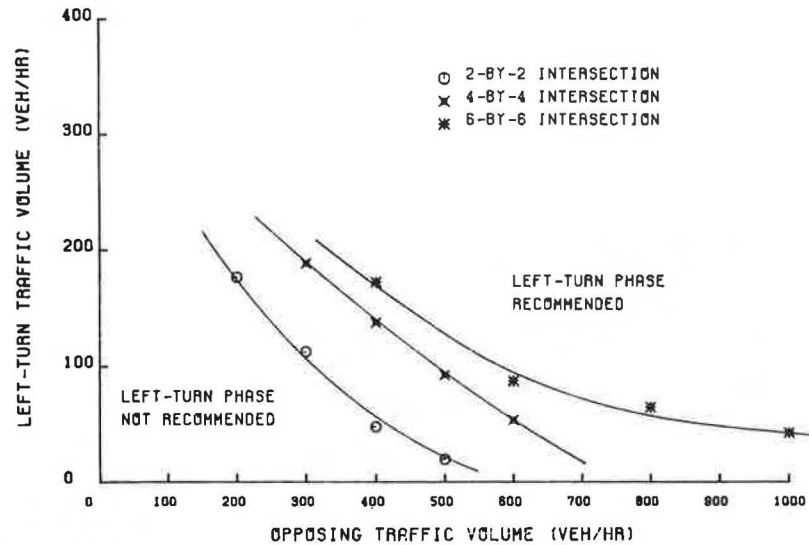


Figure 5. Decision chart for implementing separate left-turn phase at signalized intersections for which $G/C = 0.5$ and $C = 60$ s.

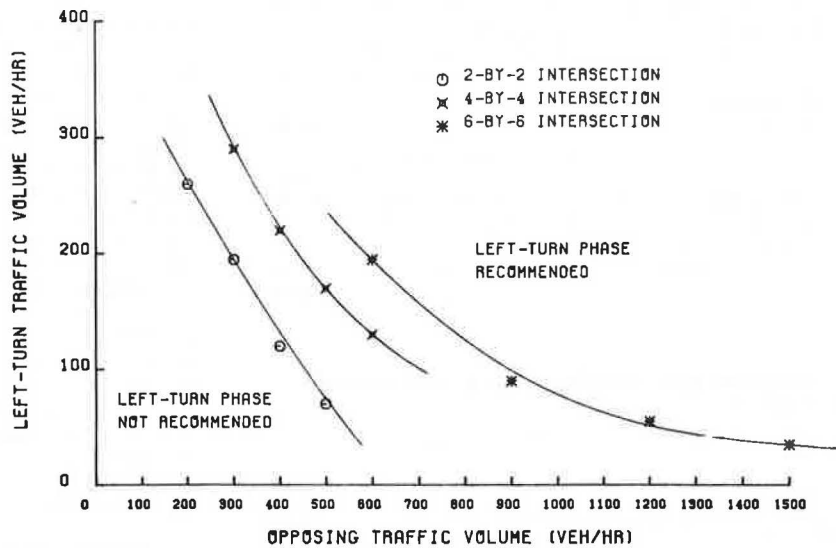


Figure 6. Decision chart for implementing separate left-turn phase at signalized intersections for which $G/C = 0.6$ and $C = 60$ s.

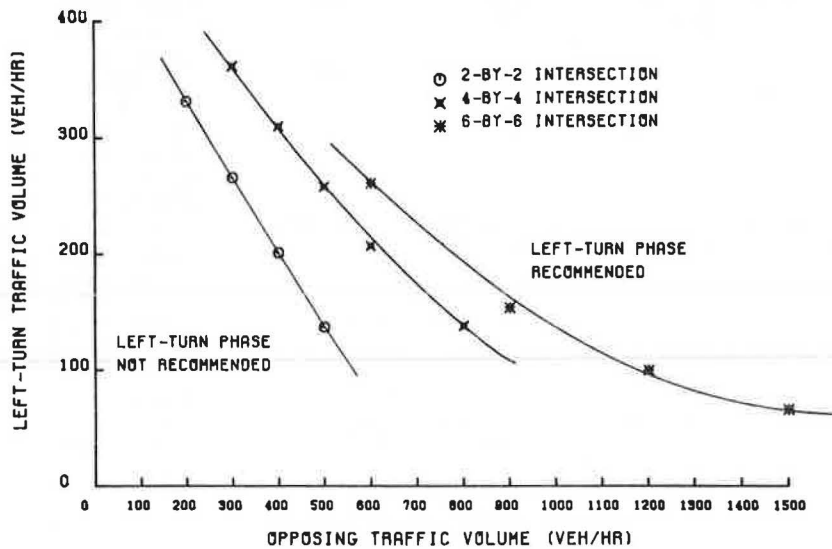


Figure 7. Comparisons among different warrants for separate left-turn phase at two-by-two signalized intersections.

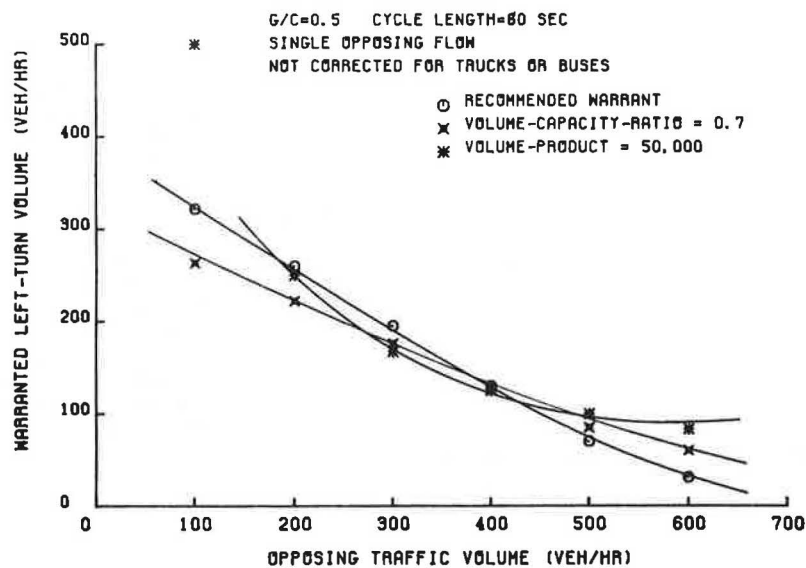


Figure 8. Comparisons among different warrants for separate left-turn phase at four-by-four signalized intersections.

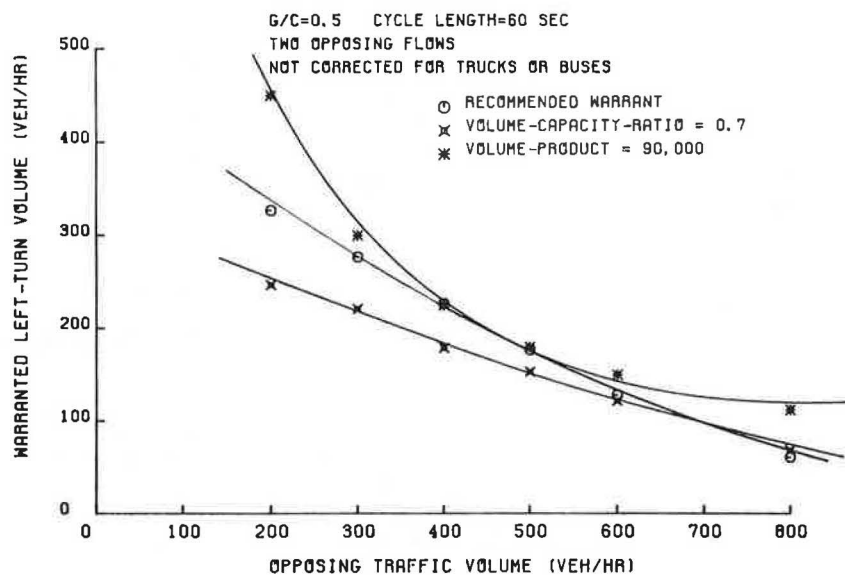


Figure 9. Comparisons among different warrants for separate left-turn phase at six-by-six signalized intersections.

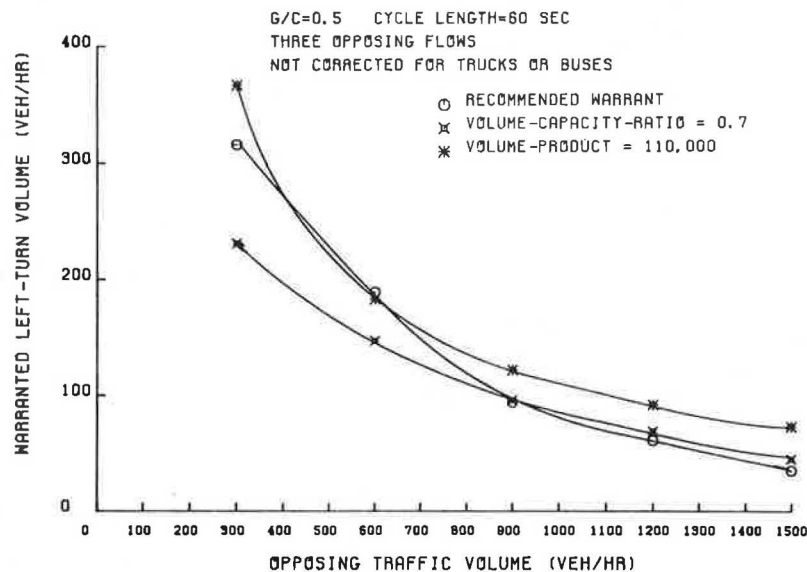


Table 6. Values of \tilde{e}_L , \tilde{e}_0 , \tilde{Q}_c , and \tilde{f}_c for single opposing flow.

Opposing Volume Q_0 (vph)	Through Volume in Median Lane (vph)	\tilde{e}_L	\tilde{e}_0	\tilde{Q}_c	\tilde{f}_c
$0 < Q_0C/G < 1000$	100	1.6	0.634	855	0.84-0.87
	200	1.7	0.593	820	0.84-0.87
	300	1.9	0.526	680	0.84-0.87
	400	2.2	0.455	560	0.84-0.87
$0 < Q_0C/G < 800$	500	2.9	0.340	415	0.84-0.87
$1000 < Q_0C/G < 1350$	100	3.2	0.310	530	0.79-0.82
	200	3.7	0.270	460	0.79-0.82
	300	4.5	0.220	375	0.79-0.82
	400	5.6	0.180	300	0.79-0.82
$800 < Q_0C/G < 1350$	500	4.0	0.250	295	0.79-0.82

Table 7. Values of \tilde{e}_L , \tilde{e}_0 , \tilde{Q}_c , and \tilde{f}_c for two opposing flows.

Opposing Volume Q_0 (vph)	Through Volume in Median Lane (vph)	\tilde{e}_L	\tilde{e}_0	\tilde{Q}_c	\tilde{f}_c
$0 < Q_0C/G < 1000$	100	2.0	0.507	910	0.86-0.92
	200	2.1	0.483	840	0.86-0.92
	300	2.3	0.443	740	0.86-0.92
	400	2.6	0.380	615	0.86-0.92
$0 < Q_0C/G < 800$	500	3.3	0.305	455	0.86-0.92
$1000 < Q_0C/G < 1600$	100	2.7	0.370	770	0.82-0.87
	200	2.9	0.340	695	0.82-0.87
	300	3.4	0.290	590	0.82-0.87
	400	4.4	0.230	465	0.82-0.87
$800 < Q_0C/G < 1600$	500	5.3	0.188	365	0.82-0.87
$1600 < Q_0C/G < 2000$	100	6.3	0.160	435	0.79-0.84
	200	7.1	0.140	375	0.79-0.84
	300	8.7	0.115	310	0.79-0.84
	400	11.1	0.090	240	0.79-0.84
	500	16.7	0.060	160	0.79-0.84

Table 8. Values of \tilde{e}_L , \tilde{e}_0 , \tilde{Q}_c , and \tilde{f}_c for three opposing flows.

Opposing Volume Q_0 (vph)	Through Volume in Median Lane (vph)	\tilde{e}_L	\tilde{e}_0	\tilde{Q}_c	\tilde{f}_c
$0 < Q_0C/G < 1000$	100	2.2	0.450	910	0.91-0.96
	200	2.3	0.430	840	0.91-0.96
	300	2.5	0.400	745	0.91-0.96
	400	2.9	0.343	615	0.91-0.96
$0 < Q_0C/G < 800$	500	3.6	0.280	460	0.91-0.96
$1000 < Q_0C/G < 1600$	100	3.2	0.317	775	0.88-0.94
	200	3.4	0.297	705	0.88-0.94
	300	3.9	0.260	605	0.88-0.94
	400	4.8	0.210	485	0.88-0.94
$800 < Q_0C/G < 1600$	500	5.8	0.173	375	0.88-0.94
$1600 < Q_0C/G < 2000$	100	9.1	0.110	445	0.72-0.84
	200	10.0	0.100	395	0.72-0.84
	300	11.1	0.090	335	0.72-0.84
	400	14.3	0.070	260	0.72-0.84
	500	20.0	0.050	105	0.72-0.84

from charts provided. As to signalized intersections, Dart (8) performed a computer simulation to develop warrants for a left-turn bay. If delay is used as a design criterion, the need for a bay can be ascertained. In this section, warrants for a left-turn bay will be explored from its relation to left-turn capacity.

Before warrants for a left-turn bay are developed, criteria for defining critical conditions when there is no bay must be chosen. The four left-turn delay criteria used in developing warrants for a

separate left-turn phase seem to remain relevant in this case. However, the through delay in the median lane should also be an important concern since through vehicles in the median lane will be impeded by left-turning vehicles if there is no bay. The question is what an appropriate through delay criterion would be. It has been found (1) that the average through delay in the median lane is not considerably greater than that in the curb lane so long as no more than 5 percent of the left-turners are delayed more than two cycles. In view of this, the four left-turn delay criteria alone would be appropriate for developing warrants for a left-turn bay.

Since the same criteria are used, warrants for a left-turn bay can be derived through an approach similar to that for a separate left-turn phase. For the convenience of discussion, left-turning vehicles in the opposing flows are ignored first and then taken into consideration later.

Case 1: No Left-Turning Vehicles in Opposing Flows

If we refer to Lin's study (1), the left-turn capacity for the no-bay case when there are no left-turning vehicles in opposing flows in general can be obtained as follows:

$$\tilde{Q}_L = \tilde{Q}_c (G/C) - \tilde{e}_0 Q_0 \quad (11)$$

By the same argument as in the previous section, warrants for a left-turn bay can be expressed as follows:

$$\tilde{Q}_w = \tilde{Q}_L - (1 - \tilde{f}_c) \tilde{Q}_c (G/C) \quad (12)$$

Typical values of \tilde{e}_L , \tilde{e}_0 , \tilde{Q}_c , and \tilde{f}_c are summarized in Tables 6 through 8.

Case 2: Left-Turning Vehicles in Opposing Flows

The left-turn capacity when there is no bay and there are V_{0L} and Q_0 left-turning and through vehicles per hour in opposing flows will be as follows (1):

$$Q_L = \tilde{Q}_L - aQ_0 \quad (13)$$

where

\hat{Q}_L = left-turn capacity with no bay when there are V_{0L} left-turning vehicles per hour in opposing flows,

\tilde{Q}_L = left-turn capacity with no bay when there are no left-turning vehicles in opposing flows as defined in Equation 11,

$a = 0.317 (P_C - 1.0/N)$,

P_C = percentage of opposing traffic in curb lane, and

N = number of opposing lanes.

Thus, the warrant for a left-turn bay when there are left-turning vehicles in opposing flows can be obtained as follows:

$$\tilde{Q}_w = \tilde{Q}_w - aQ_0 \quad (14)$$

Since the warranted left-turn volume Q_w can be obtained from Equation 12, the left-turn volume Q_w required for construction of a bay when there are left-turning vehicles in opposing flows can be determined from Equation 14.

REQUIRED LENGTH OF LEFT-TURN BAY

Once a decision has been made regarding the con-

Figure 10. Maximum number of left-turn vehicles stored in bay under various traffic conditions at two-by-two signalized intersections.

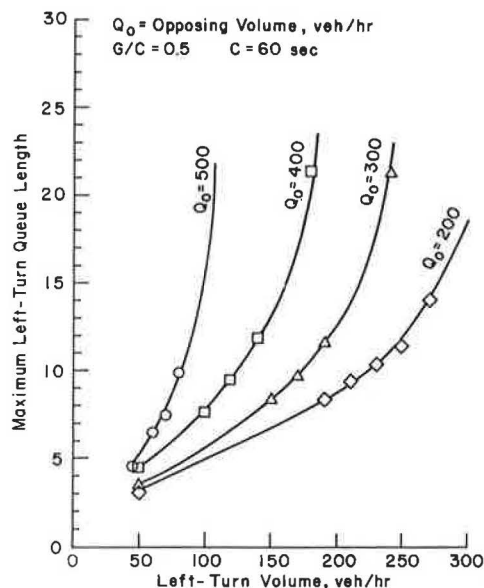
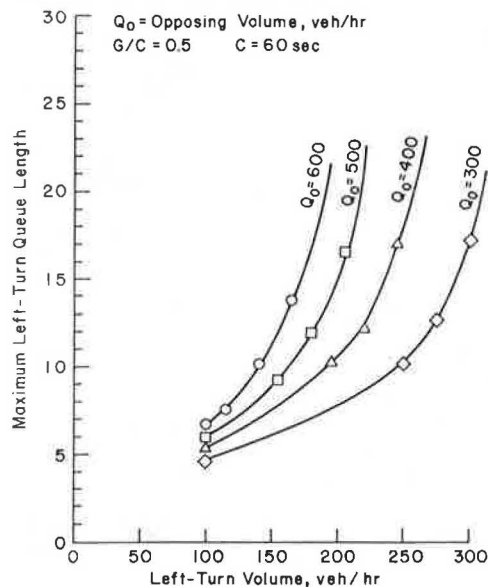
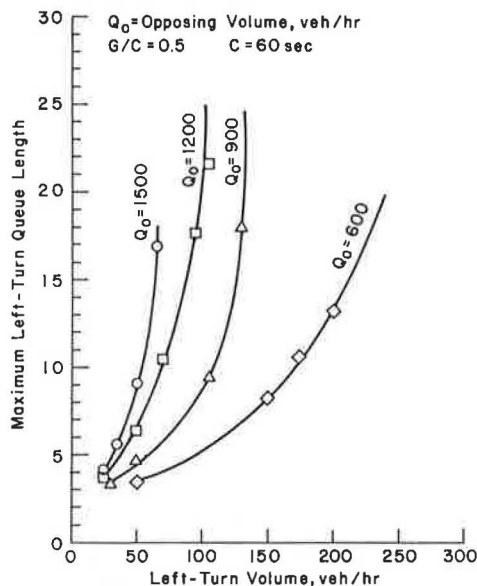


Figure 11. Maximum number of left-turn vehicles stored in bay under various traffic conditions at four-by-four signalized intersections.



struction of a left-turn bay at a signalized intersection, the next step would be to determine how long the left-turn bay should be. The American Association of State Highway and Transportation Officials (AASHTO) (9) states that storage length should be 1.5-2.0 times the average number of vehicles that would be stored per cycle based on design volume. Unfortunately, this guideline may not clearly recognize the fact that the average number of left-turning vehicles stored per cycle will depend on the opposing volume and the signal-timing scheme. For the same left-turn demand, the number of left-turning vehicles stored in the bay for high opposing volume will be much larger than that for low opposing volume. Messer (10) used a combination of theory and traffic simulation to develop the relation between left-turning volume and left-turn

Figure 12. Maximum number of left-turn vehicles stored in bay under various traffic conditions at six-by-six signalized intersections.



bay length required for a protected left-turning movement. In this section, the bay length required for an unprotected left-turn movement will be derived based on the simulation results from the TEXAS model.

From the study of left-turn queuing (1), it was found that the relationship between the average and the maximum values of left-turn queue length can be approximately represented by the following equations:

Based on the average condition:

$$L_m = 5.5\bar{L}^{0.58} \quad (R^2 = 0.95) \quad (15)$$

Based on 95 percent confidence level:

$$L_m = 7.4\bar{L}^{0.55} \quad (R^2 = 0.86) \quad (16)$$

where L_m is the maximum left-turn queue length in vehicles and \bar{L} is the average left-turn queue length in vehicles.

If the bay length is designed based on the average condition, the bay length will be exceeded under a given traffic condition with a probability of 0.5. On the other hand, if the bay length is designed based on the 95 percent confidence level, the bay length will be exceeded with a probability of 0.05. Any bay length in between will have a probability greater than 0.05 but less than 0.5 of being exceeded.

On the assumption that a passenger car and a truck or a bus will occupy w_c ft and w_T ft of bay length, respectively, the required bay length can be determined from the following equation:

$$L_B = w_T p_T L_m + w_c (1 - p_T) L_m \quad (17)$$

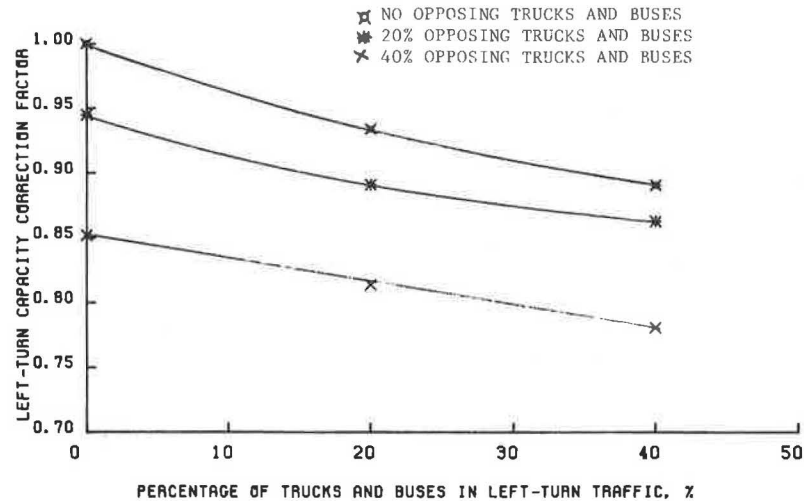
where p_T is the percentage of trucks in the left-turning traffic flow (decimal).

Figures 10 through 12 are charts for determining the required bay length based on the 95 percent confidence level.

CORRECTIONS FOR TRUCKS AND BUSES

So far, it has been assumed that the traffic population consists of passenger cars only. For traffic

Figure 13. Factors for adjusting left-turn capacity for different combinations of opposing and left-turn truck percentages.



flows in which passenger cars are mixed with trucks and buses, the left-turn warrants obtained in the previous sections have to be modified. From Lin's study (1), the left-turn capacity for mixed traffic flows can be obtained by adjusting the "truck-free" capacity as follows:

$$Q_L^* = f_T Q_L \quad (18a)$$

where

- Q_L^* = left-turn capacity for mixed traffic flows (vph),
- Q_L = left-turn capacity for traffic without trucks and buses (vph), and
- f_T = correction factor for trucks and buses obtained from Figure 13.

Therefore, the left-turn warrant for mixed traffic will be

$$Q_w^* = Q_L^* - M \quad (18)$$

DISCUSSION

Although the four left-turn delay criteria adopted in this study have been suggested by researchers and practicing engineers, it is recognized that different criteria and methodologies might bring out different left-turn warrants. It seems appealing to have simplified left-turn warrants such as constant volume-capacity ratios or cross products of volumes. However, simulation results from the TEXAS model show little evidence of such simple relations. Alternatively, this study reveals a new type of capacity warrant. The warranted left-turn volume is at some margin from the left-turn capacity, whereas the margin may have different constant values over different ranges of opposing volume. This type of left-turn warrant, though more complicated, is more reasonable. It is hard to believe that complicated left-turn operations can be characterized by a single numerical value with reasonable accuracy, especially over a wide range of traffic conditions.

Another important problem would be how to establish the phasing plan for a signal that has a separate left-turn phase. As in the case of no left-turn bay, the required bay length must be known once a left-turn bay is warranted. In fact, to know how to implement a left-turn treatment effectively is more important than to know when to implement it. In many field studies, it has been found that the

left-turn delay is increased after a separate left-turn phase has been implemented. This might happen when the left-turn phase is not really justified or, more likely, the left-turn phase is not properly designed. Hence, it is necessary to have guidelines for phasing the left-turn signal.

The Texas Transportation Institute (5) provided guidelines for choosing a phasing scheme, such as leading, lagging, or skipping (actuated) left-turn phase for a single left-turn movement. However, how to determine the cycle length and duration of a left-turn phase is not quite clear. A simple guideline for timing the left-turn phase might be as follows: The total time available for left turns in 1 h after addition of a separate left-turn phase should not be less than that before it was added. For example, for an opposing traffic flow of 400 vph at a two-by-two signalized intersection, the transparency [ratio of accepted gap time to total time (1)] is 0.22. That means the total time available for left turns in 1 h is 792 s. Thus, the total time for the separate left-turn phase, either pretimed or actuated, should not be less than 792 s. Otherwise, the left-turn delay would be increased after a separate left-turn phase had been implemented. When left turns are prohibited during the through green phase, the duration of the left-turn phase should be at least 11 s if a cycle length of 70 s is used. For protective or permissive left turns, the duration of the left-turn phase can be less than 11 s since some part of the green phase for opposing traffic is used by left-turners. As to an actuated left-turn phase, the maximum extension for the phase can be set at 11 s. When the left-turn demand is saturated, the actuated signal would perform as pretimed so that the total available time for left turns in 1 h would be 792 s. Moreover, for the pretimed signal, whether the left-turn phase should be added to or taken out of the original cycle length is not a trivial question. Since the cycle split for opposing flows is critical to the left-turn delay, a rule of thumb would be to keep the cycle split as near the original cycle split as possible.

ACKNOWLEDGMENT

This paper was prepared in cooperation with the Federal Highway Administration, U.S. Department of Transportation. The contents reflect our views, and we are responsible for the facts and the accuracy of the data presented. The contents do not necessarily

reflect the official views or policies of the Federal Highway Administration. This paper does not constitute a standard, specification, or regulation.

REFERENCES

1. H.-J. Lin. A Simulation Study of Left-Turn Operations at Signalized Intersections. Univ. of Texas at Austin, dissertation, Aug. 1982.
2. K.R. Agent and R.C. Deen. Warrants for Left-Turn Signal Phasing. Department of Transportation, Commonwealth of Kentucky, Frankfort, Res. Rept. 505, Oct. 1978.
3. F.L. Orcutt. Primer for Traffic Selection. Public Works, Vol. 106, No. 3, March 1975, pp. 76-80.
4. G. Gurnett. Intersection Delay and Left Turn Phasing. Traffic Engineering, Vol. 19, No. 7, June 1969, pp. 50-53.
5. Guidelines for Signalized Left-Turn Treatments. Texas Transportation Institute, Univ. of Texas at Austin; FHWA, July 1981.
6. R.W. Failmezger. Relative Warrants for Left-Turn Refuge Construction. Traffic Engineering, Vol. 13, April 1963, pp. 18-20.
7. M.D. Harmelink. Volume Warrants for Left-Turn Storage Lanes at Unsignalized Grade Intersections. HRB, Highway Research Record 211, 1967, pp. 1-18.
8. O.K. Dart, Jr. Development of Factual Warrants for Left-Turn Channelization Through Digital Computer Simulation. Texas A&M Univ., College Station, Ph.D. dissertation, Aug. 1982.
9. Policy on Design of Urban Highways and Arterial Streets. AASHTO, Washington, DC, 1973.
10. C.J. Messer and D.B. Fambro. Effects of Signal Phasing and Length of Left-Turn Bay on Capacity. TRB, Transportation Research Record 644, 1977, pp. 95-101.

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

Determining Capacity and Selecting Appropriate Type of Control at One-Lane Two-Way Construction Sites

PANOS G. MICHALOPOULOS AND ROGER PLUM

The problem of determining the most appropriate type of traffic control at one-lane two-way construction sites (i.e., on two-lane two-way roadways where one lane is temporarily closed for repairs and the other must be shared by both directions of traffic) is addressed. Capacity and performance tables and figures are presented for stop-sign, signal, or flagger control. These were developed by a microscopic simulation program that was adjusted and calibrated from field data. Following safety and visibility constraints, selection of the most appropriate control type can be made from the capacity and performance estimations obtained from the methodology presented here along with some practical considerations. An overview of existing practices followed by most states is also presented.

Traffic control at construction and maintenance zones has become particularly important in recent years, especially in view of increased government liability for accidents and incidents on public road systems and the reduced tolerance for inefficient operating conditions. Despite the attention recently given to the development of design standards and improved traffic control at construction and maintenance zones, few guidelines are currently available for determining capacity and the most appropriate type of control at two-lane two-way roadways where one lane is temporarily closed and the other must be shared by both directions of traffic. Such is frequently the case for two-lane two-way bridges during deck repairs.

The Minnesota Department of Transportation, recognizing the need for further research in this area, sponsored a research project to (a) determine capacity and select the most appropriate type of control (including optimal timing plans in the case of signal control) and (b) develop guidelines for increasing safety by appropriate signing. In this paper only the first issue is addressed, but it should be noted that all the results of this study are described in a final report (1), which may be

consulted for further details not included here due to space limitations.

Selection of the most appropriate type of control (i.e., among stop sign, pretimed or actuated signal, and flagger) requires performance evaluation of each alternative, which in turn is dependent on capacity estimations. An extensive literature search combined with a survey of practices in all states revealed the absence of any well-established methodology for dealing with problems of capacity, performance evaluation, and selection of the most appropriate type of control. For this reason, a more systematic procedure was developed and is described here. It should be kept in mind that although the basic research was geared toward one-lane bridges during the construction or maintenance operations, the results are general and apply to any similar situation in which a single lane is alternately used by both directions of travel.

The problems of capacity determination and performance evaluation with stop-sign control were resolved by generating tables based on simulation, whereas the signal-control case (pretimed or actuated) was treated both analytically and by simulation. Finally, flagger control was assumed to be similar to actuated-signal control, at least from the capacity and performance points of view (i.e., excluding visibility and safety aspects), and therefore it was not treated separately. Naturally this assumption is only an approximation, but in view of the difficulties involved in realistically modeling flagger control, it was felt that such approximation should suffice. It should be pointed out that the simulation programs and their results were tested against actual data collected at 15 sites by time-lapse photography, and model calibrations and adjustments were made. Similar comparisons were also made with the analytical results.

Table 1. Observed capacities for various expressway lane-closure conditions and work activities.

Work Activity	No. of Lanes in One Direction		Observed Capacity (vph)
	In Normal Operation	In Work Area	
Median barrier or guard-rail repair	2	1	1500
	3 or 4	2	3200
	4	3	4800
Pavement repair, mud-jacking, pavement grooving	2	1	1400
	3 or 4	2	3000
	4	3	4500
Striping, resurfacing, slide removal	2	1	1200
	3 or 4	2	2600
	4	3	4000
Installation of pavement markers	2	1	1100
	3 or 4	2	2400
	4	3	3600
Middle lanes (for any reason)	3 or 4	2	2200
	4	3	3400

BACKGROUND

Although considerable progress has been made in recent years in improving traffic operations at construction zones, very little is known about the capacity of the one-lane situation, described earlier; this information is needed for selecting the most appropriate type of control. In this section, only the existing literature on capacity considerations is reviewed. Existing signal-timing practices are omitted due to space limitations; pertinent information on this subject can be found elsewhere (2,3).

The factors that affect roadway capacity can be divided into two categories: roadway and environmental factors and traffic factors. In the first category one may include items such as lane width, lateral clearance, horizontal and vertical alignment, pavement conditions, and weather conditions. Traffic factors in general include the percentage of large vehicles (trucks, buses, etc.), fluctuation of the demand, conflicts, lane distribution, flow interruptions, etc. Although in current literature most of these factors have been taken into account for surface streets and highways during normal operations, only a few were considered at construction sites. Table 1 (2,4) shows observed capacities of expressways with lane closures. The table suggests single-lane capacity ranging from 1100 to 1500 vehicles per hour (vph). The Highway Capacity Manual suggests that maximum capacity per lane at construction zones is 1500 vehicles/h. These figures refer to saturation-flow estimates; i.e., they assume uninterrupted flow. Experience in Europe (3) indicates lower values, which vary slightly among countries. The guidelines developed in Germany appear to be the most conservative and take speed into account. According to these guidelines, the per-lane saturation flow in a construction area is 1200 vph if average speed in the construction zone is greater than 30 mph. The above figure decreases by 5 percent when average speed drops to 25 mph and by 20 percent for speeds less than 20 mph.

The figures given above assume only one-way operation, and although they might seem unrelated to alternate use of the same lane by both directions of travel, they are in fact needed for estimating capacity when signal control is imposed. Before this section is concluded, it is noted that, to the best of our knowledge, no additional information is available in the literature concerning the capacity of the situation in question; i.e., no information

was found for determining capacity at one-lane two-way construction sites. The only exception is the case of pretimed signal control, which is partly covered by the guidelines of the Organization for Economic Cooperation and Development (OECD) (3). Even in this case, however, the capacity estimations are rather crude.

CAPACITY ESTIMATION

Capacity is the most crucial measure for determining the best type of control. This is because in most practical applications the simplest type of control is sought, provided that it can handle the demand without causing serious disruptions to adjacent intersections or resulting in excessive delays to the users. Furthermore, as will be seen in subsequent sections, knowledge of capacity is essential for estimating the performance of each type of control. In addition to the type of control, capacity is affected by geometric factors (including visibility), traffic factors, environmental conditions, speed, and length of the construction zone. The geometric factors (excluding length) can be taken into account by appropriate estimation of saturation flow (by field measurement, earlier experience, or from the literature presented in the previous section). Traffic factors can be considered by converting demands to passenger-car units (PCUs) and by incorporating the appropriate peak-hour factor or probability distributions. Finally, speed and length of the construction zone can be considered by measuring or estimating the traverse (or crossing) time (i.e., the time required to travel the entire construction zone). In the absence of empirical data, the report by Michalopoulos and others (1) may be consulted for estimating the average crossing time analytically.

As mentioned earlier, capacity with signal control can be estimated analytically; however, with stop-sign control, best solutions can only be given by either extensive data collection or simulation. In the absence of extensive data, the second approach was followed in this study. The program prepared for simulating one-lane construction sites and employed for determining capacity and evaluating control performance is described in the following sections.

For the problem of estimating saturation flow, Table 1 can be used as a guideline in the absence of field data. Based on this and the suggestion of the OECD report (3), a value of 1200 vph is recommended as more realistic than that recommended in the previous section, especially at one-lane bridges. This figure is rather conservative and could easily be exceeded under ideal conditions or it could vary according to the type of work activity as Table 1 suggests. It is also stressed that conversion of the demands to PCUs is needed for signal timing and estimating the performance of the various control alternatives. A number of suggestions are available in the literature, but in the absence of personal experience the following factors are suggested (3):

1. Trucks or buses with three axles or more, 2.25 PCU;
2. Two-axle trucks, 2.00 PCU; and
3. Motorcycles, 0.5 PCU.

These values can be increased or decreased by 3 percent for each 1 percent downhill or uphill grade accordingly.

ONELANE Computer Simulation Program

Analytical techniques for determining capacities and

measures of effectiveness at one-lane construction sites generally fail to account for the randomness of arrivals and the variability of travel times in the construction area. Consequently, computer simulation of traffic by using a random-arrival generator and a probability distribution for travel times was developed for evaluating capacities at such locations.

Previously existing computer programs that simulate traffic controlled either by stop signs or by traffic signals were designed to accommodate standard intersections, where two or more roadways intersect. Initially, employment of existing programs was attempted by proper data manipulation (e.g., by employing long clearance and extension intervals or assuming dummy phases) in order to avoid costly development of new software. However, after some experimentation two of the most widely known simulation programs [NETSIM (5) and TRAFLO (6)] proved not to be adaptable to the case under consideration. This inability to be adapted was primarily caused by the longer clearance intervals required at construction sites and/or the particular characteristics of such areas. For instance, the case of a one-lane site controlled by stop signs, where vehicles are discharged one or two at a time from alternating approaches, did not conform with either of the existing programs. Thus, a new simulation program called ONELANE was developed specifically to accommodate the particular characteristics of one-lane sites.

Following are some of the general input requirements of the ONELANE program:

1. Type of control--stop-sign, pretimed, or actuated signal control--must be specified.
2. Demand on each approach must be converted to PCUs per hour.
3. Saturation flow rate of the common lane during construction is required in passenger cars per hour.
4. Average lost time (during transitions) to reach saturation flow is given in seconds.
5. Mean traverse interval in seconds is derived empirically or by using the methods described earlier.
6. SDs of the actual traverse intervals are calculated. It is assumed in the program that actual traverse intervals are distributed normally about the mean.
7. If stop-sign control is being simulated, a maximum platoon size is chosen. This is the maximum number of vehicles that traverse the construction site at a time as a group. Under ideal conditions only one car should be crossing at a time. However, it was observed that this condition is not always met; i.e., two or more cars could be departing from the same approach as soon as it receives priority. Thus, one must measure or assume the maximum number of cars that are likely to depart at a time. The program allows maximum platoon size to depart if the queue on the approach that has priority is long enough; otherwise, fewer cars are released according to the actual demand when the approach under consideration receives priority and traffic starts moving.
8. If signal control has been selected, the user has the option of specifying the traffic signal settings or allowing the program to determine the optimum settings as described in the next section. For pretimed control, the settings include cycle length and green, amber, and all-red times for both approaches. For actuated signal control, the settings include minimum and maximum green times, amber, and all-red times as well as extension intervals.

ONELANE is a microscopic simulation program;

i.e., each vehicle is followed from the time it arrives on an approach until the time it departs the single-lane portion of the construction area. Arrivals are assumed to be random; a minimum headway is determined by the user-specified saturation flow rate and the average arrival headway is determined by the demand on each approach.

During the simulation, vehicles are discharged from an approach only if opposing vehicles have cleared the one-lane portion of the construction site. With stop-sign control, if vehicles are waiting on both approaches, it is assumed that platoons passing through the one-lane section will alternate between approaches. If vehicles are waiting on just one approach, platoons from that approach will discharge consecutively until a vehicle arrives on the opposing approach. With signal control, it is assumed that no vehicles violate the signal by departing the stop line while a red signal indication is displayed.

As each vehicle is discharged from the stop line, the delay and the number of stops encountered by the vehicle are calculated along with the current queue size (in vehicles) on that approach. Actual traverse time for each vehicle is obtained from a normal distribution of clearance intervals with the constraint that a faster vehicle cannot pass a slower one.

As the simulation progresses, the average delays and number of stops on each approach are constantly updated. At the end of the simulation, these values as well as the maximum queue size and the average energy consumed on each approach are printed along with the number of cars serviced and the final queue size. Before we conclude, it should be noted that the ONELANE program was tested against the field data collected and calibrations were made accordingly. The details of testing and calibration have been presented elsewhere (1).

Estimation of Capacity with Signal Control

Once the saturation flow and the optimal signal settings have been determined, the capacity on each approach can be estimated from

$$Q = s(g/c) \quad (1)$$

where

Q = hourly capacity flow rate [passenger cars per hour (pcph)],
 s = one-lane saturation flow rate (PCUs/h),
 g = effective green interval of approach being considered (s), and
 c = cycle length (s).

The cycle length is the sum of the green, amber, and all-red times on both approaches. Effective green interval g can be determined from

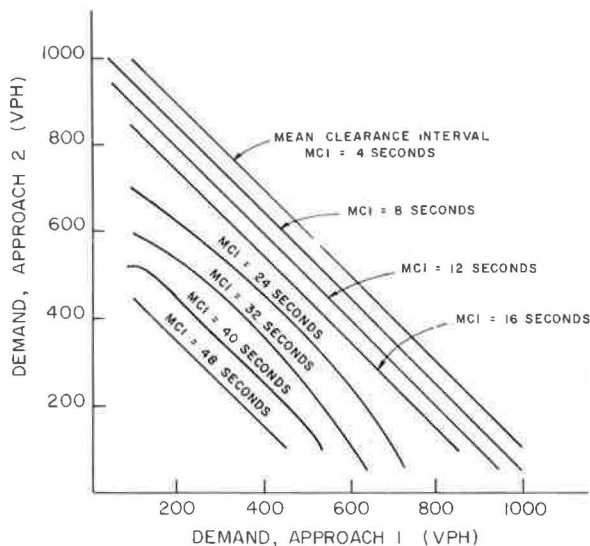
$$g = G + A - LT \quad (2)$$

where

G = actual green time (s),
 A = amber time (s), and
 LT = lost time (s).

It should be noted that this method for determining capacity is valid for both pretimed and traffic-actuated signal systems. However, in the case of traffic-actuated control, the equations must be modified slightly. Since in actuated signals capacity is achieved when both the cycle length and the green time are maximized, maximum cycle length

Figure 1. Capacity of signal control at one-lane construction sites.



(C_{max}) and maximum green time (G_{max}) should be substituted for c and G in the equations. These maxima can be determined by using the methods outlined in the next section.

Since the above method of calculating capacity is deterministic and does not take the variability of arrivals and travel times into account, the capacity of one-lane construction zones with signal control was also estimated and tabulated by using the ONE-LANE simulation program. This was accomplished by using the following approach: increments of 4 s were employed for clearances (travel times) from 4 to 48 s. For each clearance interval, a range of volumes from 100 to 1200 pcph, in increments of 100 pcph, was assumed on each approach. Each combination of demands and clearance interval was simulated for 1 h 10 times to dampen the effect of any single extraordinary simulation. The results from the 10 simulations were then averaged to arrive at a reliable estimate of each measure for that combination of clearance interval and demands. Demands were judged to exceed capacity if the actual volume serviced on either approach was less than 95 percent of the demand on that approach, which indicated a final queue size of at least 5 percent of demand. It should be noted that signal timing for each set of volume combinations was determined first (i.e., prior to simulation) by using the methodology of the next section.

Curves were developed from the simulations for each mean clearance interval. These curves were combined in Figure 1, which can be used as a guideline. It should be noted that a saturation flow of 1200 pcph and a total lost time (excluding all-red) per cycle of 7.4 s (3.7 s per approach) were used in all simulations. The latter as well as the variance in travel times were derived from the collected data.

A single set of curves is presented in Figure 1 despite the fact that two types of signal control, pretimed and actuated, were simulated. This is because the capacity results from the two types of control were so similar in most cases as to be nearly indistinguishable. This is not an unexpected result, since at volumes greater than or equal to capacity, actuated signals operate as pretimed, and if properly timed, they should yield the same capacity. Consequently, one set of results effectively serves for both methods of control.

To use the curves in Figure 1, one simply enters from the bottom of the figure with the demand on one

approach and from the left side of the figure with the demand on the opposing approach. One then proceeds vertically from the bottom and horizontally from the left to the intersection of the two demands, which indicates the maximum allowable mean clearance interval (MCI) by interpolation between the curves for various MCIs. If the actual (or estimated) MCI at the one-lane site exceeds the value found in the figure, signal capacity is exceeded. An actual MCI less than or equal to that found from Figure 1 indicates that traffic-signal control is capable of accommodating the demands without significant delays or queueing.

A final consideration in evaluating capacity estimates for traffic-signal control is determining how well the analytical methods described at the beginning of this section approximate the simulation results. If we assume a fifty-fifty split (i.e., equal demands on both approaches), a mean clearance interval of 12 s, and lost times equal to those indicated earlier (3.7 s/phase/cycle), effective green time (with the maximum cycle length) would be 71.3 s on each approach. This assumes an amber time of 3 s and an all-red time of 9 s to accommodate the 12-s clearance interval. By using Equation 1, the capacity per approach would be $(1200 \times 71.3)/168 = 509$ pcph. As can be seen in Figure 1, this analytical result compares very favorably with the results of the simulations. Further comparisons between the analytical approach and points from different curves confirmed the consistency between the two capacity-determination methods; i.e., inclusion of random effects did not (on the average) alter significantly the results of Equation 1.

Estimation of Capacity with Stop-Sign Control

As mentioned earlier, no reliable analytical methods existed for determining the capacity of one-lane construction sites controlled by stop signs. For this reason, the use of the simulation program proved to be essential.

By proceeding as in evaluating the capacity under signal-controlled conditions, a range of demands on each approach was simulated for MCIs in increments of 4 s. In addition, maximum platoon sizes of one to five vehicles were tested. A maximum platoon size of one vehicle means that every vehicle discharged comes to a complete stop at the stop line. A maximum platoon size of two vehicles means that in a queue of two or more vehicles, the first vehicle in line will stop at the stop line and then depart, whereas the second vehicle in the queue will follow the first without coming to a complete stop at the stop line. Then it is assumed that the next vehicle in line (if any) will stop. A maximum platoon size of three vehicles implies that up to two vehicles will follow the first in the platoon without coming to a full stop at the stop line, etc.

The simulation of each combination of MCI, maximum platoon size, and demands was repeated 10 times to avoid random noise possibly resulting from a single simulation. Any case in which the average actual volume for the 10 simulation runs was less than 95 percent of the demand on either approach was judged to be operating under oversaturated conditions; i.e., the combination of demands exceeded the capacity of the site.

As was done for traffic-signal control in the previous section, a series of curves was constructed that represent the demand limits that can be accommodated by using stop-sign control (1). Depicted in Figure 2 is the set of capacity curves found for a maximum platoon size of two vehicles and a saturation flow of 1200 vph. Capacity from this figure is simply found as before (Figure 1) or by entering

Figure 2. Capacity of stop-sign control at one-lane construction sites: maximum platoon size of two vehicles, $s = 1200$ vph.

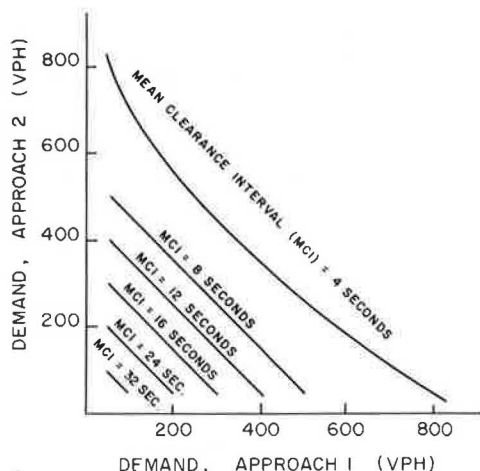
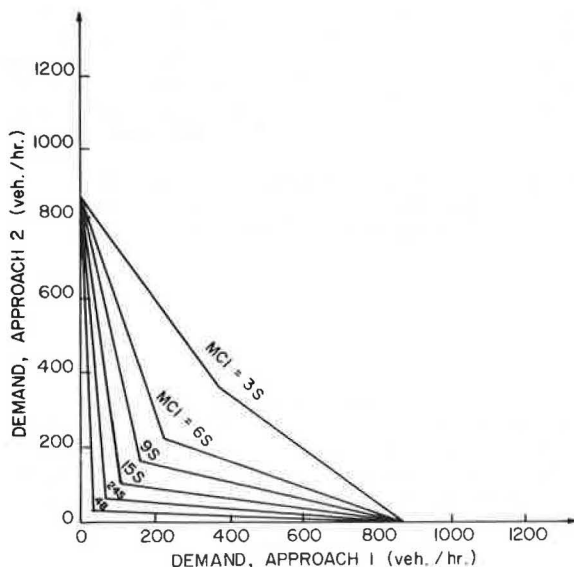
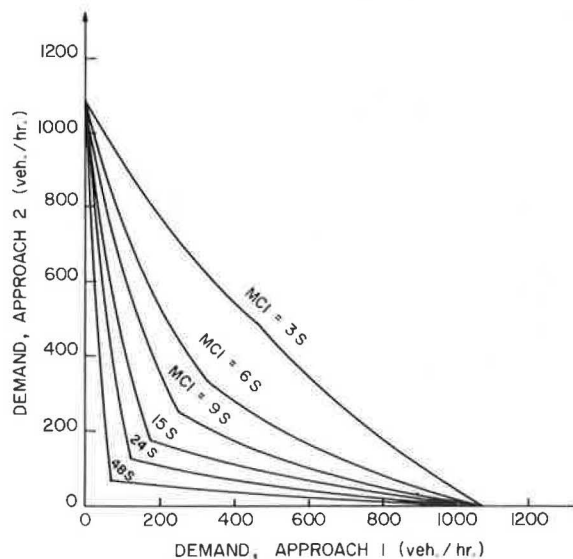


Figure 3. Capacity of stop-sign control at one-lane construction sites: maximum platoon size of one vehicle, $s = 1600$ vph.



demand on one approach, proceeding to the appropriate MCI, and finding capacity on the other axis. From the field data it was found that the maximum platoon size of 2 was most common. It is therefore recommended that Figure 2 be used in the majority of cases when employment of stop-sign control is being considered. Similar figures were derived for maximum platoon sizes of one to five vehicles; Figures 3 and 4 present the cases of maximum platoon sizes of 1 and 2, respectively, derived by assuming saturation flow of 1600 vph. Space limitations do not allow inclusion of all figures and tables presented in an earlier report (1). It should be noted, however, that in real life, queued vehicles often do not come to a complete stop at stop signs. The effect of these rolling stops would be to increase the effective capacity of a site to a point above that for the case in which all vehicles come to a full stop. Therefore, capacities found from the simulations that used a maximum platoon size of one vehicle were felt to be low. Further, in many cases the vehicle immediately behind the first one in the

Figure 4. Capacity of stop-sign control at one-lane construction sites: maximum platoon size of two vehicles, $s = 1600$ vph.



queue will follow that front vehicle past the stop sign without actually stopping or slowing down for the sign. This is especially true when the front of the leading vehicle has come to a stop at some point past the stop sign before proceeding, which leaves less than a full car length between the stop sign and the second vehicle in line. However, rarely will two vehicles in a row follow the leading vehicle without at least one slowing for the stop sign. Consequently, capacities found from the simulations that used a maximum platoon size of three or more vehicles may be a little high.

When the capacities under stop-sign control (Figure 2) are compared with those under signal control (Figure 1), it is apparent that the capacity with stop signs is significantly lower than the capacity with traffic signals, given the same traverse MCI. This result was anticipated, since signals allow for large platoons of vehicles to traverse a construction site as a group, whereas stop signs restrict the platooning effect.

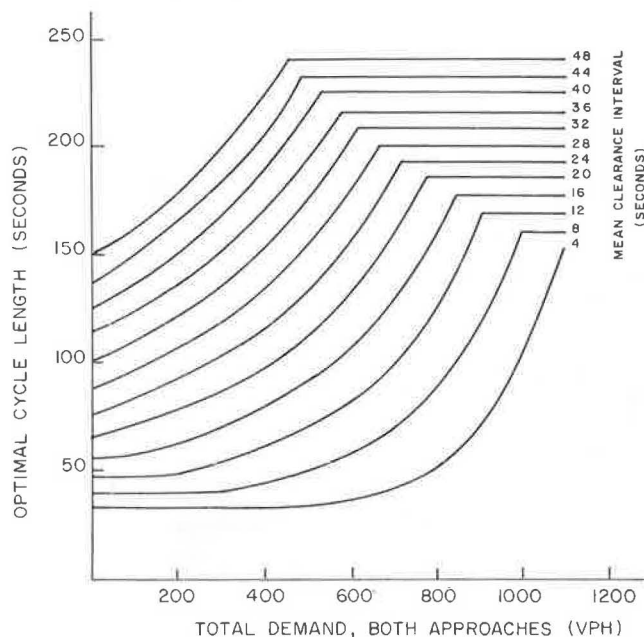
For those combinations of demands and mean traverse intervals where both stop-sign control and signal control are feasible alternatives, selection of one type of control over the other may be based on several different factors. Some guidelines in making such a decision have been included in a later section, although it must be recognized that the criteria for selecting the optimal alternative may depend not only on performance, but also on cost, accident experience, and personal judgment.

Before this section is concluded, it should be pointed out that actual demands at the test sites were insufficient to allow verification of the entire range of capacity tables developed. However, the data collected were useful in adjusting and calibrating the ONELANE program based on the actually observed measures of effectiveness (i.e., delays, stops), traverse time variability, minimum departure headways, capacity at a few locations, etc.

SIGNAL TIMING

In the previous section, optimal signal settings were required to determine the capacity of a one-lane construction site when traffic-signal control is being evaluated. In addition, after it has been determined that traffic signals are the most appro-

Figure 5. Optimal cycle lengths.



appropriate type of control, optimum settings are needed for minimizing delays.

In the following two subsections, a procedure for determining the optimal signal settings for both pretimed and actuated signals is presented. The settings derived from these procedures are based on assumptions that are commonly accepted but may not apply in any particular situation. Therefore, when traffic-signal control is implemented, it is advisable to field check the operation of the system and make appropriate adjustments.

Timing of Pretimed Signals

In order to calculate the optimal cycle length and green times for each approach, the following information is required:

1. Demand on each approach,
2. Estimation of saturation flow rate, and
3. MCI.

In order to accommodate the worst case, the demand used for each approach should be the highest hourly volume during the day. If, for example, different settings are required for the morning, the afternoon, and the off-peak periods, it is possible that three different demands will be used for each approach in calculating all the appropriate settings. In the case of single settings of the signal, the peak-hour volume should be used on each approach, regardless of its time of occurrence. This is common practice in signal timing and ensures that the peak demands will be accommodated in both directions; such treatment, however, will also tend to increase overall delays during off-peak demands. In the following discussion, the acceptable range of signal settings is given as well as an optimal value that minimizes average delay. More specifically, the minimum cycle length (in seconds) is (3)

$$c_{\min} = 2\bar{t} / [1 - (q_1 + q_2)/s] \quad (3)$$

where

- \bar{t} = MCI (s);
 q_1, q_2 = demands on approach 1 and approach 2, respectively (pcph); and
 s = saturation flow of common lane (pcph).

For safety, the minimum cycle length (c_{\min}) must be at least 30 s.

If a maximum green interval of 72 s [i.e., a value slightly higher than that used in England (7)] is assumed, the maximum cycle length should be

$$c_{\max} = 144 + 2\bar{t} \quad (4)$$

Finally, the optimum cycle length that minimizes average delay is (3)

$$c_{\text{opt}} = (3\bar{t} + 5) / [1 - (q_1 + q_2)/s] \quad (5)$$

This is the value that should be used in most cases, subject to the restrictions that the actual cycle length must be greater than or equal to c_{\min} and less than or equal to c_{\max} .

To simplify selection of the proper cycle, calculations for a wide range of demands and clearance intervals were made and are presented in Figure 5. A saturation flow rate of 1200 pcph was employed in developing this figure. To use Figure 5, enter the horizontal axis with the sum of demands on both approaches converted to PCUs. Proceed vertically until the proper MCI is intersected. From that point, proceed horizontally to the left and read the cycle length from the vertical axis.

Once the actual cycle length has been determined, the green times for both approaches are such that the degree of saturation on each approach is the same; i.e.,

$$G_1 = (c - 2\bar{t}) / [1 + (q_2/q_1)] \quad (6)$$

$$G_2 = c - 2\bar{t} - G_1 \quad (7)$$

In addition to green times, an amber clearance interval between 3 and 5 s in duration should be included. Values less than 3 s are not recommended, since vehicles usually require at least 3 s to stop. Values greater than 5 s are not advisable, since longer amber clearance intervals tend to increase the number of violations.

Finally, if the mean traverse time is greater than the amber clearance interval, an all-red interval equal to the difference should be included after the amber interval on each approach. This additional time will help to ensure that vehicles already traveling the one-lane portion in one direction will be cleared before vehicles on the opposing approach receive a green signal indication.

Timing of Actuated Signals

The use of traffic-actuated signals at one-lane construction sites offers the same basic advantage as at intersections. When demands on the approaching roadways are varying considerably with time, traffic-actuated control will usually decrease the delay and inefficiency.

The guidelines for the yellow and all-red times given in the last section are still valid for actuated control. These guidelines may also be compared with those given by the Department of Transportation in England (7).

The minimum green time should be set to 12.0 s, i.e., the same as in pretimed signals. Any phase duration less than this limit is considered unsafe (7). A vehicle extension interval (following the

minimum green) of 7.0 s per actuation is recommended. The objective here is to allow a vehicle to clear the detector (it takes about $3.0 + 1$ s for this at $s = 1200$ PCU/h) and to allow a reasonable time for another arrival. This is also justified by the long traverse times, which make it unwise to end a phase prematurely (this would result in decreased capacity and increased delay). It is of course assumed that only one detector per approach is employed and that it is placed at the stop line. If more than one detector is employed, the extension interval should equal the travel time from the first upstream detector to the stop line. The maximum green time should be in the range of 12-72 s (7) so that the maximum cycle should not exceed $(144 + 2\bar{t})$ s (Equation 4). Calculation of the optimal maximum green interval (G_{max}) can easily follow from the design volumes, which can be obtained from the guidelines of the previous section. With the design volumes known, Equations 3-7 can be used for calculating optimum cycle and green times as if the signal were pretimed. The values so obtained can be used as optimal maximum green times provided that they fall within the bounds specified earlier (12-72 s); otherwise, the boundary values should be employed. It should of course be noted that the maximum extension interval is $G_{max} - 12$ s. In the absence of volume information and/or for comparing the maximum green times calculated from the above guidelines, specific recommendations for initial values of the signal settings as well as for field adjustments have been given elsewhere (7).

ESTIMATION OF MEASURES OF EFFECTIVENESS

Although in most rural and low-volume situations selection of the simplest type of control subject to capacity considerations may suffice, this is not necessarily the case at urban or high-volume sites. Thus, as the complexity of the construction site increases, control performance becomes important and further analysis is needed. This analysis involves estimation of measures of effectiveness such as delay, queue size, number of stops, and energy consumption. The objective is, of course, selection of a traffic-control strategy that both meets the demands adequately and optimizes the measure(s) of effectiveness judged to be most important in a particular instance.

In the remainder of this section, procedures for estimating the most common measures of effectiveness for different types of control are presented. It should be noted, however, that due to the lack of sufficient information, safety considerations are not included.

Analytical Methods

For pretimed and actuated signal control, analytical methods exist for estimating the measures of effectiveness mentioned above. These methods were originally developed for use at isolated signals and are also applicable to one-lane construction sites (3). It should be kept in mind, however, that these analytical techniques are only approximations. This is due to the simplifications made in the process of obtaining closed-form solutions of a rather complex process.

Estimation of delay, stops, and maximum queue size with pretimed signal control can be made by employing Webster's methodology (8) as suggested in the OECD report (3). Since this methodology is widely known, it is not presented here; suffice it to mention that further details of this method along with numerical examples may be found in an earlier

report (1), which also gives guidelines for saturated conditions [not treated by Webster (8)].

A similar procedure for estimating delays, stops, and maximum queue size (analytically) was developed by Courage and Papapanou (9) for actuated signals, and it was adopted in this study. According to this procedure, average delay at a particular approach of a traffic-actuated signal can be estimated from

$$\bar{d} = 0.9 \left\{ \bar{c} (1 - \lambda)^2 / 2(1 - \lambda x) \right\} + \{3600 x_1^2 / 2q(1 - x_1)\} \quad (8)$$

where

\bar{d} = average delay per vehicle (s);

\bar{c} = average cycle length under actuated operation, which is cycle length as if signal were pretimed; average cycle can be calculated from Equation 5 subject to maximum and minimum values suggested in that section;

Δ = \bar{G}/\bar{c} ;

\bar{G} = average green time provided to approach being analyzed calculated from average cycle and Equations 6 and 7;

q = arrival flow rate at approach under consideration;

x = degree of saturation = $q/\Delta s$;

x_1 = $(q/c_{max})/(G_{max}-s)$;

c_{max} = maximum cycle length at which controller is set determined from guidelines of section on signal timing;

G_{max} = maximum green time provided by controller to approach being analyzed determined from guidelines given in section on signal timing; and

s = saturation flow (PCU/h).

As in the case of pretimed signals, Equation 8 is valid for undersaturated conditions, i.e., as long as $x_1 \leq 0.95$. If the average cycle resulting from the computations tends to exceed or equals the maximum cycle, then the signal operates in a pretimed mode. In such a case, the procedures of pretimed control apply.

The maximum queue size under actuated control is (9)

$$N = q(c_{max} - G_{max})/3600 \quad (9)$$

where N represents the maximum queue size in cars that should be expected at the approach being analyzed and the remaining parameters are as defined in Equation 8. Finally, the average number of stops per car, or equivalently the percentage of vehicles stopping, is

$$P = [1 - (\bar{G}/\bar{c})]/[1 - (q/s)] \quad (10)$$

Once delays and stops have been determined, excess energy consumption can easily be estimated by assuming average consumption factors for idling and stops (1,9). In conclusion, it is noted that no reliable analytical estimations of the measures of effectiveness with stop-sign control are available at this time.

Simulation

In the previous subsection, analytical methods for estimating the measures of effectiveness were presented. Because these analytical methods will yield only approximations, caution must be exercised in applying the results derived from these techniques. Naturally, more accurate results can be obtained by employing the ONELANE simulation program, which is specifically designed for this purpose. In order to

Table 2. Measures of effectiveness for stop-sign control.

Avg Clearance Interval (s)	Demand (pcph)		Avg No. of Stops per Vehicle		Avg Delay per Vehicle (s)		Maximum Queue Size ^a (vehicles)		Energy Consumed (gal/1000 vehicles)	
	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2	Approach 1	Approach 2
4	100	100	1.0	1.0	2.3	2.2	1	1	10.4	10.4
	200	100	1.0	1.0	2.4	2.7	2	2	10.4	10.4
	200	200	1.0	1.0	2.9	2.9	2	2	10.5	10.5
	300	100	1.0	1.0	2.5	3.5	2	2	10.4	10.4
	300	200	1.0	1.0	3.4	3.7	2	2	10.6	10.6
	300	300	1.0	1.0	4.9	5.2	3	3	11.0	11.0
	400	100	1.0	1.0	2.7	4.0	2	2	10.5	10.4
	400	200	1.0	1.0	4.3	4.9	3	2	10.8	10.7
	400	300	1.4	1.0	14.5	8.3	7	4	16.1	12.9
	400 ^b	400	7.6	8.0	141	148	32	33	99.6	104
	500	100	1.0	1.0	3.4	4.7	4	2	10.7	10.6
	500	200	1.4	1.0	11.2	6.4	7	3	15.5	11.9
	500 ^b	300	15.5	1.1	252	9.6	72	4	197	52.6
	600	100	1.1	1.0	5.3	5.7	5	2	11.7	10.9
	600 ^b	200	13.2	1.0	170	7.5	58	3	161	38.3
	700	100	5.8	1.0	58.2	6.8	24	2	67.6	19.7
8	800 ^b	100	25.5	1.0	267	7.2	122	2	300	54.5
	100	100	1.0	1.0	3.5	3.3	2	2	10.6	10.6
	200	100	1.0	1.0	4.1	5.1	3	2	10.7	10.7
	200	200	1.0	1.0	7.4	7.3	3	3	11.4	11.4
	300	100	1.0	1.0	5.4	6.9	4	2	11.2	10.9
	300	200	1.9	1.1	33.9	14.2	9	4	24.7	16.4
	300 ^b	300	9.9	9.9	265	264	44	44	143	143
	400	100	1.2	1.0	8.8	9.1	6	2	13.0	11.5
	400 ^b	200	17.9	1.1	403	15.1	88	4	246	77.0
	500 ^b	100	23.8	1.2	188	12.5	54	3	270	42.9
12	100	100	1.0	1.0	5.1	5.3	2	2	10.8	10.9
	200	100	1.0	1.0	8.8	10.0	4	3	11.9	11.5
	200	200	2.4	2.5	69.7	75.0	9	10	35.6	37.0
	300	100	1.5	1.0	20.4	13.6	9	3	18.2	13.5
	300 ^b	200	16.7	2.8	592	84.7	101	10	266	126
	400 ^b	100	7.1	1.0	128	16.4	33	3	92.3	31.5
16	100	100	1.0	1.0	9.3	8.6	3	3	11.6	11.6
	200	100	1.3	1.0	25.2	18.9	7	3	17.5	14.6
	200 ^b	200	8.6	8.3	368	357	42	42	147	145
	300 ^b	100	8.1	1.1	208	23.0	37	4	116	45.2
20	100	100	1.0	1.0	16.7	15.0	3	3	13.2	13.1
	200	100	4.1	1.1	136	30.5	19	4	64.1	33.7
	200 ^b	200	11.7	11.8	608	612	67	68	218	219
	300 ^b	100	17.2	1.1	617	31.0	103	4	275	114
24	100	100	1.1	1.1	32.0	29.7	4	4	16.7	16.6
	200 ^b	100	10.7	1.2	522	44.4	60	5	194	99.2
28	100	100	1.5	1.5	66.1	68.1	5	6	25.7	26.1
	200 ^b	100	12.6	1.7	835	84.4	96	6	266	156
32	100 ^b	100	3.4	3.3	235	224	13	12	73.3	72.0

^aMaximum platoon size is two vehicles.^bIndicates that demands exceed capacity.

assist the designer, tables and figures for estimating the measures of effectiveness similar to those presented in the earlier sections were also developed by simulation with the same volume and traverse MCI combinations. Signal settings, lost times, saturation flow, and other parameters were the same as in Figures 1 and 2. For stop-sign control different tables were generated for each maximum platoon size. Table 2 can be used for estimating the measures of effectiveness when maximum platoon size is 2. The tables for the remaining maximum platoon sizes and control alternatives are not presented here due to space limitations.

After completion of the tables that estimate the measures of effectiveness, a comparison was made between the analytical and the simulation results for the case of signal control. The following conclusions were drawn from this comparison:

1. In practically every case, the analytical methods tended to overestimate average delay per vehicle. This overestimation ranged from less than 10 percent at demand combinations close to capacity to more than 40 percent at relatively low total demands. In general, the percentage of overestimation was higher for the minor approach than for the major approach.

2. The analytical estimates of the average number of stops per vehicle also were generally higher than the simulation results, although the magnitude of overestimation was significantly less than that for delays. In fact, the analytical methods actually matched the simulation results in several cases. As was found for delays, the tendency to overestimate was more pronounced on the minor approach.

3. Estimates of maximum queue size by using the analytical methods were generally lower than those found by using simulation. Regardless of the magnitude of the maximum queue size, the differences between the two methods were consistently between two and four vehicles. One possible explanation for the discrepancy is that the analytical methods do not adequately account for the variability of demand within the control period. The aforementioned difference was only for undersaturated conditions.

4. Because the analytical methods tended to overestimate both the average delay and the average number of stops per vehicle, analytical energy consumption estimates were higher.

Perhaps of greater interest to the reader is the comparison of the measures of effectiveness between stop-sign control and traffic-signal control. This comparison is addressed next along with other con-

siderations for determining the desirable control alternative.

GUIDELINES FOR DETERMINING MOST APPROPRIATE CONTROL

Practical Considerations

Although only the few most widely employed control alternatives were analyzed in detail in this study, it should be realized that other options are also available. In what follows, a brief discussion on the practical considerations of each alternative is presented followed by a more rational approach based on operational efficiency.

The self-regulating type of control is generally nothing more than warning signs that allow motorists to determine priority for themselves. This control is generally used only in very short zones and under very low demands. Where there is a potential for conflict, the system breaks down quite easily, especially in terms of legal assignment of right-of-way. Although self-regulating control is used successfully in some states, it was not included among the controls considered during this study.

Yield signs on one or both approaches have also been used in other states. They too offer some advantages under very low volumes and in short construction areas. Where a yield sign is posted on one approach only, the assignment of right-of-way is positive and understandable. When yield signs are posted on both sides, assignment of right-of-way may not be fully understood by motorists. Because the capacity of the yield type of traffic control is limited, it also was not considered.

The posting of stop signs on both approaches to the construction area has been used in Minnesota and several other states. The installation of such signs is inexpensive and requires very little special equipment and time. In theory, the stop control should provide for a safe environment since all motorists will be stopping in advance of the construction site and therefore reducing their speeds through the construction area. The full stop should also allow motorists to observe and react to any construction equipment or workers that may be present on the site.

In reality, there are many violations of the stop sign, although most are rolling stops where motorists are reacting to the stop-sign control and to the conditions present for traffic and construction. Many of the observed violations were motorists who sought to catch up to a previous vehicle traveling in the same direction rather than stop and wait for traffic from the opposite approach to pass across the bridge. There were very few flagrant violations of the stop sign. The stop-sign control appears to be understood quite well by motorists. Observations at the several sites did not indicate that motorists were confused by the unusual placement of a stop sign compared with that which they normally encounter at intersections. Motorists also tended to react quite well to the need to allow for clearance intervals and to wait for traffic from the other side to clear the site.

Traffic signals are often installed at one-lane sites, primarily because of their positive control of traffic. Both fixed-time and full actuated signals will provide a very visible and positive control over approaching motorists. Full actuated signals can be very efficient in view of the varying demands. Fixed-time signals, although they have very positive control over motorists, are relatively inefficient since they cannot react to any variations in demand. Because of this and the long clearance periods required for bridges, delays can be rather substantial.

In signal control, however, it is necessary to design, install, and remove the traffic signal, at a considerably higher cost than that of stop signs. Signals also require a source of power, which may be a problem in some outlying areas. There is also a continual maintenance responsibility and a potential for equipment malfunctions that could totally upset the positive control of traffic. Fixed-time signals are relatively maintenance free when properly installed. Actuated signals do require some periodic checks to make certain that detectors are working properly and that intervals are properly set. An additional problem with traffic-signal control is that the green light is an invitation for motorists to proceed into the construction zone at a speed that might be greater than desirable. There is also a concern that drivers may speed up to make the green light or to run a yellow light. Perhaps it is worthwhile to note that observations indicated that red-light violations were quite infrequent.

Flaggers are probably the most efficient type of traffic control that can be used. They will provide a very positive control over motorists and can react to changing conditions immediately. A flagger can also react to necessary stoppages of traffic for work operations or for equipment maneuvering. Flaggers can react to varying clearance intervals required for traffic and can practically ensure the highest efficiency over any type of traffic control through the construction area. Flaggers, however, are expensive in that their cost is directly related to both size and duration of the project. Further, it is necessary to provide not only flaggers at either end of the site but also periodic relief during their working period. When flaggers are used at night, lighting should be provided that again increases overall cost and requires a source of electrical power.

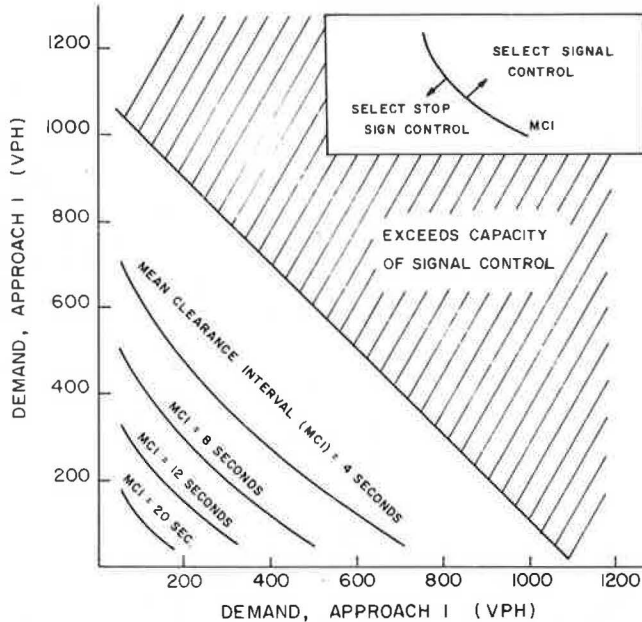
It is also possible to use a combination of controls. A fixed-time traffic signal can be equipped with a remote-control device to override the controller and provide manual control. This would allow manual regulation of traffic during peak hours and higher efficiency. With this option, it is necessary to ensure that the clearance intervals not be modified or that the individual operating the system has a very clear understanding of the operation of the signal and the legality of the clearance intervals.

A similar combination to provide increased efficiency can be made by using stop-sign control supplemented by flaggers during peak hours. This eliminates the need for night-time lighting but provides for a highly efficient method of moving peak-hour volumes. The many advantages of flaggers can be made use of during the day in both peak-traffic and construction periods. The relatively inexpensive stop-sign control can be used at other periods, assuming that the capacity is available. It is also possible to use flaggers during the working day and during the heavier traffic hours but to revert to yield control or self-regulating control during other hours.

Further Considerations

Perhaps the most important factor in determining traffic control at a one-lane construction site is sight distance. Unless vehicles can see each other at opposite ends of the one-lane section of roadway, stop-sign and possibly signal control should be excluded. The guidelines of the section on capacity estimation could be considered in determining maximum sight distances for stop-sign and signal control. After sight-distance and safety constraints, determination of the most desirable control scheme

Figure 6. Optimal control of one-lane construction site based on minimum total delay.



could be made on capacity considerations alone, which may result in elimination of one or more alternatives; capacity can be determined from the methodology presented in this paper. If sight-distance and capacity requirements are met by more than one type of control, the next determining factor should be a comparison of the costs and benefits of each type of control. Aside from the direct costs of installing and maintaining stop signs or signals or of paying flaggers, the user costs in terms of energy consumption and delay may also be considered.

A comparative analysis between stop-sign and signal control was performed for each of the measures of effectiveness. Based strictly on total delay during the design hour (the hour for which the signal settings were determined), Figure 6 was derived. To use this figure, find the point of intersection of the demands of the two approaches. If this point of intersection lies to the left and below the appropriate MCI curve, stop-sign control will yield lower delays than will signal control. On the other hand, if the point lies to the right and above the MCI curve, signal control will yield lower delays than stop-sign control. From this figure, it can be seen that for MCIs greater than about 20 s and demands greater than 200 vph, the total delay when stop-sign control is used will always be higher than that for signal control.

When the average number of stops and the average energy consumed for the two types of control were compared, it was found that in all cases signal control yielded a lower number of stops and a lower energy consumption than stop-sign control. This result is at least partly attributable to the platooning effect of signals, in which queued vehicles rarely are required to stop more than once, whereas queued vehicles controlled by stop signs often stop several times before being discharged.

CONCLUSIONS

Safety along with demand-capacity considerations constitute the basic criteria for selecting the most appropriate type of control at one-lane construction

sites. The guidelines presented here allow estimation of capacity for each control scheme so that exclusion of one or more options that fail to accommodate demand is now possible. Among the control strategies that pass the basic capacity and visibility tests, further screening can be accomplished from safety considerations. Lack of sufficient experimental data did not allow recommendation of safety guidelines sufficient for detailed design. Thus, in addition to visibility one could adopt the OECD recommendations (3) concerning the maximum lengths of the construction zone for stop-sign and signal control (60 and 250 m, respectively). Although no specific data are available for accident predictions, it should be expected that accidents decrease as the sophistication of the control strategy increases.

After the basic capacity and safety tests, further screening can be accomplished from the practical considerations of the previous section, cost-effectiveness analysis, and performance evaluation following the procedures described earlier. It should be kept in mind that for situations out of the realm of the tables and charts presented here, employment of the simulation program developed is highly advisable. This program can also be used for further accuracy or experimentation and it was originally run on a Cyber 730 computer at an average cost of less than \$5/h of simulation (the low cost is due primarily to the event scan structure of the program). Recently the program was rewritten in BASIC, and it can run on an Apple II microcomputer (at a considerably higher execution time but at practically zero cost). Employment of the program is particularly recommended for simulating saturated conditions, which are not covered here.

As mentioned earlier, prediction of accident and safety performance in general was not studied in sufficient detail due to the lack of sufficient data. This subject along with further testing of the program results for the entire volume-capacity range were left for future research. Further, it should be noted that calibration of the program was based on the data collected at rural situations. Adjustments of certain program parameters may be needed for urban conditions.

ACKNOWLEDGMENT

We would like to thank the Minnesota Department of Transportation for financial support of this research.

REFERENCES

1. P.G. Michalopoulos, G. Van Warner, H. Preston, and R. Plum. *Traffic Control for One-Lane Bridges*. Short-Elliott-Hendricson, Inc.; Minnesota Department of Transportation, St. Paul, Study 07-151, 1981.
2. *Traffic Control for Street and Highway Construction and Maintenance Zones*. Byrd, Tallamy, McDonald and Lewis; FHWA, Final Rept., 1978.
3. *Traffic Operation at Sites of Temporary Obstruction*. OECD, Paris, France, 1973.
4. R. Kermod. *A Field Procedure for Determining the Effects of Lane Closures*. California Division of Highways, Sacramento, 1969.
5. E. Lieberman, R.D. Worrall, D. Wicks, and J. Woo. *NETSIM Model: Volume 4--User's Guide*. FHWA, Final Rept., 1977.
6. E. Lieberman, B. Andrews, M. Davilla, and M. Yedlin. *Macroscopic Simulation for Urban Traffic Management: The TRAFLO Model*. FHWA, Final Rept., 1980.
7. *Traffic Signal Control Equipment for Use at*

Roadworks. Traffic Control and Communications Division, Department of Transportation, London, England, Specification MCE 0111, 1979.

8. F.V. Webster and B.M. Cobbe. Traffic Signals. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, TRRL Rept. 56, 1966.

9. K.G. Courage and P.P. Papapanou. Delay of Traffic-Actuated Signals. TRB, Transportation Research Record 630, 1977, pp. 17-21.

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

Abridgment

Single-Lane Capacity of Urban Freeway During Reconstruction

ROBERT E. DUDASH AND A. G. R. BULLEN

Lane capacities of an urban freeway under various traffic-flow configurations during reconstruction are determined. The urban freeway studied was the Penn-Lincoln Parkway, Interstate 376, located in Pittsburgh, Pennsylvania. The study locations on the freeway were in the vicinity of the entrance and exit portals of the Squirrel Hill Tunnel. Traffic flows were studied in the right lane of a two-lane section of the highway for three distinct operating conditions. The first condition consisted of both lanes traveling in the same direction. For the second condition, the left lane was closed because of construction, which left only a single lane open to traffic. In the third condition, the left lane had a single lane of traffic traveling in the opposite direction. The results of this study determined the flow, average speed, and density at capacity for each of the operating conditions. A comparison of the data indicated that the single-lane capacity of both sides of the tunnel was significantly lower during the second and third operating conditions; during the third condition, the lowest level of capacity was attained. Generally, for a two-lane, two-way facility under forced flow, the sustained capacity for a single lane was about 1200 vehicles/h. With both directions under forced flow, a two-way flow of 2400 vehicles/h could be sustained.

Various procedures have been published through the years to evaluate the capacity of a roadway. These include, in the United States, the 1950 Highway Capacity Manual (1), the 1965 Highway Capacity Manual [Highway Research Board Special Report 87 (2)], and, most recently, Interim Materials on Highway Capacity [Transportation Research Circular 212 (3)].

These evaluations of capacity, however, do not consider the effect on traffic flow of construction work zones adjacent to a roadway. Within the past several years, the trend of constructing new highways has decreased, and the trend of reconstructing inadequate highways has increased. Unfortunately, the development of the evaluation of traffic flow through construction work zones has not developed at the same rate. The main examination of the subject has been by Dudek (4), who reports on capacity studies at urban freeway maintenance and construction work zones in Houston and Dallas, Texas, for five-, four-, and three-lane freeway sections.

CONSTRUCTION WORK-ZONE EVALUATION

A study was conducted to compare lane capacities of an urban freeway while it was under various traffic-flow configurations during reconstruction. The urban freeway studied was the Penn-Lincoln Parkway, Interstate 376, located in Pittsburgh, Pennsylvania. The locations chosen for comparison were in the vicinity of the entrance portal (site A) and exit portal (site B) of the Squirrel Hill Tunnel.

The parkway is a four- to six-lane divided urban freeway traversing east and west. It has an average

daily traffic volume of 92 000 vehicles. The Pennsylvania Department of Transportation embarked on a 2-year safety improvement project to update the facility for a length of 6 miles; this consisted of the placement of a new 8-in concrete overlay road surface, new shoulders and concrete median barrier for the entire length of the project, and the rehabilitation of both tubes of the Squirrel Hill Tunnel.

The construction phasing consisted of reconstructing the westbound lanes during the 1981 construction season and the eastbound lanes during the 1982 construction season. The construction required that the westbound and the eastbound lanes be completely closed during various stages of their reconstruction. Since no convenient alternative route was available to detour parkway traffic, it was necessary to maintain two-way opposing traffic on the lanes opposite those being reconstructed.

Traffic flows were studied for the two locations in the right lane of a two-lane section of the highway for three distinct operating conditions. The first condition was for both lanes traveling in the same direction (two lanes, one way). For the second condition, the left lane was closed to traffic because of construction, which left only a single lane of traffic (one lane, one way). In the third condition, the left lane had a single lane of traffic traveling in the opposite direction (two lanes, opposing). Figures 1, 2, and 3 depict the three traffic conditions.

The horizontal alignment at the sites consisted of horizontal curves that were designed for vehicle speeds of more than 65 mph. The vertical alignment, traveling west to east, consisted of 0.5 mile of +4.5 percent grade approaching site A and 1 mile of -2.5 percent grade approaching site B.

The roadway section for sites A and B varied for each condition. During the two-lane one-way condition, two lanes, each 12 ft wide, existed. The right edge of pavement was paralleled by a curb 6 in high and a beam guardrail. The left edge of pavement was paralleled by a 6-in mountable curb and a grass median.

For the one-lane one-way condition, one lane 12 ft wide existed. The right edge of pavement remained unchanged for the two-lane, one-way condition. The left edge of roadway was paralleled by 55-gal drums spaced 100 ft center to center.

The two-lane opposing condition consisted of one lane 12 ft wide for the eastbound traffic and one

Figure 1. Two-lane one-way condition.

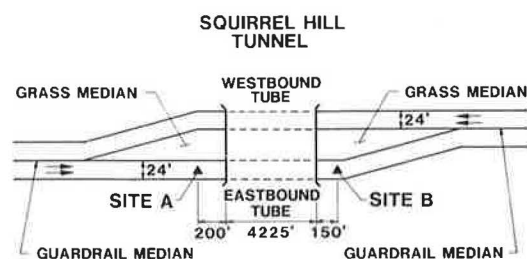


Figure 2. One-lane one-way condition.

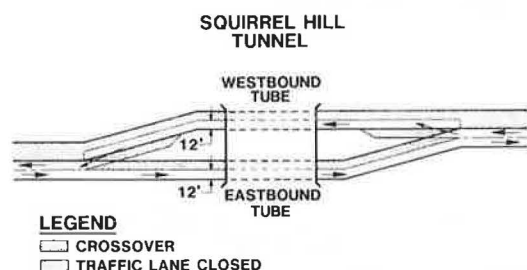
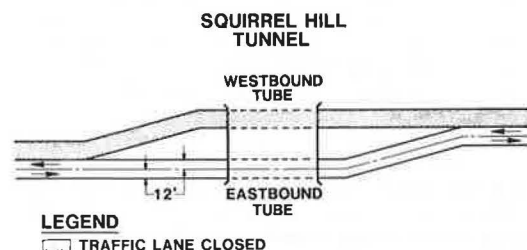


Figure 3. Two-lane opposing condition.



lane 12 ft wide for the westbound traffic. The two-way opposing traffic was separated by a 4-in double yellow center line, raised pavement markers, and traffic guideposts 18 in high.

The section for the 4225-ft-long eastbound tube of the tunnel consisted of two lanes 12 ft wide. The right edge of pavement was paralleled by a curb 8 in high and a parapet 1.5 ft high. The left edge of pavement was paralleled by a curb 8 in high and a parapet and pedestrian walkway 1.5 ft high. The vertical clearance was 19 ft at the face of the portal. Inside the tunnel 92 ft, the vertical clearance dropped to 14 ft 2 in. The same traffic control devices used for the roadway during each traffic condition were the same ones installed in the tunnel for each of the same traffic conditions.

The data were collected by a Stevens PPRII print punch traffic classifier. Volume, speed, and vehicle classifications were monitored. The speed data were reduced to space mean speed and the counts were reduced to flow. From the space mean speed and flow, density was then determined and the relationships of flow versus density, speed versus density, and speed versus flow were plotted.

Usually, data were collected for each day from about 6:00 a.m. to about 7:00 p.m. Two days of data collection were carried out at each site for each traffic condition.

RESULTS AND ANALYSIS

Total data results are summarized in Table 1, which shows the maximum 5-min flow rate along with the speed and density conditions for each site and traffic condition. Also given are the observed maximum sustained 1-h traffic flows. All maximum flow rates were monitored during forced-flow conditions with approximately 9 percent trucks. A comparison of the data indicated that the single-lane capacity for sites A and B was significantly lower during the one-lane one-way and the two-lane opposing conditions; during the two-lane opposing condition the lowest level of capacity was attained.

The data show that for normal operating conditions during the two-lane one-way condition, the maximum lane flow was about 1550 vehicles/h for 5 min and about 1350 vehicles/h for 1 h. This gives a peak hour factor of 0.87.

Under the one-lane one-way condition, the maximum 5-min flow was slightly more than 1400 vehicles/h, whereas the sustained 1-h flow was about 1200 vehicles/h. These volumes represent the maximum throughput of a single lane of traffic through the construction zone and are about 90 percent of the normal volume.

The two-lane opposing condition provided a further reduction of traffic flow to a 5-min flow of 1350 vehicles/h and a 1-h flow of just less than 1200 vehicles/h, which represented a further reduction of about 5 percent in maximum throughput.

The overall results of the study can be seen in Figures 4 and 5, which show the speed-flow envelope curves for these traffic conditions. Observation of

Table 1. Summary of traffic-flow data.

Condition	Site A (tunnel entrance)		Site B (tunnel exit)	
	Day 1	Day 2	Day 1	Day 2
Two lanes one way				
Maximum 5-min flow (vehicles/h)	1512	1560	1596	1644
Mean speed (mph)	25.6	37.0	45.3	42.4
Density (vehicles/mile)	59	42	35	39
Maximum 1-h flow (vehicles/h)	1355	1359	1402	1414
One lane one-way				
Maximum 5-min flow (vehicles/h)	1428	1416	1536	1488
Mean speed (mph)	27.2	26.1	39.1	36.4
Density (vehicles/mile)	53	54	39	41
Maximum 1-h flow (vehicles/h)	1169	1208	1264	1258
Two lanes opposing				
Maximum 5-min flow (vehicles/h)	1368	1308	1344	1356
Mean speed (mph)	25.9	28.8	39.6	39.5
Density (vehicles/mile)	53	45	34	34
Maximum 1-h flow (vehicles/h)	1188	1153	1187	1223

Figure 4. Speed-flow relationship, entrance portal, for all three conditions.

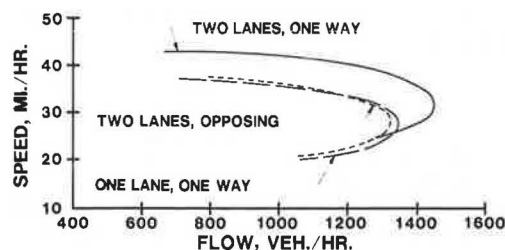
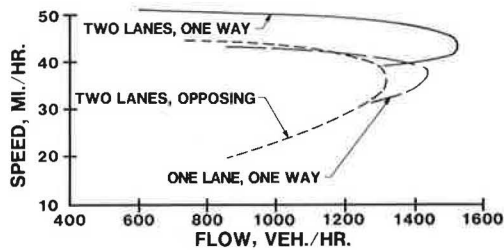


Figure 5. Speed-flow relationship, exit portal, for all three conditions.



the curves provides the following conclusions for site A (tunnel entrance):

1. The shape of the curves for each condition delineates wide envelopes, which is an indication that the entrance of the tunnel was a bottleneck location.
2. The curves were influenced by the reduction in the number of lanes available for traffic. The plot for the one-lane one-way condition and the two-lane opposing condition resulted in a shift of the curves down and to the left.
3. The opposing traffic during the two-lane opposing condition had a slight influence on the speed-flow curve. The curve moved left from the curve for the one-lane one-way condition.

The conclusions for site B (tunnel exit) were as follows:

1. The shape of the curves for the two-lane one-way condition and the one-lane one-way condition delineates narrow envelopes, which is an indication of free flow. The wide envelope of the two-way opposing condition indicates a bottleneck location.
2. The speed-flow curves were influenced by the reduction in the number of lanes available for traffic. The plot for the one-lane one-way condition and the two-lane opposing condition resulted in a shift of the curves down and to the left of the two-lane one-way condition. The overall traffic flows were controlled, however, by the discharge past the tunnel entrance.

CONCLUSIONS

The study was mainly concerned with determining the single-lane capacity of an urban freeway while under various reconstruction configurations, but it also gives some quantitative insight into capacity flows of a two-way two-lane highway.

The study shows that single-lane traffic flow

through a construction zone is significantly less than single-lane traffic flow as part of a multilane flow under normal conditions. Single-lane flow alongside an opposing traffic stream has a somewhat lower capacity than single-lane flow with no opposing traffic stream.

Generally, it seems that for single-lane traffic flow under the given geometric conditions and 9 percent trucks, a flow of only 1200 vehicles/h can be sustained; short-term flow rates go up to 1350 vehicles/h. These agree reasonably well with the Texas study conducted by Dudek (4). During forced-flow conditions, however, traffic flow is very variable and follows no particular speed-flow relationship.

During the two-way opposing traffic condition, forced flow existed from both directions. Some manual-count samples showed that with two-way traffic, similar flows were being achieved in the opposite lane. This would indicate that the capacity of two-way two-lane sections of roadway with the given geometrics under forced-flow conditions can reach 2400-2500 vehicles/h.

Comparison of the data for each of the traffic-flow configurations indicates that the tunnel entrance was a restraint to traffic flow during the two-lane one-way condition and for the one-lane one-way condition. However, during the one-lane opposing condition, the opposing traffic caused the restraint to the traffic flow.

ACKNOWLEDGMENT

This paper is based largely on a graduate project carried out at the University of Pittsburgh by Robert E. Dudash. The field work was carried out with the permission and assistance of Roger E. Carrier from the Pennsylvania Department of Transportation and Arthur Hopperstead from Trumbull-Denton Joint Venture Corporation. The services of the data-collection equipment were donated by James Martin from Jamar Sales Company, Inc.

REFERENCES

1. Highway Capacity Manual. Bureau of Public Roads, U.S. Department of Commerce, 1950.
2. Highway Capacity Manual 1965. HRB, Special Rept. 87, 1966.
3. Interim Materials on Highway Capacity. TRB, Transportation Research Circular 212, Jan. 1980.
4. C.L. Dudek and S.H. Richards. Traffic Capacity Through Urban Freeway Work Zones in Texas. TRB, Transportation Research Record 869, 1982, pp. 14-18.

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

Comparative Analysis of Signalized-Intersection Capacity Methods

ADOLF D. MAY, ERGUN GEDIZLIOGLU, AND LAWRENCE TAI

The primary objective of this study was the evaluation of eight methods currently available for capacity and traffic-performance analysis at signalized intersections. The eight methods included the U.S. Highway Capacity Manual method, British method, Swedish method, Transportation Research Board (TRB) Circular 212 planning method, TRB Circular 212 operations and design method, Australian method, National Cooperative Highway Research Program (NCHRP) planning method, and NCHRP operations method. The evaluation was based on applying the eight methods to five intersection data sets for which not only input data were available but also field measurements of lane saturation flows, delays, and percentage of vehicles stopped. The five intersections varied from simple geometric designed intersections with pretimed two-phase signals to more complicated intersections with actuated multiphase signals. Cost-related and effectiveness-related criteria were established in order to evaluate the eight methods. The cost-related criteria included time needed to learn the method, input requirements, clarity and completeness of methodology, and time needed to apply the method. The effectiveness-related criteria included degree of disaggregation, capacity-performance outputs, flexibility of use, and accuracy of predictions. The conclusions included the identification of the limitations of the study, significant results, and future research directions. The NCHRP operations method and the Australian method were found to be the most cost-effective methods. The other methods were about equal in their cost-effectiveness. The NCHRP planning method was found to be an acceptable method when level of effort available is limited and only overall intersection evaluation is needed.

Since an intersection is a point that has to serve all approaches in turn on a highway network, generally the capacities of intersections determine the capacities of a highway network. In order to increase the safety and the capacity of intersections, signals have been used since the 1920s (1). The first U.S. Highway Capacity Manual (HCM) in 1950 contained a chapter for estimating the capacities of signalized intersections (2). Numerous studies have been undertaken to evaluate the different aspects of signalized intersections, and many capacity methods have been developed. Each method corresponds to certain conditions. There is a need for a comparative evaluation to determine which methods should be chosen to evaluate a certain aspect of a signalized intersection. Therefore, a study was designed to apply the eight most accepted methods for capacity analysis to five comprehensive intersection data sets in order to draw conclusions about the eight methods.

A literature search was undertaken and resulted in the identification of eight leading methods for the capacity analysis of signalized intersections. These methods are

1. U.S. HCM method [1965 (3)],
2. British method [1966 (4,5)],
3. Swedish method [1977 (6,7)],
4. Transportation Research Board (TRB) Circular 212 planning method [1980 (8)],
5. TRB Circular 212 operations and design method [1980 (8)],
6. Australian method [1981 (9)],
7. National Cooperative Highway Research Program (NCHRP) planning method [1982 (10)], and
8. NCHRP operations method [1982 (10)].

The next step in the research was to make a comparative analysis of the eight methods in which their input requirements and output options were considered. Inputs were classified as approach width, approach volume, exclusive left- or right-turn lane and length of this lane, percentage of

heavy vehicles, number of local buses, parking conditions, peak-hour factor, pedestrians, and several others. Outputs were classified as saturation flow, capacity, delay, percentage of stopped vehicles, queue length, degree of saturation, level of service (LOS), signal timing, and several others. This analysis and a classification scheme of the methods are given next.

The third task was to obtain intersection data sets that included all input data requirements for all eight methods as well as selected traffic-performance measurements that could be compared with the predictions of the various methods. Five intersections were selected from an NCHRP study; all five intersections were located in the San Francisco Bay Area. They included simple geometric designed intersections with two phases and extended into more complicated actuated intersections with multiphases.

The fourth task was to apply the eight methods to the five intersection data sets. Each method was carefully studied in terms of clarity of use, technical completeness, and time requirements for learning and applying the method.

The fifth task was the comparative analysis and evaluation of the eight methods. The criteria for evaluation consisted of input requirements, output flexibility, accuracy of results, and level of effort.

INPUT REQUIREMENTS AND OUTPUT OPTIONS

Input requirements and output options are a mixed blessing. On the one hand, higher levels of input or output generally require more data collection and preparation, more complicated analytical procedures, and greater time required to learn and use the method. On the other hand, lower levels of input or output generally limit the breadth of application, degree of accuracy, and flexibility of output. Hence users must carefully select that method that meets their desired objectives at minimum cost. This discussion is included to aid users by providing a comparative analysis of input requirements and output options of the various methods. Input requirements and output options are summarized in two tables and explained in the next two sections. A classification scheme of the methods is developed and is discussed in the third section.

Input Requirements

The input requirements of each of the methods were reviewed and are summarized in Table 1. The following paragraphs describe the contents of Table 1 under three subcategories: supply inputs, demand inputs, and control inputs.

Supply inputs consist of geometric dimensions, lane configurations, street alignment, bus-stop location, parking, and site characteristics. The only supply input required by all methods is the presence of turning lanes and the permitted movements by lane. The two TRB Circular 212 methods and the NCHRP planning method require little additional supply input. On the other hand, the other five methods require considerably more supply, particularly the NCHRP operations method.

Table 1. Input requirements of eight methods.

Input Requirement	Method							
	U.S. HCM	British	Swedish	TRB Circular 212		Australian	NCHRP	
				Planning	Operations and Design		Planning	Operations
Supply input								
Geometric dimension								
Approach width	X	X	X		X	X		X
Lane width			X		X	X		X
Turning-lane length			X			X		X
Distance to parking		X	X			X		X
Lane configuration								
No. of lanes			X	X	X	X	X	X
Turning lanes	X	X	X	X	X	X	X	X
Permitted movements	X	X	X	X	X	X	X	X
Street alignment								
Horizontal		X						
Vertical		X	X			X		X
Bus-stop location	X							X
Parking								
Presence	X	X				X		X
Turnover								X
Site characteristic								
Population	X							
Location	X	X				X		X
Demand input								
Approach volume								
Through	X	X	X	X	X	X	X	X
Right-turn	X	X	X	X	X	X	X	X
Left-turn	X	X	X	X	X	X	X	X
Vehicle classification								
Truck	X	X	X		X	X		X
Bus	X	X	X					X
Motorcycle or bicycle		X	X					
Traffic pattern								
Peak-hour factor	X				X			X
Arrival platoon								X
No. of local buses	X				X			X
Pedestrian volume			X		X	X		X
Control input								
Signal-timing plan								
Cycle length	X	X	X	X	X	X	X	X
No. of phases	X	X	X	X	X	X	X	X
Duration of phases	X	X	X	X	X	X	X	X
Other signal information								
Pedestrian time			X					X
Pedestrian pushbutton								X

Note: X indicates maximum input. Some methods do not require some inputs.

Demand inputs consist of approach volumes, vehicle classifications, number of local buses, pedestrian volumes, and traffic patterns. The only demand input required by all methods is the approach volume. The TRB Circular 212 planning method and the NCHRP planning method require no additional demand input data. On the other hand, the other six methods require considerably more demand input, particularly the NCHRP operations method.

Control inputs include the signal-timing plan and other signal information. In most methods the signal-timing-plan input is optional. That is, the signal-timing-plan data can be used in the method if provided or can be determined by the method if all other input data are provided except for the signal-timing plans. This option is most readily available in the British, Swedish, Australian, and NCHRP operations methods.

In summary, the TRB Circular 212 planning method and the NCHRP planning method require the least input, whereas the NCHRP operations method requires the greatest amount of input. It should be noted that some of the methods that require the greatest amount of input have incorporated default values when selected input data are not available.

Output Options

The output options of the methods were reviewed and

are summarized in Table 2. The following paragraphs describe the output options of the various methods on a comparative basis.

The upper portion of Table 2 provides a comparison between the methods in terms of disaggregation of analysis: lane, lane group, approach, and total intersection. The TRB Circular 212 planning method, the TRB Circular 212 operations and design method, and the NCHRP planning method analyze the total intersection in an aggregate form. At the other extreme in output flexibility, the NCHRP operations method permits the analysis to be made on an individual-lane basis and then provides for aggregation of results on a lane-group, approach, and intersection basis. The HCM method and the British method perform their analyses on an approach and/or turning-lane basis. The Swedish and Australian methods perform their analyses on an individual-lane basis, which could be but is not formally aggregated.

The major outputs of most methods are capacity related and performance related. The TRB Circular 212 planning method, the TRB Circular 212 operations and design method, and the NCHRP planning method do not provide any capacity-related or direct performance outputs. All other methods provide for capacity-related outputs and (with the exception of the HCM method, which provides load-factor information) vehicle-performance-related measures such as delay, percentage of stopped vehicles, and queue

Table 2. Output options of eight methods.

Output Option	Method							
	U.S. HCM	British	Swedish	TRB Circular 212		Australian	NCHRP	
				Planning	Operations and Design		Planning	Operations
Degree of disaggregation								
Lane	X ^a	X ^a	X			X		X
Lane group								X
Approach	X	X						X
Total intersection				X	X		X	X
Capacity-related								
Saturation flow	X	X	X			X		X
Capacity	X	X	X			X		X
Vehicle-performance-related								
Avg delay		X	X			X		X
Percentage of stopped vehicles		X	X			X		X
Queue length		X	X			X		X
Load factor	X							
Pedestrian-performance-related								
Delay						X		
No. stopped						X		
LOS	X			X	X		X	X
Preliminary design	X	X	X	X	X	X	X	X
Improved signal timing								
Cycle length		X	X			X		X
No. of phases						X	X	X
Duration of phases		X	X			X		X
G/C ratio	X	X			X	X		X

^aTurn lanes only.

length. The Australian method also provides pedestrian performance output.

All U.S. methods provide for an LOS determination and all methods provide for some preliminary design evaluation. The British, Swedish, Australian, and NCHRP operations methods provide for signal-timing determination with the capacity-performance analysis.

Classification of Methods

A review of the last two sections shows that these eight methods fall into a classification scheme that consists of three groups of methods determined by their input requirements and output options. The first group is called the most comprehensive group and consists of the British method, the Swedish method, the Australian method, and the NCHRP operations method. These methods require extensive inputs and analysis time but result in wide and extensive output information. Group 2 is intermediate according to work, input requirements, and output options and consists of the HCM method and the TRB Circular 212 operations and design method. The third group includes the least-complicated methods, which require the least input and working time, and consists of the TRB Circular 212 planning method and the NCHRP planning method. These methods can be used for a very fast overview investigation of the performance of intersections. The variety of comprehensiveness between methods complicates the comparative analysis by considering both the cost and the effectiveness and by restricting the appropriate method selected by the method's dependence on application, input data availability, and output requirements. These issues will be addressed later.

INTERSECTION DATA SETS

The next task of the study was to obtain intersection data sets that included all input data requirements for all eight methods as well as selected traffic-performance measurements that could be compared with the predictions of the various methods. Five such data sets were obtained from an NCHRP

study. These particular intersection data sets were selected by considering completeness of field measurements, variety of design and control features, and proximity to the San Francisco Bay Area.

The location and characteristics of these five intersections and their input and traffic-performance measurements will be presented and described in the following two sections.

Location and Characteristics of Intersections

The first intersection is located in Berkeley, California, at Sacramento and Dwight. Sacramento is a divided north-south major arterial with two through lanes in each direction plus an exclusive left-turn lane. Dwight is an undivided east-west collector-type street with one lane in each direction. The intersection is located in a residential area and some distance from adjacent signalized intersections. The signal is a two-phase pretimed controlled signal with a 70-s cycle. The traffic flows are generally light and the percentage of trucks is of the order of 1-2 percent. Parallel parking is permitted on all approaches and near-side bus stops occur on the Sacramento approaches with a far-side bus stop eastbound on Dwight.

The second intersection is located in Berkeley, California, at San Pablo and University. San Pablo is a divided north-south major arterial with two through lanes and an exclusive left-turn lane in each direction. University is a divided east-west major arterial with two through lanes and an exclusive right-turn lane (left-turning movements are prohibited). The intersection is located in an outlying business district and adjacent signalized intersections are approximately 0.25 mile away with an average level of coordination. The signal is a two-phase pretimed controlled signal with a 65-s cycle. The traffic flow is moderately heavy--4-6 percent trucks. Parallel parking and near-side bus stops are located on all approaches.

The third intersection is located in El Cerrito, California, at Carlson and Central. Carlson is a north-south major arterial with two through lanes

and an exclusive left-turn lane in each direction. Central is an east-west major arterial with two through lanes in each direction. The intersection is located in an urban fringe area and has little signal coordination with adjacent signals. The signal is a three-phase pretimed controlled signal with an 80-s cycle (the third phase is for exclusive left-turn movements from Carlson). The traffic flow is moderately heavy--1-3 percent trucks. Parking is not permitted on any of the approaches and near-side bus stops exist on the eastbound and southbound approaches.

The fourth intersection is located in Richmond, California, at San Pablo and McDonald. San Pablo is a divided north-south major arterial with two through lanes and with exclusive left-turn and right-turn lanes. McDonald is an east-west major arterial with one through lane and with an exclusive left-turn lane in each direction (in addition, eastbound McDonald has an exclusive right-turn lane). The intersection is located in an outlying business district and there are no nearby adjacent signals. The signal is a six-phase fully actuated controlled signal with a maximum 95-s cycle. The sequence of phasing is eastbound (all movements), westbound (all movements), northbound and southbound left-turn movements only, continued northbound or southbound left-turn movement with adjacent through and right-turn movement, and finally the through- and right-turn movements only on San Pablo. The traffic flow is very heavy with 0-2 percent trucks. Parking is not permitted on any approach and the only bus stop (near side) is located on the eastbound approach.

The fifth and last intersection is located in Berkeley, California, at Grove and Rose. Grove is a north-south collector-type street with a single-lane approach in each direction. Rose is an east-west collector-type street with a single-lane approach in each direction. The intersection is located in an urban fringe area with no nearby adjacent signals. The signal is a two-phase pretimed controlled signal with a 65-s cycle. The traffic flow is generally light--1-5 percent trucks. Parking is permitted on all approaches and there are far-side bus stops on Grove.

In summary, these five intersections have some common features and at the same time each has some unique features. All intersections are located in flat topography, have four approaches, provide for two-way operations on all crossing streets, and have less than 6 percent truck traffic. The unique features of the intersection at Sacramento and Dwight are that two approaches have only single lanes and the traffic flow is generally light. The intersection at San Pablo and University has fairly heavy pedestrian and local bus activities and special turn features (no left turns on two approaches and an exclusive right-turn lane). The intersection at Carlson and Central has a three-phase pretimed signal controller and parking is not permitted on any of the approaches. The intersection at San Pablo and McDonald has a six-phase fully actuated signal controller and operates under very heavy traffic conditions. The intersection at Grove and Rose has only single-lane approaches with relatively heavy pedestrian activities.

Input and Traffic-Performance Measurements

The field measurements that are necessary in order to apply the eight methods to the five intersections are given in Table 3. Not all input measurements were required in all of the eight methods.

Selected traffic-performance measurements were taken for selected lane groups at the five intersections. These field measurements consisted of satu-

ration flow, average delay, and percentage of vehicles stopped. These measurements are presented in Figures 1-3.

NUMERICAL RESULTS

The eight methods were then applied to the five intersection data sets. This application process provided three different ways of evaluating and comparing the eight methods. First, the application provided insights as to the clarity and the completeness of the methods. Evaluation of clarity and completeness of the methods will be discussed later.

Second, the application provided estimates of time requirements to learn and apply each method. These time-requirement estimates are presented in Table 4 for comparative purposes and not as absolute values. As noted from the table, time needed to learn and apply the methods varied considerably.

Finally, the application provided traffic-performance predictions that could be compared with field-measured traffic performances. These comparisons included saturation flow, average total delay, percentage of vehicles stopped, and LOS and will be presented and discussed in the following four paragraphs.

Five of the eight methods predict saturation flow. Measured saturation flow and directly comparable predicted saturation flow for the five methods were available for 13 lane groups. These saturated flows are presented in tabular form in Figure 1. The intersection number, direction of movement, and permitted traffic movements are identified for each lane group. Then the lane saturation flows for each of the five methods are presented as well as the measured values. The mean value and standard deviation are given for each column.

Four of the eight methods predict average total delay. Measured delay and directly comparable predicted delay for the four methods were available for 30 lane groups. These average total delays are presented in tabular form in Figure 2. The individual lane groups are identified and average total delays entered for each method. The mean value and standard deviation are given for each column.

Four of the eight methods predict percentage of vehicles stopped. Measured percentage of vehicles stopped and directly comparable predicted percentage of vehicles stopped for the four methods were available for 30 lane groups. These data are presented in tabular form in Figure 3. The individual lane groups are identified and the percentage of vehicles stopped is entered for each method. The mean value and standard deviation are given for each column.

Three of the methods could not be included in the saturation-flow evaluation and four of the methods could not be included in the evaluation of average total delay or percentage of vehicles stopped. In order to provide some form of comparison for these methods, an LOS analysis was performed. The measured LOS was based on measured average total delay and converted to the LOS scale based on the NCHRP operations method scale relationship. All five U.S. methods provided LOS predictions. These LOS indicators are presented in tabular form in Figure 4 for 30 lane groups. The individual lane groups are identified and LOS is entered for each method. Different methods provided different levels of aggregation of LOS as indicated in Figure 4.

COMPARATIVE ANALYSIS AND EVALUATION

The comparative analysis and evaluation of the eight methods are discussed in this section. The criteria selected for evaluation included level of effort required and quantitative or qualitative effective-

Table 3. Input measurements of five selected intersections.

Input ^a	Sacramento-Dwight (Berkeley)				San Pablo-University (Berkeley)				Carlson-Central (El Cerrito)			
	NB	SB	EB	WB	NB	SB	EB	WB	NB	SB	EB	WB
Approach width (ft)	45	45	17	21	39	39	31	32	34	34	22	28
No. of through lanes	2	2	1	1	2	2	2	2	2	2	2	2
No. of right-turning lanes	0	0	0	0	0	0	1	1	0	0	0	0
No. of left-turning lanes	1	1	0	0	1	1	0	0	1	1	0	0
Length, turning lane (ft)	110	110	-	-	170	165	-	-	150	100	-	-
Right-turning radii (ft)	10	10	10	10	10	10	10	10	15	10	40	20
Left-turning radii	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Gradient of approach	0	0	0	0	0	0	0	0	0	0	0	0
Peak-hour factor	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
Bus-stop location	Y	Y	Y	N	Y	Y	Y	Y	N	Y	Y	N
Site characteristic	Residential				Outlying business				Fringe			
Parking condition	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N
Parking distance to stop line (ft)	110	72	65	60	90	90	140	140	165	-	-	-
Approach volume	720	620	440	390	420	480	820	850	680	260	800	360
Load factor	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Right-turning volume	50	40	50	18	65	52	13	120	40	37	200	74
Left-turning volume	84	125	55	47	73	81	-	-	210	59	78	16
Heavy vehicles (%)	1	1	2	1	6	3	4	4	1	1	2	1
No. of local buses	11	0	0	0	8	7	8	13	0	1	7	0
Bicycles or motorcycles (%)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Pedestrian volume	0	0	9	51	57	84	7	45	1	5	0	7
No. of parking maneuvers	1	1	1	1	3	1	3	3	-	-	-	-
Arrival type	3	3	3	3	3	3	4	4	4	3	3	3
Signal timing ^b	Pt	Pt	Pt	Pt	Pt	Pt	Pt	Pt	Pt	Pt	Pt	Pt
No. of phases	2	2	2	2	2	2	2	2	3	3	3	3
Pedestrian walking time (s)	11	9	22	22	16	16	17	17	19	14	17	17

Note: Y = yes; N = no; NA = not available.

^aThese parameters are those required by the NCHRP method (10).^bPt = pretimed; Act = fully actuated.

ness. These two criteria were applied to the eight methods and the comparative analysis and evaluation are presented in the following two subsections.

Level of Effort Required

Time needed to learn the method, input requirements, clarity and completeness of methodology, and time needed to apply the method were used to assess the level of effort required. This evaluation resulted in the classification of the eight methods into three groups.

The first group, the one requiring the least level of effort, consisted of the TRB Circular 212 planning method and the NCHRP planning method. These two methods required only 1-2 h in order to learn the method and only 0.5 h to typically apply the method. These two methods required much less input data than the other six methods. The experience from this study indicated that step-by-step procedures and sample problems are slightly clearer in the NCHRP planning method than in the previous TRB Circular 212 planning method.

The second group, the one requiring a moderate level of effort, consisted of the HCM method and the TRB Circular 212 operations and design method. These two methods required 2-4 h in order to learn the method and approximately 1 h to typically apply the method. These two methods required a moderate amount of input data and are about equal in terms of clarity and quality of sample problems.

The third and last group consisted of the four methods that require the greatest level of effort. These were the British, the Swedish, the Australian, and the NCHRP operations methods. These methods required 8-20 h to learn and typically required 1-4 h per application. Input requirements were rather significant; they varied from the British method, which required the least in this group, to the NCHRP operations method, which required the greatest amount of input data. Several of the methods, notably the NCHRP operations method, provided default values when certain input data are not available.

Of these four methods, the NCHRP operations method is the most user-oriented; it uses clear step-by-step procedures and numerical examples.

Quantitative or Qualitative Effectiveness

Degree of disaggregation, capacity-performance outputs, flexibility of use, and accuracy of predictions were used as criteria in assessing the quantitative or qualitative effectiveness of the eight methods. It was more difficult to classify the eight methods into groups based on criteria of quantitative or qualitative effectiveness. Instead, the methods will be evaluated for each criterion individually in the following four paragraphs.

The criterion of degree of disaggregation provided a clear differentiation between methods. The two TRB Circular 212 methods and the NCHRP planning method are the most aggregated in that results are provided only for the total intersection. The next level of disaggregation was provided by the HCM method and the British method on an intersection-approach basis and for special turn lanes. The Swedish and the Australian methods provide analysis on an individual-lane basis. The greatest degree of flexibility in degree of disaggregation is provided in the NCHRP operations method, in which the user may perform analysis by individual lane, lane group, approach, and/or total intersection.

The criterion of capacity-performance output applied to the eight methods provided a three-level classification. The two TRB Circular 212 methods and the NCHRP planning method provided the least output; only the overall LOS was specified. The HCM method provided saturation-flow and capacity values but the only performance values were load factor and LOS. The other four methods (British, Swedish, Australian, and NCHRP operations methods) provided saturation flow, capacity, delay, percentage of stopped vehicles, and queue-length outputs. In addition, the Australian method provided pedestrian-performance outputs.

The criterion of flexibility of use included LOS,

Figure 1. Results for lane saturation flow.

San Pablo-McDonald (Richmond)				Grove-Rose (Berkeley)			
NB	SB	EB	WB	NB	SB	EB	WB
44	43	39	23	20	20	18	20
2	2	1	1	1	1	1	1
1	1	1	0	0	0	0	0
1	1	1	1	0	0	0	0
300	160	250	110	-	-	-	-
10	40	20	40	10	10	10	10
NA	NA	NA	NA	NA	NA	NA	NA
0	0	0	0	0	0	0	0
0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
N	N	Y	N	Y	Y	N	N
Outlying business				Fringe			
N	N	N	N	Y	Y	Y	Y
-	-	-	-	5	18	37	16
1210	560	790	300	350	450	150	170
NA	NA	NA	NA	NA	NA	NA	NA
95	78	260	56	60	20	31	21
210	108	320	113	18	23	30	39
2	1	1	0	4	1	5	1
0	0	4	0	3	3	0	0
NA	NA	NA	NA	NA	NA	NA	NA
13	4	0	11	51	56	4	79
-	-	-	-	4	4	2	0
3	3	3	3	5	3	3	2
Act	Act	Act	Act	Pt	Pt	Pt	Pt
6	6	6	6	2	2	2	2
12	16	21	18	10	9	13	10

LOCATION			LANE SATURATION FLOWS (VPH OF GREEN)					
INTERSECTION NUMBER	DIRECTION OF MOVEMENT	TRAFFIC MOVEMENTS	MEASURED	PREDICTED METHOD				
				HCN	BRITISH	SWEDISH	AUSTRALIAN	NCHRP
1	EB		1494	1234	1898	1820	1498	1604
1	WB		1457	1411	2367	1360	1530	1680
3	NB		1636	1248	1620	1670	1810	1573
3	EB		1386	1430	1532	1770	1596	1647
4	NB		1446	1360	1638	1650	1793	1484
4	NB		1626	1300	2078	1810	1814	1673
4	SB		1660	1278	2101	1800	1775	1631
4	EB		1446	1679	1886	1830	1775	1476
4	WB		1857	1714	2470	1860	1614	1614
5	NB		1104	1480	2132	1630	1348	1353
5	SB		1475	1802	1775	1880	1429	1401
5	EB		1286	1012	1805	1160	1284	1270
5	WB		1525	1157	1488	1760	1219	1373
MEAN			1490.6	1392.7	1906.9	1692.3	1575.8	1521.5
STD. DEVIATION			181.9	229.4	308.9	211.1	211.0	137.1

Figure 2. Results for average total delay.

LOCATION			AVERAGE TOTAL DELAY (SECS/VEH)				
INTERSECTION NUMBER	DIRECTION OF MOVEMENT	TRAFFIC MOVEMENT	MEASURED	PREDICTED METHODS			
				BRITISH	SWEDISH	AUSTRALIAN	NCHRP
1	NB		11	12	12	10	32
1	NB		10	9	13	11	11
1	SB		15	9	11	9	31
1	SB		7	9	11	10	10
1	EB		29	12	27	19	42
1	WB		39	14	24	18	37
2	NB		22	11	13	13	40
2	NB		16	12	13	13	13
2	SB		14	11	13	13	40
2	SB		9	12	13	14	13
2	EB		22	15	19	15	11
2	WB		6	13	20	14	14
3	NB		32	25	30	28	35
3	NB		21	20	22	25	18
3	SB		28	24	24	25	42
3	SB		25	18	19	20	21
3	EB		34	18	24	21	29
3	WB		17	17	18	18	29
4	NB		41	32	31	35	56
4	NB		39	23	22	28	27
4	SB		40	43	17	42	66
4	SB		31	28	15	33	33
4	EB		47	63	30	34	59
4	EB		77	37	35	35	38
4	WB		40	37	23	37	71
4	WB		49	38	35	39	36
5	NB		3	5	6	6	7
5	SB		4	5	5	6	12
5	EB		47	21	26	22	43
5	WB		19	19	21	21	53
MEAN			26.5	20.4	19.7	21.1	32.3
STD. DEVIATION			16.8	13.0	8.0	10.4	17.4

Figure 3. Results for percentage of vehicles stopped.

LOCATION			VEHICLE STOPPING (PERCENT)				
INTERSECTION NUMBER	DIRECTION OF MOVEMENT	TRAFFIC MOVEMENT	MEASURED	PREDICTED METHOD			
				BRITISH	SWEDISH	AUSTRALIAN	NCHRP
1	NB		80	62	56	51	68
1	NB		47	53	55	55	46
1	SB		54	52	56	55	66
1	SB		29	52	55	52	45
1	EB		43	74	70	77	78
1	WB		70	70	68	73	74
2	NB		86	59	64	59	77
2	NB		40	63	63	60	51
2	SB		56	59	64	60	77
2	SB		38	64	63	60	51
2	EB		35	74	68	66	48
2	WB		17	67	69	68	54
3	NB		65	87	83	80	74
3	NB		75	78	74	72	58
3	SB		75	79	78	72	78
3	SB		71	72	71	66	62
3	EB		61	83	70	74	69
3	WB		56	71	68	65	69
4	NB		78	93	79	82	84
4	NB		65	79	72	78	68
4	SB		94	97	83	86	88
4	SB		58	80	69	79	72
4	EB		84	96	83	82	87
4	EB		84	91	85	83	75
4	WB		83	90	86	81	90
4	WB		87	93	89	86	75
5	NB		20	41	37	43	36
5	SB		22	44	37	46	50
5	EB		85	81	76	78	79
5	WB		58	83	76	77	84
MEAN			60.5	73.0	69.0	68.9	67.8
STD. DEVIATION			22.1	15.7	12.8	12.3	14.5

Table 4. Comparison of level of effort for eight methods.

Level of Effort	Method							
	U.S. HCM	British	Swedish	TRB Circular 212			NCHRP	
				Planning	Operations and Design	Australian	Planning	Operations
Time needed to learn (h)	4	8	20	2	4	20	1	15
Time needed to apply (h)	1	1	4	1/2	1	3	1/2	3

Figure 4. Results for LOS.

LOCATION				LEVEL OF SERVICE				
INTERSECTION NUMBER	DIRECTION OF MOVEMENT	TRAFFIC MOVEMENT	MEASURED ^b	HCM	PREDICTED METHOD			
					CIRCULAR		NCHRP	
					PLANNING	OPERATIONS DESIGN	PLANNING	OPERATIONS
1	NB	↑↑↑	A	A	A	B	A ^a (A-C)	B
1	SB	↑↑↑	A	A	A	B	"	B
1	EB	↑↑↑	C	E	A	B	"	D
1	WB	↑↑↑	C	B	A	B	"	C
2	NB	↑↑↑	B	A	A	A	A ^a (A-C)	D
2	SB	↑↑↑	B	A	A	A	"	B
2	SB	↑↑↑	A	A	A	A	"	B
2	EB	↑↑↑	B	B	A	A	"	A
2	WB	↑↑↑	A	B	A	A	"	A
3	NB	↑↑↑	C	D	B	B	A ^a (A-C)	D
3	SB	↑↑↑	B	D	B	B	"	B
3	SB	↑↑↑	C	A	B	B	"	D
3	SB	↑↑↑	B	A	B	B	"	B
3	EB	↑↑↑	C	E	B	B	"	D
3	EB	↑↑↑	C	E	B	B	A ^a (A-C)	B
3	WB	↑↑↑	B	A	B	B	"	D
3	WB	↑↑↑	B	A	B	B	"	B
4	NB	↑↑↑	D	E	B	C	C ^a (A-C)	E
4	NB	↑↑↑	C	E	B	C	"	C
4	SB	↑↑↑	D	C	B	C	"	F
4	SB	↑↑↑	C	C	B	C	"	C
4	EB	↑↑↑	D	E	B	C	"	E
4	EB	↑↑↑	E	E	B	C	"	D
4	WB	↑↑↑	D	C	B	C	"	E
4	WB	↑↑↑	D	C	B	C	"	E
5	NB	↑↑↑	A	A	A	A	A ^a (A-C)	A
5	SB	↑↑↑	A	A	A	A	"	A
5	EB	↑↑↑	D	A	A	A	"	D
5	WB	↑↑↑	B	A	A	A	"	D

a) Using the criteria in the "old" report of NCHRP 3-28 Project.

b) Measured values are derived from measured stopped delay values based on the NCHRP 3-28 criteria.

preliminary design, and signal-timing applications. The five U.S. methods all provided LOS results. All eight methods could be used in preliminary design investigations. The major difference between methods was in signal-timing applications. The British and the Swedish methods provided considerable flexibility with signal-timing investigations. However, the Australian and the NCHRP operations methods provided the greatest flexibility with signal timing. The NCHRP operations method was especially user-oriented; it provided step-by-step instructions and accompanying work sheets.

The criterion of accuracy of predictions provided a very quantitative comparison between methods. Not all eight methods provided numerical results that could be compared with the various measurements taken in the field. Therefore, it was necessary to use different measurements to compare different methods. Lane saturation flow was used to compare the HCM, British, Swedish, Australian, and NCHRP operations methods. Percentage of delay and vehicles stopped was used to compare the British, Swedish, Australian, and NCHRP operations methods. LOS

was used to compare the five U.S. methods. The comparison between these measurements and predictions will be discussed in the next four paragraphs.

The results for lane saturation flow were presented in Figure 1. The measured and predicted lane saturation flows are shown for 13 lane groups and the calculated mean and standard deviation are indicated at the bottom of the column for each method. In the following analyses, it is assumed that the measured values are correct and any differences between the predicted and measured values are considered to be errors in the predicted methods. Statistical analyses were performed and their results are given in Table 5. The two best prediction methods for lane saturation flows were the NCHRP operations method and the Australian method. However, even these methods had mean errors of the order of 8-12 percent of the mean measured lane saturation flows.

The results for average total delay were presented in Figure 2. The measured and predicted average total delays are shown for 30 lane groups and the calculated mean and standard deviation are indicated at the bottom of the column for each method.

Statistical analyses were performed and their results are shown in Table 5. Average stopped delay was actually measured in the field and predicted by the NCHRP operations method. A factor of 1.3 [based on an earlier Federal Highway Administration (FHWA) study and also reported in the NCHRP study] was applied to the stopped delays in order to compare them with the other methods that are based on total average delay. The best prediction method for total average delay was the Australian method, followed by the British and the Swedish methods. The NCHRP method performed the least effectively and inspection of the regression plots indicated two subgroups of data points: separate left-turn lanes and other lane groups. Further linear-regression analysis of these two subgroups provided the following additional results:

Separate left-turn lanes:

$$d_{obs} = 0.74d_{pred} - 5.95 \quad \text{with } r = 0.91 \quad (1)$$

Other lane groups:

$$d_{obs} = 0.91d_{pred} + 2.48 \quad \text{with } r = 0.82 \quad (2)$$

These results clearly indicate that the major deficiency in predicting total average delay in the NCHRP method lies with separate left-turn lanes.

The results for percentage of vehicles stopped were presented in Figure 3. The measured and predicted percentages of stopped vehicles are shown for 30 lane groups and the calculated mean and standard deviation are indicated at the bottom of the column for each method. Statistical analyses were performed and their results are shown in Table 5. The NCHRP operations method was slightly better than the other three methods, which were about equal. However, all four methods exhibited mean errors of the order of 20-25 percent of the measured mean of percentage of stopped vehicles.

The LOS results were presented in Figure 4. Because of the integer nature of LOS, the comparative analyses were performed in a slightly different manner, as given in Table 6, in which frequencies of differences in LOS between the indicated method and the field results are displayed. For example, if we consider the HCM method, the LOS of 11 of the 30 lane groups was predicted to be one LOS higher than that obtained in the field. The NCHRP operations method predicts slightly lower (worse) LOS than field conditions, whereas the other four methods predict one LOS higher (better) than field conditions. The percentages of the frequencies for each method are tabulated below. Three levels of percentages are given: predicted LOS equals field results, predicted LOS within plus or minus one LOS of field results, and predicted LOS within plus or minus two LOS of field results. The TRB Circular 212 operations and design method and the NCHRP operations method represented field conditions slightly better than the other three methods.

Level	Method				
	U.S. HCM	Plan- ning	Opera- tions and Design	NCHRP Plan- ning	Opera- tions
Predicted equal to field (%)	27	33	33	27	40
Predicted within ±1 LOS of field (%)	77	66	94	74	83

Level	Method				
	U.S. HCM	Plan- ning	Opera- tions and Design	NCHRP Plan- ning	Opera- tions
Predicted within ±2 LOS of field (%)	97	93	97	97	100

CONCLUDING REMARKS

This final section contains some concluding remarks about the limitations of the study, significant results, and future research directions.

There were a number of factors that limited the effectiveness of this study. All data sets were obtained at intersections located in the San Francisco Bay Area and did not include locations elsewhere in the United States or abroad. Although the characteristics of the five intersections varied, a greater variety of intersections and traffic characteristics would have enhanced the study. Additional traffic-performance measures such as total delay, load factor, and queue length should have been obtained in the field in order to more fully evaluate prediction methods. Greater precision in the field measurements and more lane groups would have strengthened the statistical analysis and its interpretation. Finally, some of the procedures in some of the methods were not explicit, which caused the users to apply judgment and not always get identical results.

The eight methods were compared on the basis of their cost-effectiveness. The NCHRP operations method and the Australian method were found to be the most cost-effective. The other methods were about equal in their cost-effectiveness.

Table 7 attempts to summarize the most significant results of this comparative study by assessing the level of effort required and quantitative or qualitative effectiveness. In regard to level of effort required, the eight methods were classified into three groups: least effort, moderate effort, and most effort. In regard to quantitative or qualitative effectiveness, the comparisons are more difficult because of the variety of ways effectiveness could be measured and because of strengths and weaknesses in the various methods. The NCHRP operations method and the Australian method were found to be the most effective, followed by the Swedish and British methods. The HCM method was the next most effective, followed by the two TRB Circular 212 methods and the NCHRP planning method. The NCHRP planning method was found to be an acceptable method when level of effort available is limited and only overall intersection evaluation is needed.

Future research should have two major directions. First, this comparative study should be extended to eliminate or at least reduce the previously identified limitations. These include broadening the environments and intersection types, increasing the number of lane groups, including additional field traffic-performance measurements, and providing greater field measurement precision. Second, the two NCHRP methods should be enhanced. Some deficiencies have been identified in this study and it is suggested that user surveys be undertaken to identify other possible difficulties. The major research effort should be directed toward increasing the accuracy and reducing the level of effort required by the NCHRP operations method. Although some improvement in accuracy of predicting lane saturation flow and percentage of stopped vehicles is

Table 5. Statistical analysis of results for lane saturation flow, total average delay, and percentage of stopped vehicles.

Calculation	Lane Saturation Flow					
	Measured	Prediction Method				
		U.S. HCM	British	Swedish	Australian	NCHRP
Overall mean value	1491	1393	1907	1692	1576	1522
Significant difference between measured and predicted means (0.05 level)	-	No	Yes	Yes	No	No
SD	182	229	309	211	211	137
Significant difference between measured and predicted SDs (0.05 level)	-	No	No	No	No	No
Mean error	-	249	424	236	174	117
Root-mean-square error	-	280	526	280	208	147
No. of high estimates	-	4	11	11	9	7
No. of low estimates	-	9	2	2	4	6
Linear regression						
Y-intercept	-	+1333	+1212	+792	+807	+293
Slope	-	+0.113	+0.146	+0.413	+0.434	+0.787
Correlation coefficient (r)	-	0.14	0.25	0.48	0.50	0.59
Significant difference from slope = 1.00	-	Yes	Yes	Yes	Yes	No
Significant difference between r and r = 0	-	No	No	No	No	Yes

Table 6. Analysis of results for LOS.

LOS Difference	Method				
	U.S. HCM	TRB Circular 212		NCHRP	
		Planning	Operations and Design	Planning	Operations
+3	1	2	1	1	0
+2	1	8	1	7	0
+1	11	10	16	14	3
0	8	10	10	8	12
-1	4	0	2	0	10
-2	5	0	0	0	5
-3	0	0	0	0	0

Table 7. Summary of comparative analysis.

Item	Method							
	U.S. HCM	British	Swedish	TRB Circular 212		Australian	NCHRP	
				Planning	Operations and Design		Planning	Operations
Level of effort required	Moderate	Most	Most	Least	Moderate	Most	Least	Most
Quantitative/qualitative effectiveness								
Degree of disaggregation	2	2	3	1	1	3	1	4
Capacity-performance outputs	2	3	3	1	1	4	1	3
Flexibility of use	2	3	3	1	1	4	1	4
Accuracy								
Lane saturation flows	2	2	2	NA	NA	3	NA	4
Total avg delay	NA	3	3	NA	NA	4	NA	2
Percentage of stopped vehicles	NA	2	2	NA	NA	2	NA	3
LOS	2	NA	NA	2	3	NA	2	3

desired, the greatest inaccuracies were in predicting delay. A limited investigation revealed that the magnitude of errors in estimating delay was greatly influenced by type of lane group and especially delays associated with exclusive left-turn lane groups. One of the major objections to the NCHRP operations method is the level of effort required. The primary ways of reducing the level of effort required are through improved use of default values, employment of nomographs, and/or computerization.

ACKNOWLEDGMENT

A special word of appreciation is extended to NCHRP

and to JHK and Associates for their approval and their provision of the five intersection data sets for use in this study. Robert Spicher, William Reilly, and Stephen Bolduc were particularly helpful and supportive of this work.

These analyses of the eight methods applied to the five intersection data sets were first undertaken in a graduate course at the University of California in early 1982. We would like to acknowledge the contributions made by the following graduate students in that class: S. Au, J. Jen, S.Y. Wong, G. Patrick, Y.J. Sun, H.J. Stander, M. DeRobertis, T. Folks, J. Bierstedt, J.H. Rhee, J. Rupe, and J. Kahng.

Total Avg Delay					Percentage of Stopped Vehicles				
Measured	Prediction Method				Measured	Prediction Method			
	British	Swedish	Australian	NCHRP		British	Swedish	Australian	NCHRP
26.5	20.4	19.7	21.3	32.3	60.5	73.0	68.9	68.9	67.8
-	Yes	Yes	Yes	Yes	-	Yes	Yes	Yes	Yes
16.8	13.0	8.0	10.4	17.4	22.1	15.7	12.8	12.3	14.5
-	No	Yes	Yes	No	-	Yes	Yes	Yes	Yes
-	8.4	9.3	7.6	12.2	-	15.8	12.7	14.2	12.6
-	12.4	12.4	11.6	15.9	-	19.8	16.5	18.7	15.5
-	9	10	11	20	-	25	22	19	21
-	21	20	19	10	-	5	8	11	9
-	+6.5	-8.2	-1.4	+7.3	-	-12.25	-24.46	-18.85	-21.28
-	+0.98	+1.76	+1.32	+0.59	-	+0.998	+1.234	+1.153	+1.207
-	0.76	0.83	0.82	0.61	-	0.70	0.71	0.64	0.79
-	No	Yes	No	Yes	-	No	No	No	No
-	Yes	Yes	Yes	Yes	-	Yes	Yes	Yes	Yes

We also wish to thank Louis Torregrosa, David Kuan, and Tsutomu Imada for help in the final preparation of this paper.

REFERENCES

1. A.D. May. Intersection Capacity 1974: An Overview. TRB, Special Rept. 153, 1975, pp. 50-59.
2. Highway Capacity Manual. TRB, Washington, DC, 1950.
3. Highway Capacity Manual--1965. TRB, Special Rept. 87, 1965.
4. F.V. Webster. Traffic Signal Settings. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, TRRL Tech. Paper 39, 1958.
5. F.V. Webster and B.M. Cobbe. Traffic Signals. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, TRRL Tech. Paper 56, 1966.
6. B.E. Peterson and E. Imre. Beräkning av kapacitet, kölängd, fördröjning i vägtrafik-anläggningar (Swedish Capacity Manual). Statens Vägverk, Stockholm, Sweden, 1977.
7. K.-L. Bang. Capacity of Signalized Intersections. TRB, Transportation Research Record 667, 1978, pp. 11-21.
8. Interim Materials on Highway Capacity. TRB, Transportation Research Circular 212, 1980.
9. R. Akcelik. Traffic Signals: Capacity and Timing Analysis. Australian Road Research Board, ARRB Res. Rept. 123, March 1981.
10. JHK and Associates. Urban Signalized Intersection Capacity. NCHRP [Project 3-28(2)], unpublished, May 1982.

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

Speeds and Flows on an Urban Freeway: Some Measurements and a Hypothesis

V.F. HURDLE AND P.K. DATTA

Speeds and flows were measured in a bottleneck section of a six-lane urban freeway near Toronto, Canada, on three successive mornings. The average capacity flow was 1984 passenger-car units per lane per hour, very close to the traditional value of 2000, but at an average speed of 80 km/h (49 mph), a much higher speed than is usually indicated in textbooks and manuals. Frequency distributions of the observed flows and speeds at capacity are reported and used as a part of a general discussion of the meaning of the term "capacity." In order to study the relationship between speed and flow, measurements were also made before the section reached capacity. At flows less than three-fourths of capacity, the average speed was 100 km/h (62 mph); there was no apparent decrease in speed as the flow increased. Between three-fourths of capacity and capacity, a gradual reduction in speed from 100 km/h to the 80-km/h speed observed at capacity was expected, but no such smooth speed transition was observed. The nature of the data leads to the hypothesis that the average speed on an urban freeway with a speed limit, where neither grades nor curvature is severe and where the traffic is not affected by downstream bottlenecks,

may not vary as a function of flow but may depend only on whether the traffic is or is not a capacity flow discharged from an upstream queue.

A good understanding of the way in which speed varies with flow is an essential prerequisite to the creation and use of any level-of-service concept for freeways. Unfortunately, misinformation about this relationship abounds. In this paper, we present some data and some ideas that we hope will help to combat some of the misinformation. As we studied the data, however, we found ourselves questioning not just the things we had intended to challenge, but also some of the things we ourselves believed. What was intended to be a straightforward presenta-

Figure 1. Typical speed-flow relationship.

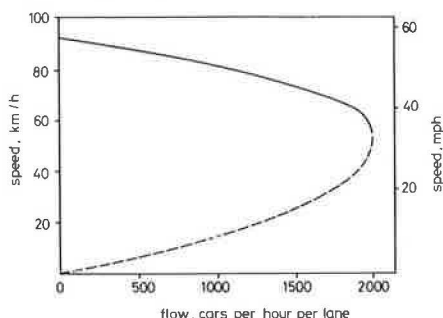


Figure 2. Speed-flow relationships and data points.

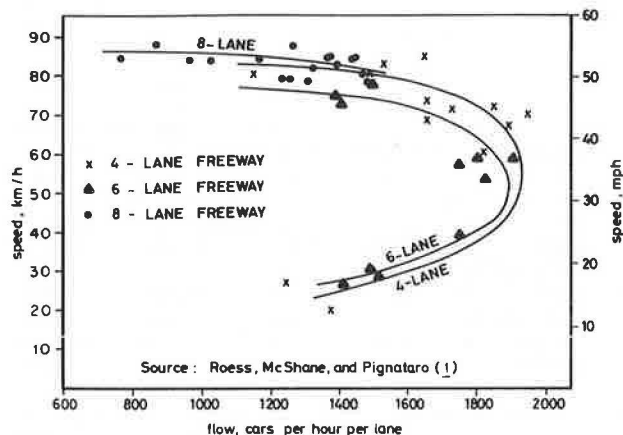
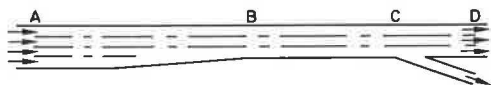


Figure 3. Simple freeway bottleneck.



tion of a simple empirical study gradually became instead an essay that asks more questions than it answers.

As an introduction to both the data set and the unanswered questions, consider Figure 1. The speed-flow relationship shown there is not based on any data set but is our impression of conventional wisdom, the sort of curve one often finds in books and papers or on the blackboard in transportation engineering classes.

That Figure 1 is not entirely correct is immediately obvious. The upper branch of the speed-flow relationship cannot have a negative slope at all points but must be horizontal at very low flows; anyone who drives knows that the first car on a multilane road does not slow down just because a second car appears. It would also seem to be obvious that the speed does eventually drop. What is not obvious, however, is how large the flow becomes before the drop begins or how far the speed drops. Some theorizing may be possible, but these questions must ultimately be answered by making measurements.

Over the years, many measurements have been made and various curves have appeared in the literature or have been circulated privately. Some indicate a rapid drop-off in speed near capacity, whereas others do not; some show a single, continuous curve, whereas others show the upper and lower (dashed)

branches as separate curves. No attempt will be made here to summarize this large number of reports (although we will pause to bemoan the fact that many are of limited use because detailed information about the locations and traffic conditions is not readily available). One recent report (1), by Roess, McShane, and Pignataro, is of particular interest here, however, because it shows high speeds at flows of at least three-fourths of the roadway capacity. A speed-flow relationship from this report is reproduced in Figure 2.

This curve is a major step in the right direction; we hope it will lay to rest forever the myth that speeds on North American freeways vary as a function of flow even at moderate flows. [On freeways without speed limits or with speed limits well above 100 km/h (62 mph), the situation may be quite different. Thus the applicability of some of what we say must be limited to the North American scene, where the fact that it becomes increasingly difficult to drive at very high speeds as the flow increases is not a serious issue.] However, we were still not satisfied when this curve appeared. We did not believe that capacity flows on urban freeways typically move at speeds of the order of 50 km/h (30 mph) but rather that they move at something like 70-80 km/h (45-50 mph). This disbelief was based primarily on the personal experiences of one author as an engineer and on extensive examination of speed and density contour charts for California freeways [2; reports on various Los Angeles and San Francisco area freeways by the California Department of Transportation (Caltrans)]. The Ontario data in this paper were gathered in the hope that they would confirm this view. This they do, but they also raise some new questions that we had not foreseen.

SOME PRELIMINARIES

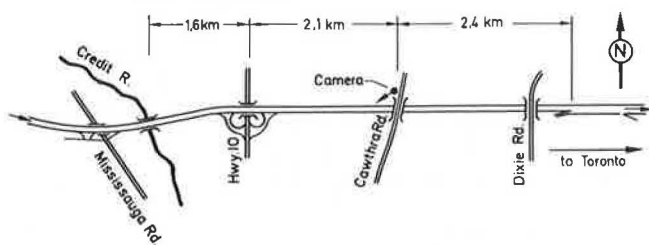
Before the data are presented, it is essential that authors and readers agree on just what they represent or at least that the readers understand the authors' intentions. Therefore it is necessary to explain what we mean by capacity, speed-flow curves, etc. We shall do this by considering what happens at the very simple freeway bottleneck shown in Figure 3. Traffic enters from the left at some rate λ . Between B and C, however, the freeway is narrower, so it is possible that λ vehicles per unit time cannot get through but only some smaller flow, say, 3μ vehicles per unit time, where μ is the average of the capacities of the three individual lanes.

The remaining $\lambda - 3\mu$ vehicles per unit time will queue up behind point B and await their turns to pass through the bottleneck. Needless to say, they will not move at high speeds while waiting but will travel very slowly or in a stop-and-go fashion and the drivers and passengers will complain about the congestion. [This condition is known as level-of-service F (2-4).]

Much of what would happen on this stretch of freeway can be represented by a speed-flow curve like Figure 1 if we measure the flows in vehicles per lane per hour rather than in vehicles per hour. Then the flow at the nose of the curve must be μ , the capacity of one lane, and the lower, dashed curve obviously represents the congested conditions encountered within the queue.

The conditions in section BC, the bottleneck, can be anywhere on the upper part of the curve but never on the lower (unless an accident or other incident occurred downstream, in which case BC would no longer be the controlling bottleneck). Conditions in section AB, on the other hand, can be on either branch, depending on whether the point observed is

Figure 4. Location of observations.



within or behind the queue, but can never lie to the right of $3u/4$ on the lower branch, since it is physically impossible for the average total flow within the queue to exceed the bottleneck capacity ($3u$).

It is, of course, possible to observe average lane flows between $3u/4$ and u behind the queue in section AB but not for very long, since the queue would then grow and the conditions at the observation point would abruptly change to the lower branch of the curve as soon as the end of the queue arrived.

It should be clear from the preceding discussion that it is very difficult to obtain enough data at any one observation point to plot the entire speed-flow curve. It would seem, in fact, to be virtually impossible unless one regulated the flow by metering ramps both upstream and downstream. In the simple experiment described in this paper, the observations were all taken within the bottleneck section, so all the data points lie on the upper branch of the curve. Furthermore, the flow increased from half of capacity to capacity in about half an hour, so each day of observation provided rather little data at flow levels approaching capacity, the most interesting region of operation.

The two branches of the speed-flow curve in Figure 1 join to form a single, smooth curve. It should be noted, however, that the nose of the curve is not necessarily smooth nor even continuous. The two branches may be quite separate curves, one the result of driver behavior while in queue and the other the result of quite different behavior when not in queue (5-8).

Whether the two branches form a single, smooth curve is not an issue we wish to address in this paper, but it should be noted that even if they do, one should not expect to see a sequence of observations that traverses around the nose of the curve (7,8). Such a sequence would represent a gradual transition between conditions within and outside of a queue at capacity or near-capacity flows. This is something that could happen on a real roadway only under very extreme conditions, if at all. Certainly it could not happen in the kind of uncontrolled experiment described here.

A second point to be noted is that the data points cannot be expected to lie nicely along some smooth curve but will be scattered about any curve one may draw. Thus, the speed-flow curves we are talking about are really some kind of average. In this paper, we shall presume that the scatter is entirely stochastic, the result of differences between the individual drivers and the vehicles on the roadway. In fact, however, there are probably also systematic departures from "average" behavior because of differences between increasing and decreasing flow conditions, cyclic fluctuations in operating conditions, etc. We shall not worry about such things in this paper; our interest here is the general shape of speed-flow curves and the magnitude, rather than the pattern, of variations in ca-

capacity flows. Systematic departures from average behavior may, however, be important to those whose work requires more detailed knowledge of freeway flows. Conversely, the stochastic variations may be sufficiently large to make such detailed knowledge very difficult, if not impossible, to obtain.

DATA GATHERING

The data presented in this paper were collected on the Queen Elizabeth Way, a six-lane freeway in Mississauga, Ontario, Canada (9). All data are for eastbound, weekday morning traffic passing beneath Cawthra Road on May 25, 26, and 27, 1977, bound for Toronto. This location, shown in Figure 4, is near the middle of a long bottleneck section that begins at the on ramp from Highway 10 and ends near Dixie Road, where a fourth eastbound lane is added. The lanes are 3.66 m (12 ft) wide with ample side clearance, grades are minimal, there is no horizontal curvature, and the speed limit is 100 km/h. All observations were made in good weather.

The Highway 10 on ramp is metered but not severely enough to completely eliminate upstream congestion. There is also a secondary bottleneck upstream, where the freeway crosses the Credit River, that sometimes meters the flow sufficiently to keep flows at the observation site below capacity.

Neither the metered ramp nor the secondary bottleneck was desirable for this sort of experiment, since we wanted to observe capacity flows. However, sensors operated just downstream from the Credit River Bridge by the Ontario Ministry of Transportation and Communications usually indicate speeds less than 32 km/h (20 mph) for at least an hour during the morning peak, which indicates that a queue almost always builds up behind the bottleneck where our measurements were taken. Thus, since the flow patterns observed on the three days of data gathering were very similar, we feel confident that we did measure capacity flows. However, future experiments of this type should include observation of upstream and downstream traffic, so that there can be no doubt as to whether the flows observed are capacity flows.

All observations were taken from the bridge or bridge approaches at Cawthra Road. All flow data are based on 2-min counts made by two observers located above the freeway lanes on the sidewalk along the west side of Cawthra Road. These observers were clearly visible to drivers, but they were not particularly conspicuous and the presence of pedestrians on the bridge is not unusual, so we do not feel that the presence of observers had an appreciable effect on freeway speeds. If there was an effect, it would almost certainly have been to slow traffic down, an effect that does not seem likely in light of the experimental results.

Speeds were measured by time-lapse photography by using a 35-mm camera equipped with a motor drive and intervalometer. The camera was set up on the abutment fill at the north end of the bridge, a location where very few drivers would notice it. In order to keep film and data-processing costs low, we took a sequence of four photographs every 2 min. The individual photographs were taken 2.63 s apart. Once the flow reached capacity, the 2-min sampling interval was changed to 5 min; this allowed us to record data from 6:00 to 9:00 a.m. on a single film cassette 10 m (33 ft) long.

The photographs were projected onto a grid and the speed of every vehicle that appeared on the grid in at least two successive pictures was measured. The gridded zone was approximately 88 m (270 ft) long with grid lines at 3-m (10-ft) spacings. The arithmetic mean of the speeds of all the vehicles

caught in a single series of four pictures was associated with the count during the 2-min interval immediately preceding, or including, the time when the pictures were taken to give a single data point for the speed-flow relation. The plotted speed was usually the average of the speeds of four to six vehicles but sometimes the speed of only one vehicle at low flows and the average speed of as many as nine vehicles at high flows. It is a space-mean speed rather than the time-mean speed obtained by most sampling methods.

This procedure kept the film cost low and avoided the problem of losing data during film changes. However, more frequent sampling of speed would be desirable.

No formal check on the accuracy of the speed measurements was made. However, a car was driven through the section at various times on another day at speeds consistent with the data. Speeds measured by detectors at other locations on the Queen Elizabeth Way at moderate to high flow levels are also consistent with our data, so we have no reason to believe there was serious error in the speed-measurement procedure. In any case, errors of a few kilometers per hour would not affect the nature of our conclusions unless the error was different at different flows, a problem unlikely to occur with our procedures.

Our primary interest at the beginning of this research was speed, not capacity, so trucks were not counted separately. Later, when we examined the data, we decided that we would like to be able to express the flows in passenger-car units. This was accomplished by counting cars and trucks on the film. Since this is a sampling procedure rather than a continuous count and because there was a very noticeable variation in the truck percentage with time of day, it was necessary to fit a curve to the observed data. The fourth-degree polynomial shown in Figure 5 is therefore only a rough approximation of the actual percentage of trucks, but it seemed adequate for the purpose of estimating equivalent passenger-car flows. Since the roadway is very nearly level, each truck was considered equivalent to two passenger cars (3,4).

FLOW MEASUREMENTS AND CAPACITY

The average flow during each 2-min count interval, averaged over the three days, is shown in Figure 6. In this figure and all that follow, the flow is expressed in equivalent passenger-car units per lane

per hour; the raw counts have been averaged over the three lanes and then adjusted on the basis that one truck is equivalent to two passenger cars (3,4). The flows observed on the individual days are shown in the lower parts of Figures 7, 8, and 9.

It can be seen that the flow increases steadily and quite rapidly from about 6:20 until shortly before 7:00 and then levels off and fluctuates about a mean of approximately 2000 PC/(lane·h) until nearly 9:00 a.m. This is what one would expect to see in a bottleneck section; the flow stops increasing when the capacity is reached, not because the demand levels off but because the roadway cannot carry any more traffic.

There is, of course, a good deal of fluctuation evident in the capacity flow observations. The capacity of a roadway is determined by the driving style of the individual drivers, so counts of capacity flows must necessarily be random variables. Some knowledge of their distribution would thus seem to be necessary if one is to understand what the word "capacity" means.

The distribution of the 120 flow observations made between 7:00 and 8:20, the period we feel is clearly and conservatively identifiable as a period of capacity flow, is shown in Figures 10 and 11. A normal distribution with the same mean and variance is also shown in Figure 11.

The mean capacity is 1984 PC/(lane·h); the individual means for the three days of observation are 1927, 2004, and 2020 PC/(lane·h). A 90 percent confidence interval for the mean capacity is 1953-

Figure 5. Approximation of truck percentage used to calculate equivalent passenger-car flows.

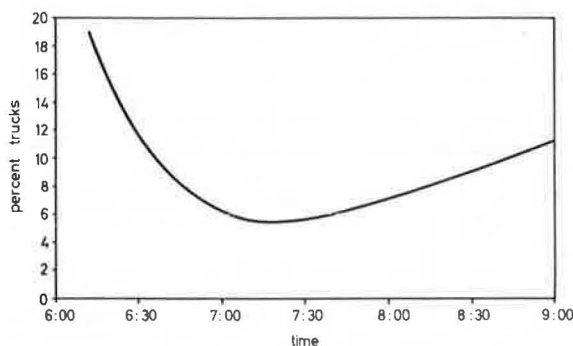
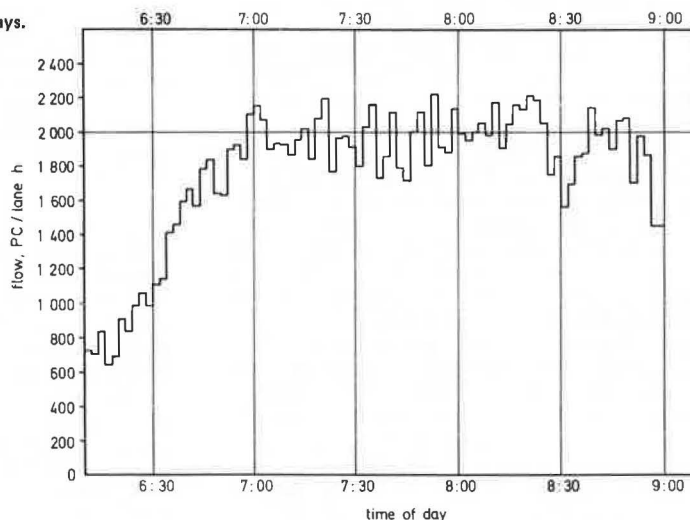


Figure 6. Average of flows measured on three successive days.



2015 PC/(lane·h), a range that includes the commonly accepted value of 2000.

It can be seen in Figures 10 and 11 that flows throughout the range from 1750 to 2200 PC/(lane·h) were frequently observed, whereas flows outside of

this range were relatively rare but did occur. It is important to understand that the shape of the distribution was determined not only by the characteristics of the roadway and the drivers, but also by the way in which we made our counts. We chose to

Figure 7. Speeds and flows observed on Wednesday, May 25, 1977.

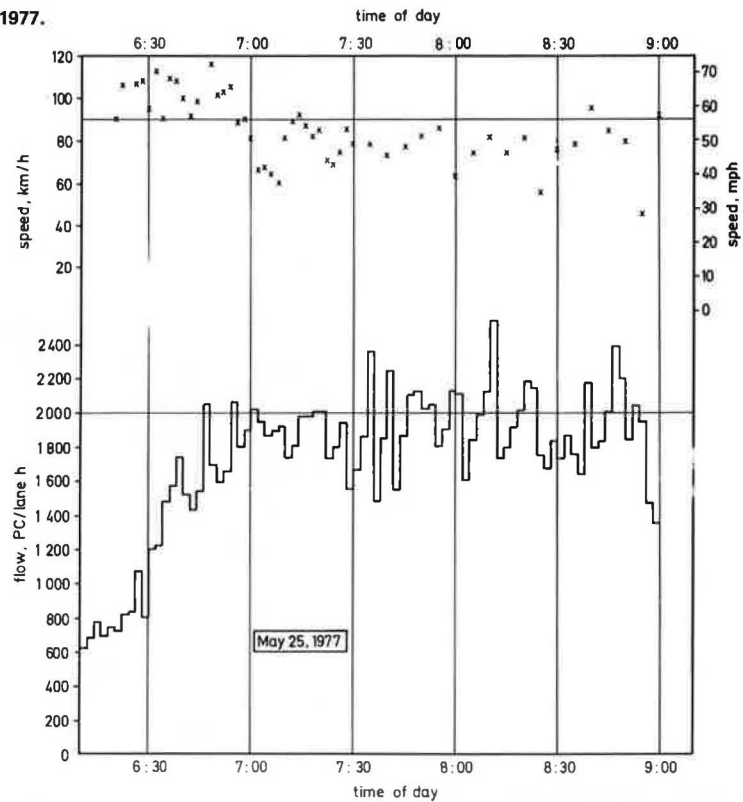


Figure 8. Speeds and flows observed on Thursday, May 26, 1977.

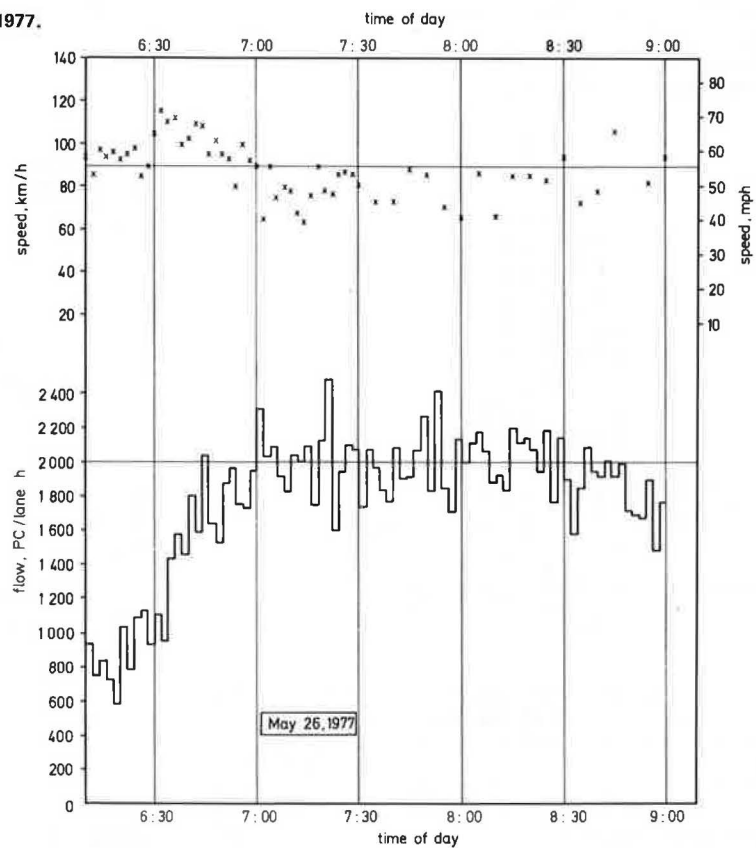


Figure 9. Speeds and flows observed on Friday, May 27, 1977.

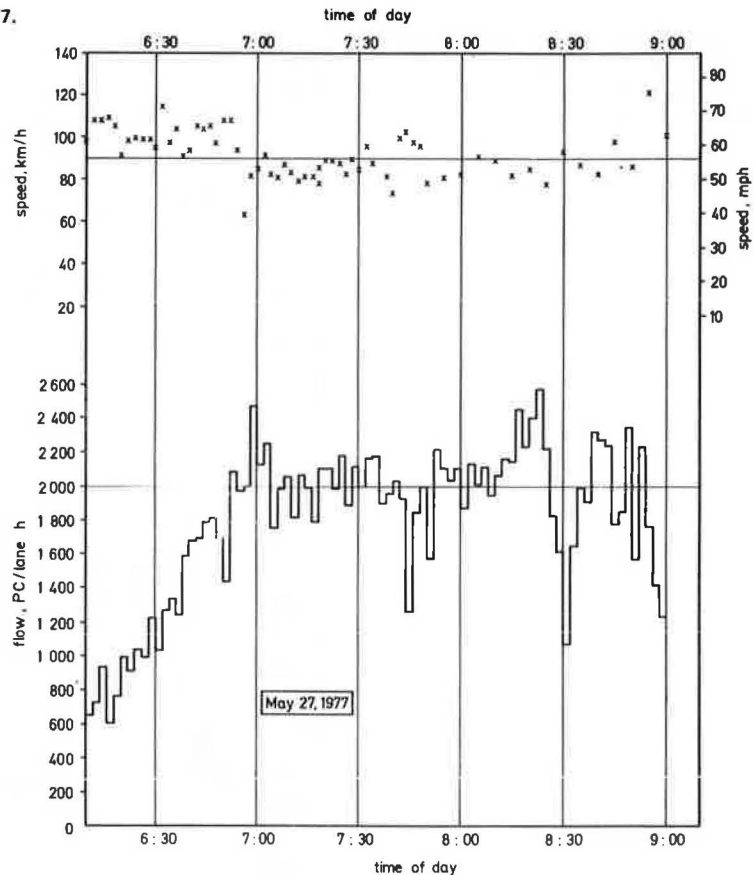
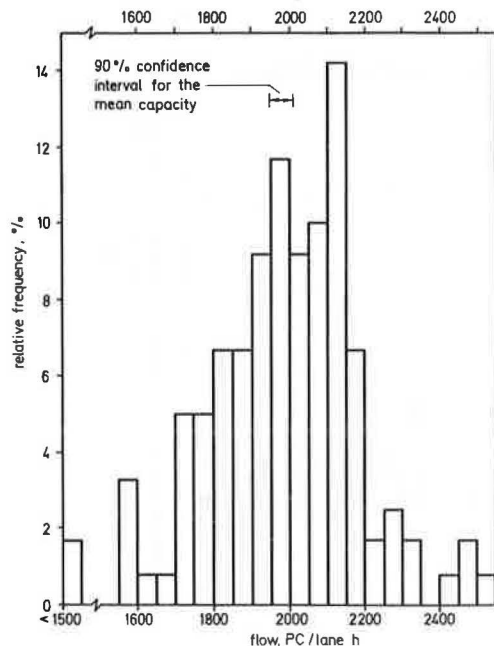
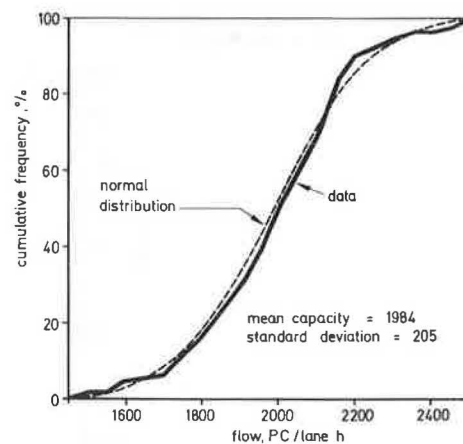


Figure 10. Frequency histogram for capacity flows.



make 2-min counts and then computed the average flow over each 2-min interval by multiplying the count by 30. (We also made a truck correction, which increased the variance-to-mean ratio of the calculated flows by 11 percent. For the sake of simplicity, the effect of that adjustment will be ignored in the discussion that follows.) Had we instead made 1-min

Figure 11. Cumulative frequency polygon for capacity flows.



or 5-min counts, the average capacity flow would still have been 1984 PC/(lane·h), but Figure 10 would have had a quite different appearance.

If the counts made during different time intervals are assumed to be independent random variables from the same distribution—a reasonable assumption for capacity flows—it is easy to see the effect of changing the length of the count interval. We made 2-min counts and found that the standard deviation of the flows thus measured was 205 PC/(lane·h). Had we made our counts m times as long, one would expect the standard deviation of the flows to be $1/(m)^{1/2}$ as great, as shown in the second column of Table 1. The implications of this fact are apparent in the right-hand side of Table 1, where we

have also assumed that the counts are normally distributed. Here it can be seen that very high flows will be observed rather often if the counting interval is short but almost never if it is long. For example, the average flow would exceed 2200 PC/(lane·h) in about one out of every four 1-min counts but in only one out of every twenty 5-min counts, one out of every five hundred 15-min counts, and in less than one out of every ten thousand 0.5-h counts.

We have defined capacity as the average flow through a bottleneck when a steady supply of cars is assured by the existence of an upstream queue. We think this is a useful definition and that it is compatible with the definition in the Highway Capacity Manual (3). However, it is clearly not the only possible definition. One might say that the capacity is the highest flow ever observed or perhaps the 90th-percentile flow. However, it is clear from Table 1 that such definitions are only usable if the length of the counting interval is specified. Even if a suitable length could be agreed on, we doubt that these are useful definitions.

At the lower end of the distribution, however, there is a more interesting possibility. One might define a "practical" capacity as the flow that will manage to get through at least P percent, perhaps 90 or 95 percent, of the time. Again, the length of the count interval must be specified, but it is easy to see that someone designing a traffic control system might be interested in such a number. Unfortunately, it is also easy to see that a designer interested in short time intervals must either choose a value considerably lower than the average capacity or run a high risk of system failure.

Still another possible definition of capacity is based on the idea that considerably higher flows are

possible before a queue forms than can be maintained after it forms. [See, for example, Interim Materials on Highway Capacity (4, p. 256).] Under such a definition, the capacity is that higher flow, not the average flow out of the queue that we have called the capacity.

While this is a conceptually sound definition of capacity, we do not feel that it is a good definition. In the first place, we do not see the usefulness of a capacity that can only be used for an extremely short period of time (unless one can implement far more delicate traffic control schemes than are now in use). In the second place, we question the existence of these considerably higher flows. Certainly they are not evident in Figures 6, 7, and 8. Figure 9 does show a very high flow just before 7:00, but similar flows were observed later as well. To really know whether such higher flows occur prior to the formation of a queue, one would have to have many days' data, not just three. One might also want to use a shorter count interval than 2 min. On the basis of the data presented here, we can only say that we see no reason to be believers.

SPEED-FLOW RELATION

Figure 12 shows the average speed and the corresponding flow for each of the 144 observations made between 6:10 and 8:20 on the three days. As was expected, the speed seems to be virtually constant until the flow reaches at least 1500 PC/(lane·h). In fact, if one tries to fit a straight line to the data by linear regression, the slope turns out to be positive, not only for flows up to 1500 PC/(lane·h), but even if all flows up to 1700 PC/(lane·h) are included. For flows above 1850 PC/(lane·h), however, the speeds are clearly lower; in only one of the 78 observations at a flow of more than 1850 PC/(lane·h) was the speed as high as 100 km/h, the speed limit and the average speed observed when the flow was less than 1700 PC/(lane·h).

This research was undertaken with the primary purpose of studying the pattern of speed reduction that occurs as freeway flows increase. One might reasonably expect such research to lead to a speed-flow curve, but none appears in Figure 12. One reason is that we want readers to examine the data themselves, uninfluenced by any curve we might draw. There is, however, another more basic reason: We simply do not know how to draw a curve that satisfactorily represents the data shown in Figure 12.

To illustrate the difficulties, we have drawn four candidate curves in Figure 13. The lower, congested branch has been omitted in each case,

Table 1. Effect of counting-interval length on observed frequencies of high flows.

Interval Length (min)	SD of Flow Measurements	Percentage of Flow Measurements Exceeding Q PC/(lane·h)			
		Q = 2100	Q = 2200	Q = 2300	Q = 2400
0.5	410	39	30	22	15
1	290	34	23	14	8
2	205	28	15	6	2
5	130	18	5	0.7	0.07
10	92	10	0.9	0.03	x ^a
15	75	6	0.2	x	x
30	53	1.4	x	x	x
60	37	0.1	x	x	x

^ax indicates less than 0.01 percent.

Figure 12. Speed and flow observations.

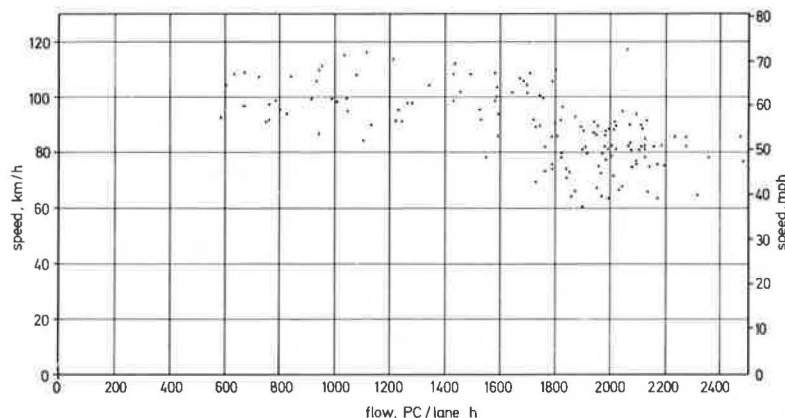
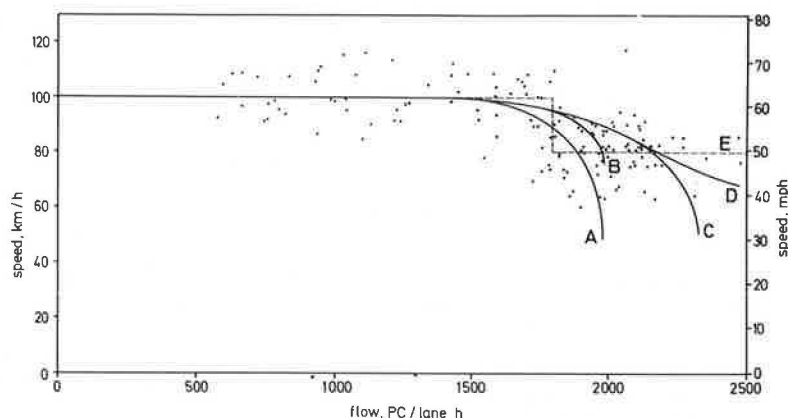


Figure 13. Some possible shapes for a speed-flow curve.



since we have no data for congested flows. Curves A, B, and C are all of the same general shape as those in Figure 2: a horizontal line up to about three-fourths of capacity and then a sharp bend. Each curve becomes vertical at some flow that might be regarded as the capacity. For curves A and B, this capacity is the average capacity flow observed in the experiment; i.e., it is the capacity of the roadway as we have defined the word. In curve C, on the other hand, a higher capacity has been used, a flow closer to the maximum flow that was observed.

Obviously none of these three curves adequately represents the data; an analyst with no a priori ideas about the shape of the curve would surely draw a curve more like D. This did not bother us; our prejudices favored a curve like D anyway.

What did bother us, however, is the fact that curve D does not reflect the nature of the data very well either. If one examines curve D closely, it becomes apparent that the curve systematically predicts speeds higher than those actually observed at flows between 1830 and 2000 PC/(lane·h). There are 30 observations in this range but not a single one lies above the curve. Furthermore, if one were to lower the curve enough to obtain reasonable predictions in this range, the resulting curve would systematically predict speeds that were too slow at flows greater than 2000 PC/(lane·h) and/or less than 1700 PC/(lane·h). Ultimately, we came to the conclusion that no reasonably smooth curve does an adequate job of representing the data. If one must have a curve, it must look more like "curve" E.

Curve E, in fact, fits the data quite well, certainly better than any conventionally shaped curve one might draw. This is disturbing; one must ask why it should be true.

A partial explanation lies in the way in which the data were gathered. Each speed measurement represents only what was happening during a particular 8 s within the 2-min count interval, not an average over the entire length of the count interval. In retrospect, we recognize this as a defect in our experimental procedure. Clearly we could have reduced the scatter in the data by averaging the speeds measured on several pairs of photographs taken at regular intervals throughout the 2-min count interval.

However, this defect in the experimental procedure is clearly not an adequate explanation for what happened. If the speed really dropped significantly as the flow increased, we should have observed more low speeds during intervals with high counts than during those with low counts. Except for the sudden drop in speed at about 1800 PC/(lane·h), Figure 12 shows absolutely no evidence of such a phenomenon. In fact, the average speed for the 42 observations

made during intervals with flows in excess of 2000 PC/(lane·h) was faster than the average for the 35 observations made during intervals with flows between 1800 and 2000 PC/(lane·h). This surprising result was almost certainly a matter of chance rather than a thing likely to be observed in all such experiments (though one could hypothesize that clusters of risk-prone drivers produce both high speeds and high flows). It seems very unlikely, however, that we would have obtained such a result if any of curves A, B, C, or D represented the true mean speed. That the true mean speed at capacity was in the neighborhood of 50 km/h (30 mph) seems even more incompatible with the data.

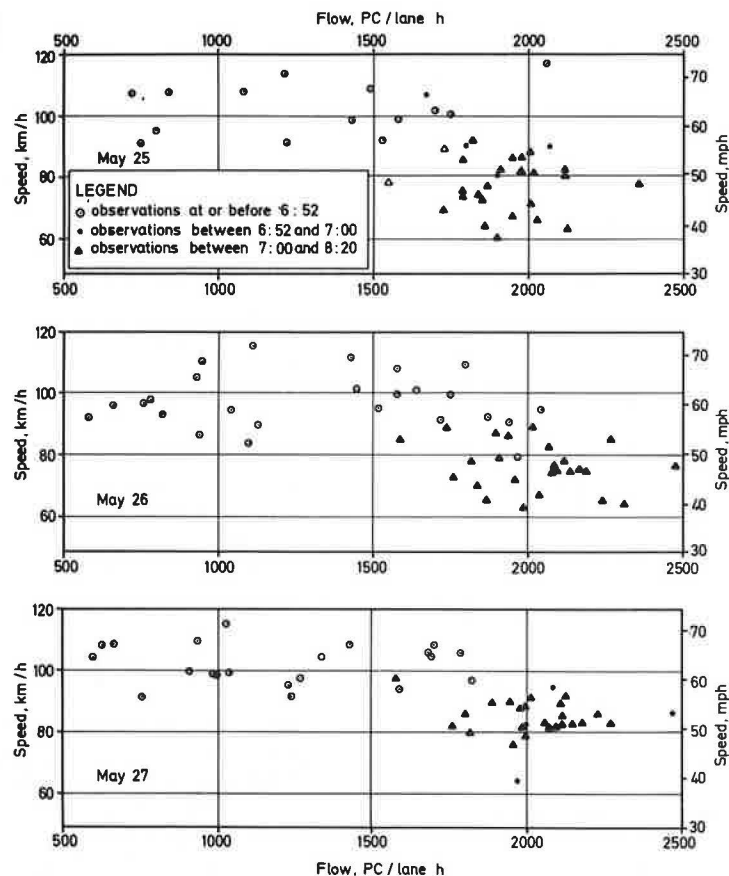
A HYPOTHESIS

Looking at Figure 13 and frustrated by the data's seeming misbehavior, we found ourselves drawn to a radical suggestion. Perhaps the whole notion that freeway traffic slows down as the flow increases is false. Maybe drivers who are able to approach the bottleneck at the speed limit just drive right on through at that speed, regardless of how high the flow may be, whereas those who have to slow down to wait in the queue only accelerate to about 70 or 80 km/h (45–50 mph) when they enter the bottleneck.

This hypothesis did not in fact arise from study of Figures 12 and 13 but from Figures 7, 8, and 9, where the observed speeds and flows are both plotted as functions of the time of day. In general, the pattern seems to be that the speed remained high until shortly before 7:00--about the time the flow reached capacity--and then dropped suddenly and remained lower until the queue vanished. (An obvious exception is the group of high speeds and some low flows observed between 7:40 and 7:50 on May 27. This group of observations could be the result of an upstream disturbance and hence not really representative of capacity flows, but we have no way of knowing.)

The hypothesis is explored further in Figure 14, where observations made at or before 6:52 are indicated by open circles and those made after 7:00 by solid triangles. The supposition is that the queue formed sometime between 6:52 and 7:00 on each of the three days, so the circles represent the situation before the queue formed and the triangles the situation after it formed. Observations made between 6:52 and 7:00 are represented by small dots; this period is not to be regarded as a transition period but as a period during which we are unable to make any statement about whether or not there was a queue. That the period of uncertainty is so long is the result of our data-collection methods, not of traffic conditions.

Figure 14. Speed and flow observations (those made while queue present shown as triangles).



The dichotomy works surprisingly well as a means of explaining the variation in speeds. On each of the three days, the triangles form a distinct group below the circles, with amazingly little overlap. If it were not for the speed of 97.5 km/h (58.5 mph) observed at 7:45 on May 27 at a flow of 1260 PC/(lane·h) and a few low speeds observed on May 26 at flows in the neighborhood of 1000 PC/(lane·h), the separation of the triangles and circles at 90 km/h (55 mph) would be very nearly total.

Particularly interesting is the small group of observations made before the queue formed but at flows in excess of 1700 PC/(lane·h), the open circles that lie above the triangles on the right-hand side of Figure 14. This group of points would seem to provide the key to testing the hypothesis that freeway bottleneck speeds are not a function of flow but only of whether there is or is not a queue up-stream. If the speeds of such points are of the same magnitude as those observed at lower flow levels and consistently higher than the speeds observed at high flow levels when there is a queue, then the hypothesis is a good one; otherwise it is not.

In our data set, there are only nine such points plus two at flows just under 1700 PC/(lane·h). One of these 11 points has a speed of 93 km/h (58 mph); the others are all faster than 95 km/h (59 mph). Thus all 11 are the sort of data points that tend to confirm the hypothesis.

To make a really convincing argument, one would need many more such data points. Unfortunately, they are difficult to obtain. The flow increases so rapidly that only a very few observations can be made on any given day at high flow levels in the absence of a queue. A first step in obtaining more such data points is to carefully observe upstream

conditions in order to positively identify the presence of a queue; one really cannot afford to throw away 8 min of prime data as we have done (the small dots in Figure 14). It is probably also better to observe in the evening rather than the morning; theory (10-13) predicts that evening flow levels should increase less rapidly. It would also be possible to increase the data yield by careful ramp metering, but this approach has its own problems. Not only is such tight control physically impossible in many locations, but the need for data at high flow levels both with and without an upstream queue is likely to be incompatible with the normal operating objectives of the metering system.

Still another difficulty is likely to arise if the hypothesis is tested on a freeway with an enforced 55-mph speed limit. Presumably, the no-queue speeds would then be lower, but the speeds downstream from queues would be similar to those we observed. Thus the two groups of points would overlap a great deal and it would be much more difficult to distinguish the hypothesized situation from one in which the average speed decreased with increasing flow in the way indicated by curve D in Figure 13. On the other hand, some locations may have somewhat lower queue discharge speeds than we observed. This would make the analysis easier. For example, the data (1) shown in Figure 2 might represent locations with no-queue speeds of about 85 km/h and queue discharge speeds of about 70 km/h.

AN ALTERNATIVE EXPLANATION

A second possible explanation for the sudden drop in speed at about 1800 PC/(lane·h) has been suggested to us by R. Wiedemann. The capacity at some downstream point within the bottleneck section may be

very slightly lower than at the point where our observations were made. If so, the triangles in Figure 14 could lie on the lower branch of the speed-flow curve rather than on the upper branch as we have presumed.

To decide between these two explanations, one would have to have data from at least one more point, at the downstream end of the bottleneck section, and we do not. The second explanation, however, conflicts with conventional wisdom in almost

the same ways as the first. If either explanation is correct, the triangles in Figure 14 represent a different condition than the circles. It follows that one cannot draw a smooth curve like A, B, C, or D in Figure 13 but must treat the two groups of points separately. Furthermore, if the triangles do lie on the lower branch of the speed-flow curve, the upper branch must have very high speeds at flows approaching capacity--the same conclusion one reaches if our hypothesis is accepted.

It could, of course, be argued that some of the triangles lie on each branch of the curve and that the upper branch does drop at high flows. If this is the case, however, the abrupt disappearance at 1830 PC/(lane·h) of points with speeds between 95 and 110 km/h (66 mph) becomes very difficult to explain, as does the virtually total separation of the circles and triangles. The seemingly random variation of the flows and speeds in Figures 7, 8, and 9 and the unimodal speed distribution we shall see in Figure 15 also seem incompatible with the idea that some of the observations made between 7:00 and 8:20 lie on one branch of the speed-flow relationship and some on the other. We cannot say that it is not true, but we find it far less likely than either of the two explanations offered above.

Figure 15. Frequency histograms for observed speeds.

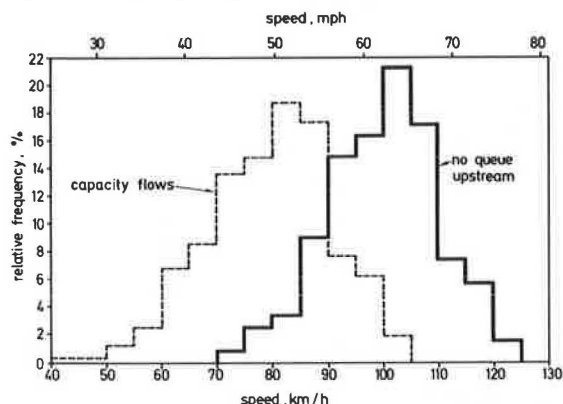


Figure 16. Cumulative frequency polygons for observed speeds.

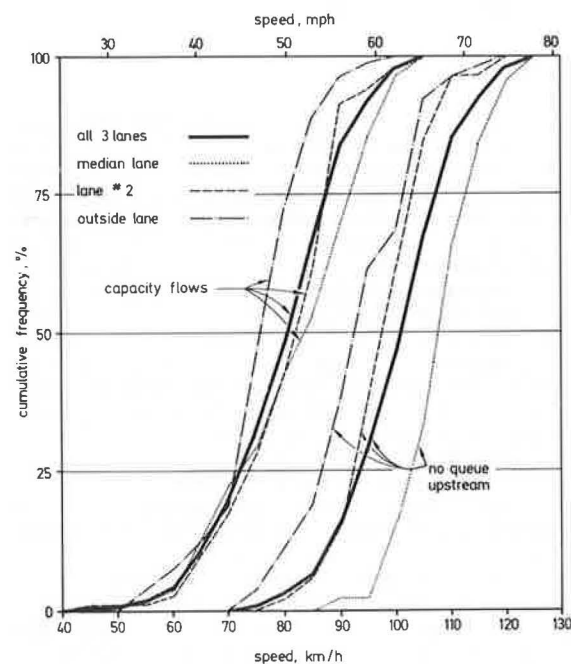


Table 2. Summary of speed observations.

Item	With Queue Upstream (Capacity) ^a (km/h)	No Queue Upstream ^b (km/h)
Mean speed (\bar{x})	79.5	100.2
SD	11.0	10.0
95 percent confidence interval for mean	$78.3 \leq \bar{x} \leq 80.7$	$98.4 \leq \bar{x} \leq 102.0$

Note: 1 km/h = 0.6 mph.

^aNumber of observations = 324.

^bNumber of observations = 122.

DISTRIBUTIONS OF SPEEDS

Under either our hypothesis or the alternative described at the beginning of the last section, there are two speed distributions: one for the condition when no queue exists upstream and the other for either capacity queue-discharge flows or for the speed within a very high-speed queue at near-capacity flow. Distributions of the speeds observed in these two situations are shown in Figures 15 and 16; the means, standard deviations, and 95 percent confidence limits on the means are given in Table 2. Figure 16 also shows the speed distributions for the three individual lanes. In Figures 15 and 16 and Table 2, and in the remainder of the paper, the speeds measured when a queue existed upstream are described as the speed of capacity flow, in accordance with our hypothesis. It should be noted, however, that if the alternative explanation is correct, the actual speed of capacity flows is higher than we have indicated.

Each speed plotted in Figures 7 through 14 was the average speed of the group of vehicles observed in a single set of four photographs. Figures 15 and 16 and Table 2, however, are based on the speeds of individual vehicles. It should be noted that the distribution for the no-queue situation is based on a sample that includes more observations at some flow levels than at others. This makes no difference if our hypothesis is correct. If, however, the speed really does vary with flow, the sample is biased. It is also biased with respect to time of day and truck percentage, since the later observations consistently included more vehicles than the earlier ones. Fewer problems arise with the capacity flow distribution, since all observations were made under reasonably similar conditions.

SUMMARY AND CONCLUSIONS

The data presented here tend to confirm other reports that urban freeway speeds remain high until the flow reaches at least 75 percent of the roadway capacity and that 2000 PC/h is still a good estimate of the capacity of a North American freeway lane under ideal conditions (2). On the other hand, we found an average speed of almost 80 km/h (50 mph) for capacity flows, much higher than the 50 km/h (30

mph) that is so often given in traffic engineering books.

Furthermore, and much to our surprise, our data led us to hypothesize that the speeds on uncongested freeway sections and in bottlenecks are not a function of flow but only of whether the vehicles are or are not being discharged from an upstream queue.

Our experiment was both simple and small, so our conclusions cannot be firm but must be checked by further experiments. If either our hypothesis or the alternative explanation that the true bottleneck was downstream is correct, however, the data indicate that the level-of-service concepts of the 1965 Highway Capacity Manual (3) need to be revised to an even greater extent than has been proposed by Roess, McShane, and Pignataro (1) and that the conclusions to be found in the large body of transport economics literature that assumes freeway speeds vary with flow in the manner indicated in Figure 1 or 2 all need to be reexamined.

ACKNOWLEDGMENT

Primary support for this research was provided by the National Science and Engineering Research Council of Canada. Funds for the data gathering were provided by the University of Toronto and the physical production of the paper was supported by the Institute für Verkehrswesen, Universität Karlsruhe, West Germany. The cooperation of the staff of the Ontario Ministry of Transportation and Communications is very much appreciated, as is the assistance lent by Chris Solecki and Dale Kerr.

REFERENCES

1. R.P. Roess, W.R. McShane, and L.J. Pignataro. Freeway Level of Service: A Revised Approach. TRB, Transportation Research Record 699, 1979, pp. 7-16.
2. A.D. May and others. Bay Area Freeway Operations Study. Institute of Transportation and Traffic Engineering, Univ. of California, Berkeley, 1968.
3. Highway Capacity Manual. TRB, Special Rept. 87, 1965.
4. Interim Materials on Highway Capacity. TRB, Transportation Research Circular 212, 1980.
5. D.C. Gazis, R. Herman, and R.W. Rothery. Non-linear Follow-the-Leader Models of Traffic Flow. Operations Research, Vol. 9, 1961, pp. 545-567.
6. A.D. May, Jr., and H.E.M. Keller. Evaluation of Single- and Two-Regime Traffic Flow Models. In Beiträge zur Theorie des Verkehrsflusses: Proc., Fourth International Symposium on the Theory of Traffic Flow (W. Leutzbach and P. Baron, eds.), Bundesministerium für Verkehr, Strassenbau und Strassenverkehrstechnik, No. 86, Bonn, 1969.
7. N.C. Duncan. A Further Look at Speed/Flow/Concentration. Traffic Engineering and Control, Vol. 20, 1979, pp. 482-483.
8. M. Koshi, M. Iwasaki, and I. Ohkura. Some Findings and an Overview on Vehicular Flow Characteristics. In Proc., Eighth International Symposium on Transportation and Traffic Theory (E. Hauer, V.F. Hurdle, and G.N. Stewart, eds.), Univ. of Toronto Press, Toronto, Canada, 1983.
9. P.K. Datta. Characteristics of Speed and Flow on a Freeway. Univ. of Toronto, Toronto, Canada, ME thesis, 1977.
10. V.F. Hurdle. The Effect of Queueing on Traffic Assignment in a Simple Network. In Transportation and Traffic Theory: Proc., Sixth International Symposium on Transportation and Traffic Theory (D.J. Buckley, ed.), A.H. and A.W. Reed, Sydney, Australia, 1974.
11. C. Hendrickson and G. Kocur. Schedule Delay and Departure Time Decisions in a Deterministic Model. Transportation Science, Vol. 15, 1981, pp. 62-77.
12. V.F. Hurdle. Equilibrium Flows on Urban Freeways. Transportation Science, Vol. 15, 1981, pp. 255-293.
13. P.H. Fargier. Effects of the Choice of Departure Time on Road Traffic Congestion. In Proc., Eighth International Symposium on Transportation and Traffic Theory (E. Hauer, V.F. Hurdle, and G.N. Stewart, eds.), Univ. of Toronto Press, Toronto, Canada, 1983.

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

Abridgment

Effectiveness Evaluation by Using Nonaccident Measures of Effectiveness

DAVID D. PERKINS AND BRIAN L. BOWMAN

The primary objective of highway safety expenditures is to improve roadway safety through reductions in accidents and accident severity. The ultimate measure of project effectiveness is therefore provided by analysis of changes in accident experience. Accident-based evaluations are, however, often impossible or undesirable due to limitations inherent in accident data bases. Low accident frequency, large lapse time, and a need to estimate ancillary benefits require the use of nonaccident measures. Nonaccident measures provide an intermediate measure that can be used to assess the effectiveness of completed highway safety projects and programs. This type of evaluation is useful when accident data are not available or are insufficient or when an indication of project effectiveness is desired sooner than the time necessary for accident-based evaluation. Nonaccident measures are considered intermediate because they are a supplement to and not a substitute for accident-based measures. No definitive quantitative relationships between changes in accident experience and nonaccident measures have been developed. A procedure for conducting an intermediate effectiveness evaluation by using nonaccident measures is described. Guidelines are presented for selecting evaluation objectives, nonaccident measures of effectiveness, experimental plans, and data requirements. The issues of statistical testing and interpretation of results as related to nonaccident evaluations are discussed.

In 1979, the Federal Highway Administration (FHWA) issued Federal-Aid Highway Program Manual (FHPM) 8-2-3 mandating the development and implementation of a continuing Highway Safety Improvement Program (HSIP) in all states. The overall objective of the HSIP is to reduce the number and severity of accidents and decrease the potential for accidents on all highways (1, Volume 8, Chapter 2, Section 3). Requirements for the structure of the HSIP include components for planning, implementing, and evaluating highway safety projects and programs. The details of the HSIP were defined in FHPM 8-2-3 and expanded in an FHWA study to develop options and procedures within each component (2). The planning component involves the selection and programming of projects through the collection and analysis of systemwide data, identification of hazardous locations, collection and analyses of site-specific accident and traffic data, and the selection of safety projects to be implemented. The implementation component involves the scheduling, design, construction, and operational review of the programmed safety projects. The evaluation component involves determining the effectiveness and efficiency of completed safety projects and programs. The evaluation component was the subject of a subsequent FHWA study (3) to develop evaluation guidelines for completed safety projects and programs within the HSIP.

The ultimate goal of evaluation within the HSIP is to improve the ability of state and local highway agencies to plan and implement future cost-effective safety programs based on the results of formal evaluations of ongoing and completed highway safety projects and programs. Effectiveness evaluation involves obtaining and analyzing quantitative information on the benefits and costs of implemented highway safety improvements. Knowledge of these benefits and costs reduces the dependence on engineering judgment and increases the ability of the agency to plan and implement highway safety improvements that have the highest probability for success. Thus, scarce safety funds can be properly allocated to high-pay-off improvements and diverted from those that are marginal or ineffective.

Effectiveness evaluation is based on an analysis

of the change in selected measures of effectiveness. Historically, the most acceptable measure for safety improvements is the change in police-reported accident experience at the project site. However, the stochastic nature of traffic accidents requires relatively large sample sizes collected over long periods of time. Other complications arise due to bias, inaccuracy, and confounding effects within the accident data base.

In response to the shortcomings of using accident experience as the sole criterion for safety evaluation, it may be necessary or desirable to conduct an interim effectiveness evaluation to obtain an indication of the short-term or intermediate effectiveness of the project based on changes in nonaccident measures of effectiveness. In such evaluations, nonaccident measures are not intended to be a substitute for accident measures since quantitative cause-and-effect relationships between accident and nonaccident measures have not been developed. If, however, nonaccident measures are selected that are logically related to accident experience or potential, the evaluation results can be used as a measure of intermediate effectiveness. The ultimate effectiveness, however, must be determined through an effectiveness evaluation based on observed changes in accident experience.

This paper describes selected elements of the procedure for evaluating completed safety improvements by using nonaccident measures as the primary measure of effectiveness.

POSSIBLE APPLICATIONS FOR NONACCIDENT MEASURES

The objective of safety expenditures is to improve safety through accident and severity reduction. Therefore, many projects are implemented to alleviate hazardous highway conditions that have caused abnormally high or severe accident experience. Many safety projects are not, however, implemented in response to abnormally high accident experience at specific locations. Rather, they are implemented to conform to accepted safety standards and practices or to prevent the emergence of an accident problem by treating potentially hazardous highway conditions and elements.

Although evaluations that examine changes in accident experience provide the most acceptable measure of project effectiveness, the requirements of accident-based evaluations often make this form of evaluation extremely difficult, if not impossible. One requirement for conducting accident evaluations is that accidents in sufficient numbers be available for use as measures of effectiveness. This requirement can generally be met for improvements at high accident locations but often accident experience at low-volume or rural locations is insufficient in number for accident-based evaluations. Another requirement is time. Usually, at least two years of accident data before and after project implementation are required for evaluation. Often, the time requirements exceed the practical limits when decisions must be made to continue, modify, or delete a particular safety project. In addition, evaluations require complete and accurate accident

data for use as measures of effectiveness. Accident data are often incomplete, erroneous, unavailable, or nonrepresentative of long-term conditions due to factors other than the improvement at the project site.

When conditions are not conducive to accident-based evaluations, nonaccident safety measures may provide valuable information on the intermediate effectiveness of a safety improvement.

Non-accident-based evaluations are applicable to many types of projects and evaluation study requirements as follows:

1. Safety projects that impact traffic performance: The primary purpose of a highway safety project is to reduce accident losses. However, in many cases problematic traffic performance, driver behavior, or other nonaccident safety measures provide the impetus for a safety project. In other cases, the improvement of traffic performance may be a secondary purpose (compared with accident reduction) of the safety project.

2. Need for a quick indication of project impacts: It is often imperative to obtain preliminary indications of project impacts soon after implementation. Previously untried projects may be evaluated based on changes in nonaccident measures as an indicator of whether the project is functioning as intended.

3. Projects implemented to reduce hazard potential: Many safety projects are implemented to meet safety standards or to eliminate specific safety deficiencies before significant accident experience develops. For these projects, accident data may not exist in sufficient numbers for accident-based evaluation. If it is not possible to obtain a sufficient accident sample through project aggregation, nonaccident evaluation procedures may provide a means of evaluating the project.

4. Projects involving staged countermeasure implementation: Individual countermeasures that make up a project may be assessed with a nonaccident evaluation when project implementation is staged. The nonaccident measures can be collected and evaluated between successive project implementation stages. This provides a means of evaluating countermeasures since the time periods between successive stages are generally too short to allow accident-based evaluation.

DEVELOPING NONACCIDENT EVALUATION PLAN

The first step in a nonaccident evaluation of a highway safety project is the development of an evaluation plan to provide overall guidance and direction. It offers the opportunity to think through the entire evaluation process in an attempt to establish the anticipated evaluation procedure and identify potential problems that may negatively impact the validity and efficiency of the evaluation effort. It is essential that the plan be developed prior to the implementation of the project so that nonaccident data may be anticipated and collected before project implementation.

The evaluation plan should address such issues as the selection of evaluation objectives, measures of effectiveness (MOEs), experimental plans, and data requirements.

Evaluation Objectives

The selection of objectives and MOEs for nonaccident evaluation is based on the ability to describe a chain of events that lead to accidents or create potential safety hazards. When the events are described, it is often helpful to consider three in-

terrelated types of factors: (a) causal factors, (b) contributory factors, and (c) the safety problem. Causal factors are defined as the predominant reasons why a safety problem exists. They are specific hazardous elements associated with the highway, environment, or vehicle. The factors can result in either the potential for accidents when a causal factor exists by itself or an accident occurrence in the presence of contributory factors. Contributory factors are elements or activities that lead to or increase the probability of a failure in the driver, the vehicle, or the environment. Safety problems are specific types of accidents or potential accidents that result from the existence of a causal factor and/or contributory factor (4).

The first step in selecting objectives is to develop the chain of causality for the highway safety project. The safety problem should be stated in terms of the actual or potential accident types to be reduced by the project. Next, the evaluator must identify the causal and contributory factors that lead to the safety problem. In many cases, the identification of these factors is straightforward since both causal and contributory factors are inherently considered when the countermeasures for the projects are developed. For example, suppose a project involves the implementation of an advance train-actuated warning flasher on an existing railroad crossing advance-warning sign at an approach with limited sight distance. The purpose of the project is to reduce the number, severity, and potential of vehicle-train and vehicle-vehicle rear-end accidents on the sight-restricted approach. The definition of the project and a knowledge of its purpose usually provide sufficient information to establish the chain of causality. Suppose the safety problem in this example is two automobile-train accidents involving two fatalities and five serious injuries during a three-year period. The major causal factor is the failure of drivers to perceive an occupied railroad crossing within sufficient time to stop and avoid an accident. The major contributory causes may be hypothesized (and verified) as limited sight distance and excessive vehicular speed (for conditions).

The intermediate objectives can be identified by perceiving how each causal and contributory factor will be affected by the introduction of the project. Thus, the correction or improvement in the causal and contributory factor provides the intermediate evaluation objectives for the evaluation. The underlying rationale of the approach is that if the intermediate objectives are achieved (i.e., if the causal and contributing factors are improved), the associated safety problem will be improved. (Verification of this rationale is subject to the results of an accident-based evaluation.)

For the rail-highway crossing example involving the installation of the flashing beacon, the intermediate objective may be defined as (a) reduction of vehicle speed at a specified point between the flasher location and the crossing and (b) increased speed reduction between points in advance of and following the warning sign after flasher installation.

MOEs

One or more MOEs should be specified for each intermediate objective. MOEs resulting from this process should be related to specific traffic operational or driver behavioral characteristics that are expected to be affected by the project. MOEs expressed as frequency, rate, and/or percentage may be appropriate.

The MOE should reflect the quantitative measure-

Table 1. Nonaccident MOEs for safety improvements at selected situations.

Situation	Nonaccident MOE	
	Behavioral	Operational
Horizontal curve	Lateral placement, shoulder encroachments, edgeline encroachments, centerline encroachments, brake applications, passing violations, speed violations	Spot speed, speed profile, deceleration profile
Vertical curve	Brake applications, passing violations	Spot speed, headway (downgrade)
Signalized intersections	Conflicts, lateral placement, brake applications, stop-bar encroachment, violations	Delay, travel time, approach speed, percentage of vehicles stopping, queue length
Stop approach	Head turns, conflicts, cross-road encroachment	Deceleration profile, spot speed
Tangent section	Lateral placement, violations	Speed, speed changes
Exit ramp	Distribution of exit points, erratic maneuvers	Mainstream spot speed, ramp spot speed, deceleration lane spot speed, deceleration profile
Weaving section	Conflicts, lateral placement, brake applications	Spot speed, speed profile
Lane drop or merge area	Distribution of merge points, erratic maneuvers	Mainstream spot speed
Railroad crossing	Head turns, brake applications	Spot speed, speed profile
Pedestrian or school crossing	Compliance, conflicts	Spot speed, delay

ments and units to be collected in the field to evaluate each intermediate objective. The evaluator should be as specific as possible when listing the MOEs. It is suggested that the evaluator refer to the state of the art of accident research when MOEs are selected. Table 1 (3) suggests several possible operational and driver behavior MOEs for safety projects at various roadway situations.

Experimental Plans

An experimental plan provides a framework for (a) estimating the expected value of each nonaccident MOE on the assumption that the project was not implemented and (b) determining the difference between the expected and actual MOEs.

An experimental plan should be selected that will maximize the validity of the evaluation. High levels of validity imply that the differences observed in the MOEs are due to the project and not a result of external factors such as changes in weather, enforcement, and other changes or improvements.

The before-and-after experimental plan generally provides very low levels of validity when applied in evaluations that span relatively long periods of time (i.e., accident-based evaluations require several years of before-and-after accident experience). However, because of the relatively short period of time between the before and after data-collection periods, the before-and-after plan is acceptable under many conditions when nonaccident measures are used in the evaluation. The time period between the before and after data collections is generally only a few months (depending on the length of the construction period) as opposed to the several years required for accident-based evaluation. Thus, it is less likely that significant changes other than the project itself will affect the MOEs and the results of the evaluation.

Evaluation plans involving control or comparison sites may be appropriate. If the time between data-collection periods becomes lengthy or if it is expected that atypical conditions may exist for either one or both periods, a control-site experimental plan should be used. If control sites are required but not available, the evaluation should not be conducted, since the validity of the evaluation results will be suspect.

Data Requirements

Given the selection of the objectives, MOEs, and

evaluation plans, specific evaluation data can be specified for collection. Nonaccident measures may consist of traffic performance variables such as travel time, delay, and speeds and/or driver behavior variables such as traffic conflicts and erratic maneuvers.

The intermediate objectives and associated MOEs provide input to determine what types of field data are required. The exact type of data, time of data collection, data-collection procedures, and data stratifications for each MOE should be specified prior to field data collection.

The magnitude of each data item must also be specified. The magnitude refers to when the data are to be collected and how much data are required to obtain a statistically reliable sample. Information on these items is contained in many traffic engineering references (5,6).

DATA COLLECTION

Because many types of field data may be needed for nonaccident evaluation, traffic engineering handbooks, manuals, and reports should be consulted to determine the specific data-collection activities to be performed in the field. Field activities require experienced data collectors and basic traffic engineering data-collection equipment. The number and level of involvement of field personnel vary with the type of field survey to be conducted, as does the type of equipment. Generally, there is sufficient flexibility in the sophistication of the study procedure and equipment requirements. Either manual or automatic procedures may be used, depending on agency resource levels, with little or no sacrifice in data quality or reliability. Data-collection costs will vary dramatically depending on the collection procedure and equipment.

Data collected before and after project implementation should be collected for similar time periods (time of day and day of week) and weather conditions and with identical data-collection procedures and personnel.

COMPARISON OF NONACCIDENT MOES

Intermediate project effectiveness is represented by the difference between the expected value of the nonaccident MOE if the project had not been implemented and the actual value of the MOE following implementation. This change provides an indication of the practical significance of the project (the determination of statistical significance is dis-

cussed in the next section). The method of determining the expected MOE and the percentage of change differs with the experimental plan selected for the evaluation. Experimental plans with higher levels of validity (i.e., control-site experimental plans) will generally improve the chances that the observed difference between expected and actual MOEs is primarily a result of the project.

Two computations are necessary: calculation of the expected value of the MOE if the improvement had not been made and calculation of the difference between the expected and actual MOE, usually expressed as a percentage. Formulas for these computations may be found in several studies on the subject (3,7).

STATISTICAL TESTING

Additional analysis is required to determine the statistical significance of the change in the selected MOEs. Statistical test results allow the evaluator to determine, with a specified level of confidence, whether the observed change can be attributed to the project or is the result of random (chance) fluctuations in the MOEs being tested.

The most important issue in statistical testing is the selection of the appropriate test. Selection is based on the type of nonaccident MOE, the evaluation objectives, the sample size, the experimental plan, and the statistical hypotheses to be tested. The type of MOE (data) refers to whether the data are discrete or continuous. The evaluation objectives refer to whether there is an interest in testing the difference in means, variances, proportions, or some other measure. The sample size aids in assessing the validity of assumptions that are made concerning the distribution of data and whether parametric or nonparametric tests are appropriate. The experimental plan provides input on the independence or correlation of the data being tested. Finally, the form of the stated hypothesis will suggest the appropriateness of the one-tail versus two-tail test. After each MOE has been specified in terms of the above factors, the selection of the appropriate test can generally be made by using engineering statistics references.

DATA-BASE DEVELOPMENT

The results of nonaccident evaluations can be organized into a data base analogous to accident-reduction-factor data bases. Such data provide feedback information useful in planning and implementing future projects to improve specific traffic performance, driver behavior, or other safety-related problems. It also provides input on quantitative cause-and-effect relationships between a project and its impact on nonaccident measures. This relationship, if analyzed in conjunction with the cause-and-effect relationship between the same project and accident measures, may provide insight into the existence of surrogates for accident experience for evaluation.

When both accident-based and non-accident-based evaluations are performed for a number of similar projects, the evaluator has the opportunity to determine whether there is a statistically significant relationship between changes in accident and nonaccident measures. If a strong, logical relationship is observed, a nonaccident measure may be feasible for use as a surrogate for accidents in future evaluations of similar projects.

INTERPRETATION OF RESULTS

The final determination of intermediate effective-

ness is based on the quantitative results of the evaluation and the ability to properly interpret the results. However, regardless of whether a conclusion on project effectiveness is positive (success), negative (failure), or otherwise, a critical assessment of the validity of the entire evaluation process as well as of the preceding planning and implementation activities and decisions should be performed. Key evaluation issues that need to be addressed prior to finalizing conclusions are as follows:

1. Was the project appropriate for achieving its intended purpose?
2. Were the chain-of-accident causality and the resulting intermediate objectives and nonaccident MOEs appropriate?
3. Was the experimental plan appropriate? What were the threats to validity that were not or could not be overcome?
4. Were the nonaccident data reliable and complete? What were the actual or suspected problems that were not correctable?
5. Were the control sites appropriate? What were the trade-offs made in control-site selection?
6. Was the statistical technique appropriate for the type of MOE and the desired evaluation objective?
7. Was the selected level of confidence appropriate?
8. Were the statistical test results reasonable?

In addition to a review of the evaluation study procedures, it is also important to review the appropriateness of decisions and activities that took place during the planning and implementation. It is important to recognize whether (a) the location was correctly identified as a hazardous location, (b) the project was properly selected and appropriate for the safety deficiency, and (c) the project was implemented as planned and designed.

If problems are observed or suspected for any of the above issues, they should be noted and an attempt should be made to correct them. If the problems are not correctable, this fact should be noted and should accompany the conclusions on intermediate project effectiveness.

CONCLUSIONS

The proper use of non-accident-based evaluation techniques can provide intermediate indications as to the effectiveness of implemented safety projects. The value of information obtained from nonaccident evaluation can extend beyond effectiveness evaluations. It can serve as an operational review tool to identify problems before they result in accident losses. It can serve to improve traffic flow and operations and to fortify contemplated remedial countermeasures during project planning activities. Caution should be exercised, however, to avoid acceptance of changes in nonaccident measures as a substitute or surrogate for changes in accident experience until such time as quantitative relationships can be identified.

These benefits and uses require that a greater appreciation for the advantages of nonaccident evaluation be obtained by traffic engineering practitioners. The procedures presented here are intended to encourage and guide practitioners in performing non-accident-based evaluations. The information obtained from both accident and nonaccident evaluations can be used to improve decisionmaking processes and increase roadway efficiency and safety.

ACKNOWLEDGMENT

This paper is based on the procedural guide Highway

Safety Evaluation, developed by Goodell-Grivas, Inc., under contract to FHWA.

The contents of this paper reflect our views, and we are responsible for the facts and accuracy of the information presented herein. The contents do not necessarily reflect the official policy of the U.S. Department of Transportation.

REFERENCES

1. Federal-Aid Highway Program Manual. FHWA, March 1979.
2. C.V. Zegeer. Highway Safety Improvement Program User's Manual. FHWA, Rept. FHWA-TS-81-218, 1980.
3. D.D. Perkins. Highway Safety Evaluation Procedural Guide. FHWA, Rept. FHWA-TS-81-219, 1981.
4. Evaluation of Highway Traffic Safety Programs--A Manual for Managers. National Highway Traffic Safety Administration, July 1977.
5. P.C. Box and J.C. Oppenlander. Manual of Traffic Engineering Studies, 4th ed. Institute of Transportation Engineers, New York, 1976.
6. M.A. Flak. Highway Safety Engineering Studies Procedural Guide. FHWA, Rept. FHWA-TS-81-220, 1981.
7. J.C. Laughland and L.E. Haefner. Methods for Evaluating Highway Safety Improvements. NCHRP, Rept. 162, 1975.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.

Surrogate Measures for Accident Experience at Rural Isolated Horizontal Curves

HAROLD T. THOMPSON AND DAVID D. PERKINS

The accident surrogate measures developed for hazardous-location identification and countermeasure evaluation at rural isolated horizontal curves are presented. An accident surrogate measure is defined as a quantifiable observation that can be used in place of or as a supplement to accident records. A list of potential accident surrogates was developed from four information sources: literature; a two-day workshop to obtain opinions and observations of highway safety professionals; analysis of an existing data base containing accident, geometric, operational, and environmental data; and selected field data collection at six rural isolated horizontal curves. Comprehensive sets of data were collected at 25 rural isolated curves. The data included measurements of operational and nonoperational characteristics and accidents. Statistical analyses of these data yielded five models for predicting specific types of accident rates. The strongest model developed in the study ($R^2 = 0.81$) indicates that the outside-lane accident rate can be predicted from measurements of the distance from the last traffic event on the outside lane and the speed differential between the approach speed and the curve midpoint speed for traffic in the outside lane. The other models (outside-lane accident rate, run-off-road accident rate, and two models for predicting rear-end accident rate) had R^2 -values greater than 0.65. The results indicate that accident surrogates can be developed through a systematic identification and measurement of roadway, driver, and traffic characteristics.

A primary goal of any highway safety agency is to reduce traffic accidents attributable to highway system failures. Historically, these agencies have relied heavily on reported traffic accidents to identify hazardous locations, to justify and prioritize safety improvements, and to evaluate their effectiveness. However, total dependence on accident history is somewhat questionable due to the limitations of these data. For example, the fact that a significant percentage of total accidents at a location are not reported often introduces error and results in suboptimal decisions. Another limitation is encountered when decisions to continue, modify, or remove countermeasures need to be made sooner than the waiting time required to collect reliable accident data.

Because of these and still other limitations, many highway safety researchers support the premise that nonaccident measures in addition to accidents should be used in the identification of hazardous locations, review of planned improvements, and evaluation of completed safety improvements. Review of several studies shows a fairly strong relationship

between accidents and various highway system characteristics such as geometrics, operations, environment, and driver behavior. However, there have been insufficient systematic efforts to investigate the feasibility of using such relationships as surrogates for accident experience in highway safety analyses.

A recent study entitled Accident Surrogates for Use in Analyzing Highway Safety Hazards (1) investigated the feasibility of using accident surrogate measures in

1. Identifying hazardous spot locations and sections of highway,
2. Evaluating the effectiveness of deployed safety countermeasures, and
3. Reviewing design plans of new facilities or improvements.

For the purpose of the study, an accident surrogate measure was defined as a quantifiable highway system feature that could be used in place of or as a supplement to accident data.

This paper presents the accident surrogates developed for highway safety analyses at rural isolated horizontal curves on two-lane roads. The surrogate development process involved (a) identifying potential highway system variables that could serve singly or in combination as surrogate measures and (b) developing explicit mathematical relationships between selected surrogate measures and accidents.

IDENTIFICATION OF CANDIDATE ACCIDENT SURROGATES

The identification of variables with potential as candidate surrogate measures was accomplished by obtaining information on actual and perceived relationships between accidents and elements of roadway, driver, and vehicle systems. Four information sources provided input on these relationships: literature; a two-day workshop to obtain opinions and observations of highway safety professionals; analysis of the Michigan Dimensional Accident Surveil-

lance (MIDAS) data base containing accident, geometric, operational, and environmental data; and selected field data collection at six rural horizontal curves.

These sources of information were synthesized to identify highway system variables worthy of further detailed analysis as surrogate measures in that a relationship between each variable and accidents had been demonstrated (or was strongly indicated). In an attempt to increase the validity and future utility of the final list of surrogate measures, members of the project team evaluated each candidate surrogate according to five criteria. The criteria include relationship to accidents, clarity of definition, credibility, ease of data collection, and affectability. The first four criteria are straightforward. However, further definition of affectability is necessary. Affectability is the likelihood that an improvement in the surrogate at a site will result in an improvement in the accident experience at that site. As an example, consider that the posted advisory speed at a horizontal curve is found to be a good indicator of the accident experience; i.e., higher accident rates become more likely as the posted advisory speed decreases. In the sense that this relationship is reasonably well established, posted advisory speed is a potential surrogate. However, it is clear that simply changing the advisory speed panel (to a higher value) will not result in an improvement in accident experience at a particular curve, because most likely this action will increase accident frequency. Hence, even though the posted advisory speed might well be rated high on relationship to accidents, clarity of definition, credibility, and ease of data collection, it will be rejected as a surrogate for countermeasure evaluation on the basis of the affectability criterion.

The selected candidate surrogate measures resulting from the final screening process are shown below. Although each surrogate did not rate high on all the criteria, each was considered at least passable on every criterion. The surrogate "speed-reduction efficiency" is defined as the ratio of the difference in actual speed reduction (average approach speed minus average speed at the curve midpoint) to the desired speed reduction (average approach speed minus the maximum permissible speed of the curve based on the friction factor).

Highway Safety Analysis

Identification		
Locations	Evaluation of Countermeasures	Design Plan Review
Speed-reduction efficiency	Speed-reduction efficiency	Design-speed differential
Curvature, grade, and distance since last curve	Physical evidence of error	Curvature, grade, and distance since last curve
Physical evidence of driver error	Erratic maneuvers	
Erratic maneuvers		

FIELD INVESTIGATIONS

The second step in the surrogate development process was to develop explicit mathematical relationships between surrogate measures and accidents. This was accomplished by analyzing candidate surrogate and accident data at a number of test sites. Regression

techniques were then used to identify the relationships between the candidate surrogates and accidents.

Selection of Candidate Surrogate Measures

Candidate surrogates were generally drawn from those tabulated above. However, some variables such as average annual daily traffic (AADT) and superelevation were added because of their logical association with accidents, whereas physical evidence of driver error was omitted due to difficulties relating to field measurement. The variables selected for field testing are listed below. The variables are identified as being either operational or nonoperational, since the intended use of these results required that the variables be separated. (In general, AADT is insensitive to highway safety treatments and thus was analyzed as a nonoperational variable.)

Nonoperational

AADT
Degree of curvature
Grade
Shoulder width
Distance since last curve
Superelevation
Slope of roadside (ditch, shoulder)
Type, location, and frequency of fixed objects

Operational

Encroachment
Speed reduction

Selection of Study Sites

A number of control variables were established to facilitate study-site selection. An attempt to reduce accident variance due to factors other than those selected for testing was made by limiting the range of these control variables rated as being either possible surrogate variables or as having shown some relationship to accidents. As a result, the following criteria were used to identify test sites:

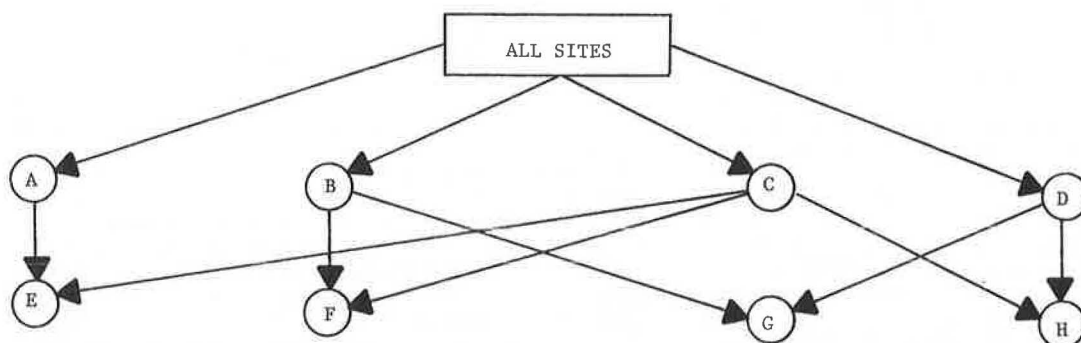
1. The curves should be located on two-lane, undivided roads and have a central angle of at least 20°.
2. Traffic volumes (AADT) should not exceed 8000 vehicles and posted speeds on curve approaches should be between 35 and 55 mph (advisory speeds on the curves may vary).
3. Lane widths should be between 10 and 12 ft and there should be gravel shoulders.
4. At least 1/4-mile distance should separate the study site from a preceding highway event that necessitates driver action to adjust vehicle path and/or speed (e.g., another curve, railroad crossing, stop sign, traffic signal, etc.).
5. The curves should not have extremely unusual roadside features.

Twenty-eight roadway sections containing isolated curves were identified through a search of the Oakland County, Michigan, inventory files. Oakland County was selected because of the availability of recent photologs, a complete file of highway improvement projects implemented since 1975, and reliable accident and volume data. Each of the sites was visited to determine whether they met all the criteria specified for test sections. Twenty-five of the sites were acceptable, and data were collected at each of these sites.

Stratification of Study Sites

The existence of complex interactions between geo-

Figure 1. Site stratification for rural isolated curves.



metric, traffic, and driver behavior variables and accident experience generally tends to mask explicit mathematical relationships between these variables and accidents. To reduce this masking effect, the test sites were stratified into subsets of sites with similar major characteristics. For example, if one or a combination of independent variables is a good surrogate for curves with restricted sight distance and another set of variables is a good surrogate for curves with no sight restrictions, these relationships can only be determined if the two categories of curves are separated during the analysis. It is possible that neither relationship would be significant for the combined sample of test sites.

The variables used to stratify the locations were sight distance, grade, land use, and the posted speed limit.

In addition to the single-variable categories, additional categories were constructed by using sight distance and land use, grade and posted speed limit, and land use and speed limit. A total of nine groups were identified for the analysis (including all sites as a group). These groups are identified by letter in Figure 1.

Group A consists of curves with sight distance limited by trees, embankments, or other obstacles close to the roadway or the inside of the curve. This group contains 19 of the 25 curves. The rationale for this stratification is that the restriction in sight distance could alter the degree to which driver expectancy is met, and this factor was identified as important in both the literature and the workshop.

Group B consists of curves on relatively flat roadway sections (less than 4 percent grade). Nearly all of the sites fall in this class (22 of 25). The rationale for this stratification is to moderate the effect of combined horizontal and vertical curvature on the accident rate.

Group C consists of roadway sections with zero or one driveway on the curve (low residential land use). As in group A, this is done to reduce the variation in the driver expectancy across the sample. Twenty of the 25 sites fall in this category.

Group D consists of all roadway sections with a posted speed of 45, 50, or 55 mph. Nineteen of the 25 curves fall in this category. This factor was used because the posted speed limit may affect driver characteristics at those sites, thus increasing the variance in the data.

Group E consists of all the sites meeting the criteria for group A (limited sight distance) and group C (few driveways). This group contains 14 of the 25 curves. Group F consists of all sites meeting the criteria for groups B and C and contains 17 curves; group G consists of all sites meeting the criteria for groups B and D and contains 16 curves;

and group H consists of all sites meeting the criteria for groups C and D and contains 15 curves.

Accident Data

Three years of accident data (1976, 1977, and 1978) were collected for each test site. Computer printouts of accidents were obtained for the specified limits of the sites plus all accidents occurring within 200 ft of the site boundaries. Each accident was examined with respect to vehicle involvement, contributory circumstances, and vehicle paths. Accidents were then stratified by type of accident and severity. Locations with unusual accident patterns, such as a high incidence of car-animal accidents, were eliminated from further consideration.

Independent/Dependent Variables

The potential surrogates (independent variables) collected and/or calculated for each of the study sites are listed in Table 1 along with the accident characteristics (dependent variables) used in the analysis.

Analysis Techniques

Regression techniques [the maximum R^2 -improvement technique (Max R^2) contained in the Statistical Analysis System (SAS) was selected as the most appropriate regression technique] were used in the analysis in which the selected candidate surrogate variables were used as independent variables and three-year accident rates for total accidents and predominant accident types were used as dependent variables. Stepwise regression was used as the analysis procedure to test for statistically significant relationships between one or a combination of candidate surrogate variables and accident experience at the test sites.

Regression analyses were performed for specific stratifications to search for statistically significant relationships between accidents and (a) combinations of nonoperational and operational variables, (b) nonoperational variables only, and (c) operational variables only. Surrogates developed from these three independent analyses were to be used for identification of hazardous locations, design plan review, and countermeasure evaluation, respectively.

RESULTS OF ANALYSIS

A total of 162 separate regression analyses were conducted on the data set by using the Max R^2 -stepwise linear regression model. This number of runs was required because of the stratification by type of independent variable (operational, nonopera-

Table 1. Independent and dependent variables included in analysis.

Dependent Variable	Independent Variable	
	Nonoperational	Operational
Accident rate: total (2), rear end (3), opposite direction (4), run off road (5), and fixed object (6)	AADT (9); degree of curvature (10); percent grade (12); superelevation error, ^a entire pavement width (69); shoulder width, average width for both shoulders (62); side-slope angle, ratio x:1, average for both sides of road (63); fixed-object rating for objects within 10 ft of pavement edge adjacent to outside travel lane (64); fixed-object rating for objects within 10 ft of pavement edge adjacent to inner travel lane (65)	Total encroachment rate ^b (17); speed differential of vehicles in outside travel lane between points on curve approach and curve midpoint, mph (38); speed differential of vehicles in inner travel lane between points on curve approach and curve midpoint, mph (41); average speed reduction efficiency ^c (66)
Inside-lane accident rate (17)	AADT (9); degree of curvature (10); percent grade (12); distance since last traffic event, inner travel lane, miles (14); superelevation error, inner travel lane (67); shoulder width adjacent to inner lane, ft (42); slide-slope angle adjacent to inner lane, x:1 (44); fixed-object rating for objects within 10 ft of edge of inner travel lane (65)	Total encroachment rate for inner-lane traffic (18); centerline encroachment rate for inner-lane traffic (34); edgeline encroachment rate for inner-lane traffic (35); speed differential of vehicles in inner lane between points on curve approach and start of curvature, mph (39); speed differential of vehicles in inner lane between points at start of curvature and curve midpoint, mph (40); speed differential of vehicles in inner travel lane between points on curve approach and curve midpoint, mph (41), speed-reduction efficiency on inner lane (60)
Outside-lane accident rate (8)	AADT (9); degree of curvature (10); percent grade (12); distance since last traffic event, outside travel lane, miles (13); superelevation error, outside travel lane (68); shoulder width adjacent to outside lane, ft (42); side-slope angle adjacent to outside lane, x:1 (45); fixed-object rating for objects within 10 ft of edge of outside travel lane (66)	Total encroachment rate for outside-lane traffic (19); centerline encroachment rate for outside-lane traffic (32); edgeline encroachment rate for outside-lane traffic (33); speed differential of vehicles in outside lane between points on curve approach and start of curve, mph (36); speed differential of vehicles on outside lane between points at start of curvature and curve midpoint, mph (37); speed differential of vehicles in outside travel lane between points on curve approach and curve midpoint, mph (38); speed reduction efficiency on outside lane (59)

Note: Variable numbers are shown in parentheses.

^aDifference between minimum superelevation required for prevailing conditions and actual superelevation (in/ft).

^bNumber of edgeline plus centerline touches per 100 vehicles entering curve.

^cRatio of observed speed reduction to desirable speed reduction due to curvature and superelevation, averaged for both directions of travel.

tional, or combined), the grouping of curves by physical attribute (nine groups), and the analyses of six stratifications of the dependent variable.

The simple correlation coefficients for each combination of one independent and one dependent variable were computed. Confidence limits of 95, 90, and 80 percent were used to test these correlations. Any independent variable for which the correlation coefficient was not significantly different than zero at the specified confidence level was rejected as a possible factor in the multiple regression model for predicting that dependent variable. Thus, only variables that are independently correlated to accidents were included in the stepwise multiple regression runs.

Residual error plots were also examined for each regression model that satisfied the statistical criteria for model selection. This check was performed to determine whether nonlinear transformations were necessary based on the variance of the residuals (constant variance is assumed in linear regression) and the existence of outliers. Transformations of the data were not indicated for any of the models presented in this section.

The analysis failed to identify a good surrogate measure for the total accident rate when all 25 locations were used. The only variable that was independently correlated with total accident rate and that remained in the Max R^2 -model at the 0.05 level of significance was degree of curvature. However, the R^2 -value for this one-variable model was only 0.16, and thus it is not considered to be a strong surrogate for total accidents.

The results are consistent with those from the literature review, the workshop, and the analysis of MIDAS in that this factor was identified as important in all three. It is not surprising that there

is no single surrogate that explains all accidents at all locations.

The most clearly defined surrogate measure for rural isolated curves, the outside-lane accident rate, resulted from the analysis of outside-lane accidents on highway sections with zero or one driveway per section and a speed limit greater than or equal to 45 mph (group H), which used both operational and nonoperational variables (Table 2). The coefficient of multiple correlation (R^2) for this model was 0.81, and the variables used were distance to last traffic event on the outside lane (V13) and speed differential between the approach speed and curve midpoint speed for traffic in the outside lane (V38).

The relatively high R^2 -value is not unexpected since both the independent variable and the dependent variable contain only a subset of the total sample. For this particular data base, then, it was possible to define a surrogate measure that is easily measured, capable of being measured immediately following implementation of a safety countermeasure, and strongly correlated to one particular type of accident.

One of the primary objectives of this study was to determine whether this could be accomplished through a logical procedure by using both the experience of practicing engineers and statistical testing. This objective has been met for this particular subset of the data.

Similar results were obtained for other accident classifications, situations, and groupings. Some of the more promising results are described in the following paragraphs. (All the models in Table 2 are constructed from variables that are significantly correlated with the relevant accident data at the 0.05 level.)

Table 2. Surrogate measures and associated mathematical models for rural isolated curves.

Accident Measure (accidents/million vehicles)	Surrogate Measure	Site Characteristic	Model
Outside-lane accident rate (V08)	Distance to last event, outside lane (V13); speed differential (V38)	Low residential land use, posted speeds ≥ 45 mph	$V08 = 0.03227 + 0.5949V13 + 0.1510V38$ $R^2 = 0.81$
Rear-end accident rate (V03)	ADT (V09), side-slope angle (V63)	Grade < 4 percent	$V03 = -0.1026 + 0.00004184V09 + 0.0001284V63$ $R^2 = 0.74$
	ADT (V09)	Grade < 4 percent, low residential land use	$V03 = -0.06900 + 0.00004595V09$ $R^2 = 0.72$
Run-off-road accident rate (V05)	Degree of curve (V10), superelevation error (V69)	Restricted sight distance, low residential land use	$V05 = -2.975 + 0.4985V10 - 1.508V69$ $R^2 = 0.68$

For rural isolated curves, reasonably good models ($R > 0.65$) were obtained for (Table 2)

1. Outside-lane accident rate for group H by using the distance to the last event and the speed differential on the outside lane,
2. Rear-end accident rate for group B by using the ADT and the side-slope angle,
3. Rear-end accident rate for group F by using the ADT, and
4. Run-off-road accident rate for group E by using the nonoperational degree of curve and the operational superelevation error.

Further examination of the correlation and regression results provides additional insight regarding variations in accident experience at horizontal curves.

1. For total accident rate, the independent variables selected by Max R^2 most frequently are speed differential on the outside lane, degree of curve, and total encroachment rate.
2. For rear-end accident rate, the independent variables selected most frequently are ADT and total encroachment rate.
3. For opposite-direction accident rate, the independent variables selected most frequently are speed differential on the inside lane, degree of curve, and fixed objects within 10 ft of inside lane.
4. For run-off-road accident rate, the independent variables selected most frequently are degree of curve and speed differential on the outside lane.
5. For inside-lane accident rate, the independent variables selected most frequently are encroachment rate on the inside edgeline and fixed objects within 10 ft of inside lane. (Neither of these shows up nearly as frequently as the independent variables for the other types of accident rates.)
6. For outside-lane accident rate, the independent variables selected most frequently are speed differential on the outside lane, distance to last event in outside lane, and degree of curve.
7. The success in developing models also varied by subgroupings of the sites. Eliminating sites with posted speeds below 45 mph enhanced success considerably; eliminating sites with grades greater than 4 percent was next most helpful.

These observations are consistent with intuition and lend credence to the data and statistical procedures. However, this was an exploratory study of accident surrogates and hence the data base for any

given situation or accident type or surrogate was limited.

SUMMARY AND CONCLUSIONS

The results of this study indicate that surrogate measures for accident experience can be identified. Further, a procedure for doing so has been developed and demonstrated to a limited degree. This procedure involved extensive review of the literature pertaining to studies of the effect of various operational and nonoperational highway, driver, and traffic variables on accident experience; the judgment of a group of highway safety experts on which variables were most promising in terms of developing mathematical relationships with accidents; the analyses of existing data bases to assess probable relationships; a limited amount of field data collection to supplement the other sources; and a synthesis of all these inputs to select the variables most likely to lead to meaningful surrogates. Results of the application of that procedure were tabulated at the beginning of this paper.

Data were collected for candidate surrogate measures and various categories of accident types at 25 study sites. Statistical analyses of these data yielded five reasonably strong models for predicting particular types of accident rates.

The strongest model developed in the study indicates that the outside-lane accident rate at horizontal curves can be predicted from measurements of the distance since the last traffic event on the outside lane and speed differential between the approach speed and curve midpoint speed for traffic in the outside lane. The model is strongest when applied to highways with a posted speed limit of 45 mph or greater.

The prediction models formulated in this study are based on data from a limited geographic area and may only be appropriate for selected safety studies within that area. Some caution should be exercised in extrapolating the models to other areas with differing laws, law enforcement, driver behavior, terrain, weather, and traffic control devices. It is quite possible that the models are applicable in wider areas (and that is certainly desirable, given the effort required to construct such models), but testing will be required to determine their suitability in other geographic areas.

With qualifications imposed by the size of the data set, the primary objective of the study, which is to demonstrate that accident surrogates can be developed through a systematic identification and measurement of roadway, driver, and traffic charac-

teristics, has been accomplished. Generalizing the surrogates formulated here and developing new surrogates can now proceed at a much faster pace with more efficient data collection and analyses.

ACKNOWLEDGMENT

This paper reports some of the findings of a study performed by Goodell-Grivas, Inc., under the sponsorship of the Federal Highway Administration, U.S. Department of Transportation.

We acknowledge with thanks the contributions of Tapan K. Datta of Goodell-Grivas, Inc.; William C. Taylor of the University of Michigan; and James I. Taylor of the University of Notre Dame.

The contents of this paper reflect our views, and we are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official policy of the U.S. Department of Transportation.

REFERENCE

1. Goodell-Grivas, Inc. Accident Surrogates for Use in Analyzing Highway Safety Hazards, 3 vols. FHWA, Repts. FHWA-RD-82-103 to 105, in preparation.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.

Candidate Accident Surrogates for Highway Safety Analysis

DAVID D. PERKINS AND HAROLD T. THOMPSON

The variables identified as potential accident surrogate measures for use in identification of hazardous locations, evaluation of safety countermeasures, and design plan review at 10 specific highway situations are presented. Situations included urban undivided tangent section, rural undivided winding section, rural isolated curve, lane drop, narrow bridge, exit gore area, urban nonsignalized intersection, rural nonsignalized intersection, rural undivided tangent section, and rural signalized intersection. Accident surrogate measures are defined as quantifiable highway system features and characteristics that can be used in place of or as a supplement to accident records. The list of candidate surrogates was developed from four information sources: literature, a two-day workshop to obtain opinions and observations of highway safety professionals, analysis of an existing data base, and selected field data collection.

Highway safety programs administered by the Federal Highway Administration (FHWA) are aimed at reducing traffic accidents attributable to highway system failures. To be effective, safety improvement programs must follow a systematic procedure to identify the safety deficiency, develop and implement a solution, and monitor the effectiveness of the implemented solution.

Historically, highway safety agencies have relied heavily on reported traffic accidents to identify problem locations, justify and prioritize safety projects, and evaluate their effectiveness. Many highway safety professionals, however, recognize significant shortcomings in the highway safety process when accidents are used as the sole criterion for highway safety planning and evaluation. One shortcoming is apparent when decisions to continue, modify, or remove countermeasures need to be made sooner than the waiting time required to collect accident statistics. In other instances, it is a shortcoming when safety problems are characterized by accident potential as opposed to the occurrence of accident patterns or trends. These situations often occur on low-volume roads, in rural areas, and at rail-highway grade crossings.

Because of these limitations, many highway safety professionals support the premise that identification of problem locations and effectiveness evaluations should consider alternative measures in addition to accidents. Past studies indicate that highway system characteristics such as geometrics, operations, environment, and driver behavior are related to accident experience. Several research efforts have identified precise relationships between individual characteristics and accidents. However,

there have been only limited systematic efforts to investigate the feasibility of using such relationships as surrogates for accident experience in highway safety analyses.

A recent study entitled Accident Surrogates for Use in Analyzing Highway Safety Hazards (1) investigated the feasibility of using accident surrogate measures in

1. Identifying hazardous spot locations and sections of highway,
2. Evaluating the effectiveness of deployed safety countermeasures, and
3. Reviewing design plans of new facilities or improvements.

Accident surrogate measures are defined as quantifiable highway system features and characteristics that can be used in place of or as a supplement to accident records. From a theoretical viewpoint, an accident surrogate measure must possess a definite relationship to accidents and be sensitive to safety-related changes in the highway system. From a practical viewpoint, surrogate measures must be relatively easy to collect with minimal training and equipment.

In this paper we present the variables identified as potential accident surrogates based on information obtained from four information sources: literature, a two-day workshop to obtain opinions and observations of highway safety professionals, analysis of the Michigan Dimensional Accident Surveillance (MIDAS) data base, and selected field data collected at five highway situations. Variables identified as candidate surrogates came primarily from the literature and the workshop. The MIDAS data were used to investigate the potential for surrogates by analyzing geometric, traffic, and environmental variables contained in that data base. For other potential surrogates, limited field studies were undertaken to provide an additional quantitative source of input. No candidate surrogate was eliminated from future consideration on the basis of either the MIDAS analyses or the limited field studies. The candidate accident surrogates were later field tested on a much larger scale to determine the strength of their relationship with accidents and utility as surrogates for accidents.

Variables identified through these sources were grouped according to their relevance to 10 specific highway situations. Situations considered were

1. Urban undivided tangent section,
2. Rural undivided winding section,
3. Rural isolated curves,
4. Lane-drop locations,
5. Narrow bridge,
6. Exit gore area,
7. Urban nonsignalized intersection,
8. Rural nonsignalized intersection,
9. Rural undivided tangent section, and
10. Rural signalized intersection.

LITERATURE REVIEW

The literature review consisted of past and current studies on the relationship between traffic accidents and elements of the highway system (geometry, roadside environment, traffic control, traffic operations, and driver behavior) for each of the selected highway situations. Reference sources included National Technical Information Service (NTIS), existing literature reviews, and the libraries of Wayne State University, University of Notre Dame, and University of Michigan.

The literature review identified 52 highway system elements as potential accident surrogates for one or more of the 10 highway situations. The variables and variable combinations were placed into two general categories--nonoperational and operational. Nonoperational variables relate to roadway geometry and cross-sectional elements, traffic control operations, driver performance, and driver behavior. Both types of variables are listed below:

Nonoperational variables:

1. Degree of curve,
2. Frequency of curves,
3. Grade,
4. Grade continuity,
5. Surface cross slope,
6. Sight distance,
7. Visibility of signal and sign,
8. Pavement width,
9. Lane width,
10. Approach width,
11. Pavement shoulder presence,
12. Shoulder width,
13. Percent shoulder reduction (between shoulder width on approach and shoulder width on bridge),
14. Median width,
15. Bridge width,
16. Ratio of bridge width to pavement width,
17. Difference between roadway width and bridge width,
18. Taper length,
19. Number of lanes dropped,
20. Length of deceleration lane,
21. Bridge safety index,
22. Structural adequacy of guardrail and bridge-rail,
23. Access control,
24. Number of commercial driveways per mile,
25. Number of intersections per mile,
26. Number of traffic signs per mile,
27. Type of delineation treatment,
28. Raised marker delineation,
29. Signing and delineation,
30. Type of advance warning,
31. Intersection design,
32. Type of traffic control device,
33. Illumination level, and
34. Skid resistance.

Operational variables :

1. Traffic volume,
2. Major and minor road volumes,
3. Opposing traffic volume,
4. Percent diverging traffic,
5. Traffic mix,
6. Volume/capacity ratio,
7. Posted speed,
8. Operating speed,
9. Speed differential,
10. Speed variance,
11. Lateral placement,
12. Traffic conflicts,
13. Erratic maneuvers,
14. Cycle length,
15. Signal phasing,
16. Number of phases,
17. Total stopped-vehicle delay, and
18. Red- and yellow-light violations.

The potential surrogates were categorized as "strong" or "other" according to the degree of convergence of research evidence and the reliability of the research studies considered during the literature review. A strong potential surrogate is a variable found to be related to accident experience in at least one reliable study. The reliability of a study was based on the acceptability of the article by the highway safety community, the validity of the experimental design, the sample size, and the number and type of variables controlled in the study. Meaningful conclusions and valid analysis procedures were requirements for classifying a measure as a strong potential surrogate. Where there were conflicting results from two or more reliable sources, the surrogate was not labeled "strong".

A potential surrogate is defined as "other" when it is a measure for which there is less empirical evidence and no specific relationship is defined in the literature. Standards and guidelines, such as AASHTO design standards, were selected as "other" potential surrogates. Other examples include length of taper at lane-drop locations and sight distance. These variables and their relationships to accidents are logical from an engineering-practices viewpoint, but often there is limited evidence of statistical validity or the studies are based on small samples.

Operational surrogates (such as erratic maneuvers) were used in several studies for evaluating the operational effects of countermeasures. These studies attempt to quantify the level of driver error that is logically related to the level of hazardness. The use of such operational variables in accident studies, based on their logical relationship to safety, justifies their selection as "other" potential surrogates, even though the relationships to accidents have not been validated.

Figures 1 and 2 show the selected nonoperational and operational variables, respectively, and the associated potential surrogate designation. An S indicates a strong potential surrogate and an O indicates an "other" potential surrogate.

WORKSHOP

The workshop was attended by 13 highway safety professionals with backgrounds in traffic engineering, highway safety research, and highway safety administration. Participants were asked to examine and critique a prepared list of geometric, operational, traffic control, and environmental factors. The list included the nonoperational and operational variables identified in the literature review together with more than 50 other variables identified by other researchers on the basis of logical (as op-

Figure 1. Potential surrogate classifications for nonoperational variables.

Non-Operational Variables		Situations																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													
Rural Isolated Horizontal Curves Rural Undivided Tangent Sections Rural Undivided Winding Sections Rural Signalized Intersections Rural Non-Signalized Intersections Urban Undivided Tangent Sections Urban Non-Signalized Intersections Lane Drop Locations Exit Gore Areas Narrow Bridge	Degree of Curve	S																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													

Note: "S" denotes strong potential surrogate; "O" denotes other potential surrogate.

Figure 2. Potential surrogate classifications for operational variables.

Situations	Operational Variables		Traffic Volume	Major and Minor Road Volumes	Opposing Traffic Volume	Percent Diverging Traffic	Traffic Mix	Volume/Capacity Ratio	Posted Speed	Vehicle Speed	Speed Differential	Speed Variance	Lateral Placement	Traffic Conflicts	Erratic Maneuvers	Cycle Length	Signal Phasing	Number of Phases	total Stopped Vehicle Delay	Red & Yellow Light Violations
Rural Isolated Horizontal Curves	0									0	0									
Rural Undivided Tangent Sections	S											0								
Rural Undivided Winding Sections	0																			
Rural Signalized Intersections		S							S					0						0
Rural Non-Signalized Intersections		S												0						0
Urban Non-Signalized Intersections	0																			
Urban Undivided Tangent Sections		S												0						
Lane Drop Locations																				
Exit Gore Areas						S									0					
Narrow Bridge	0		0		0		0	0		0			0							

Note: "S" denotes strong potential surrogate; "O" denotes other potential surrogate.

posed to statistical) relationships to accidents.

To facilitate a detailed examination, the factors were categorized under one or more hazard indices used to describe the causal chain of events leading to an actual or potential accident at the various highway situations (e.g., isolated curves, exit gore areas, and railroad crossing). The indices that make up the causal chain include information, human factors, vehicle control, congestion, and recovery. Definitions for these indices are provided below (note that the indices are defined such that higher values indicate higher degrees of hazard):

1. Information index: This is a measure of the information system deficiencies that detract from the driver's ability to select a safe speed and path as roadway conditions change. The absence of lane markings and inadequate advance-warning signs are examples of factors that contribute to a high information index.

2. Human factors index: This is a measure of the existence of conditions that fail to meet typical driver expectancies, therefore increasing the probability that a driver will respond incorrectly to a situation requiring evasive actions. A sharp horizontal curve following the crest of a vertical curve is an example of a factor that would contribute to a high human factors index.

3. Vehicle control index: This is a measure of the geometric and environmental characteristics that constrain the driver's ability to maintain control of the vehicle in a traffic stream. Inadequate sight distance and icy pavements are examples of factors that contribute to a high vehicle control index.

4. Congestion index: This is a measure of the operational characteristics that constrain the driver's ability to avoid an accident through a controlled vehicle maneuver. Congested flow and excessive numbers of driveways and parked vehicles along a roadway are examples of factors that contribute to a high congestion index.

5. Recovery index: This is a measure of the roadway and roadside characteristics that inhibit the driver's ability to avoid an accident or to reduce the severity of an accident resulting from partial or total loss of vehicle control. Narrow shoulders and roadside objects are examples of factors that contribute to a high recovery index.

The causal chain of events is based on the following scenario (Figure 3). A driver is presented with information from a variety of sources, including signing, the environment, and other vehicles. Through this information and past driving experiences, the driver develops mental perceptions and expectations of the driving environment. If these perceptions and expectations agree with the actual conditions, the driver can select an appropriate speed and path and safely maneuver the vehicle. If the actual conditions do not meet with what the driver perceives or expects, corrective adjustments in vehicle path or speed must be made. The vehicle control and congestion indices contain factors that determine the outcome of these adjustments. That is, if the vehicle remains under control and traffic conditions are such that an adjustment can be made without interference with other vehicles, an accident is avoided. If either of these conditions does not exist, the driver is faced with a recovery situation that results in either a near miss (recovery and no accident) or a single- or multiple-vehicle accident.

Participants were then asked to review a comprehensive list of variables for each hazard index for each highway situation. Factors were added, re-

moved, and/or redefined to fit the specific combination of index and highway situation. From these lists, workshop participants were asked to identify a limited set of variables that had the strongest intuitive and/or empirical relationship to accidents.

ANALYSIS OF MIDAS DATA BASE

The MIDAS system was analyzed to determine whether other highway system variables should be considered as candidate surrogates.

At the time of the analysis, the MIDAS data base contained geometric, environmental, traffic, cross-section, and accident data for 9000 miles of state roadway system in Michigan. Geometric data included laneage and horizontal and vertical alignment. Environmental data included roadside development and intersection traffic control. Traffic data included estimated hourly and daily volumes and speed limit. Cross-section data included lane width, shoulder width, curb type, median or no median, and turn lanes. Accident data included frequency of fatal plus injury accidents by type. Accident rates could be calculated directly from the volume and accident frequency data.

The analysis consisted of categorizing the data into the individual highway situations. MIDAS data were available for 7 of the 10 highway situations.

Analysis of variance (ANOVA) tests were also conducted to examine differences in mean fatal and injury accident rates resulting from various roadway and operational stratifications. The t-statistic was employed to determine the direction of the difference in cases where the ANOVA indicated a significant difference.

Because of the availability of only fatal and injury accidents, highway situations in urban areas were not considered in the ANOVA. The rural highway situations that were analyzed included isolated curves, undivided winding sections, undivided tangents, signalized intersections, and nonsignalized intersections. Many variables contained in the MIDAS data base could not be statistically analyzed due to the small number of locations for some variable categories. Summarized ANOVA findings follow:

1. Effect of average daily traffic (ADT) on rate of injury and fatal accidents was found for signalized intersections.

2. The effects of posted speed limits on rate of injury and fatal accidents were examined for all highway situations. The only statistically significant finding was that nonsignalized intersections with higher posted speed limits (50-55 mph) have a higher rate of injury and fatal accidents than intersections with lower speed limits (40-45 mph). This finding holds for ADT ranges from 2000 to more than 10 000 vehicles/day.

3. Intersections carrying 10 000 vehicles or more per day with volume-to-capacity (V/C) ratios of 0.5 to 1.0 have significantly lower mean rate of injury and fatal accidents than do intersections with V/C ratios between 0.0 and 0.5.

4. The effects of lane width on rate of injury and fatal accidents were examined only for isolated-curve sections and winding-roadway sections. The only significant finding was that winding sections with narrow pavement widths have a higher mean rate of injury and fatal accidents than sections with wider pavement widths.

5. Shoulder-width effects were examined for isolated curves and winding sections. No significant results were found for isolated curve sections, and no significant findings that would apply to the overall range of conditions for winding sections were detected.

6. No significant effects attributed to the presence or absence of a vertical curve were found for any of the highway groups.

The ANOVA generally indicates that some of the factors analyzed have a significant effect on rate of injury and fatal accidents. These factors were considered candidate surrogates. Because of the limitations of the MIDAS data base and the fact that only injury and fatal accidents were included in the analyses, candidates that did not show significant relationships were not eliminated.

ANALYSIS OF SELECTED FIELD DATA

As a supplement to the literature review, workshop, and MIDAS data base analysis, supplemental data collection and analysis activities of a limited nature

were undertaken to provide an additional source of input in the determination of candidate surrogates.

Candidate surrogate measures identified from the aforementioned sources and analyses were collected at five highway situations, including rural isolated curves, rural winding sections, urban undivided tangents, rural signalized intersections, and lane-drop locations.

Sites were selected in Oakland County, Michigan. In the selection of sites, basic cross-sectional and operational features, such as number of lanes and ADT, were limited to control for accident variance due to these characteristics.

Four statistical analysis techniques were used to test the relationships between the collected candidate surrogate measures and predominant accident types: (a) nonparametric (Spearman rho) correlation analysis, (b) parametric (Pearson) correlation anal-

Figure 3. Causal chain of events for potential accidents.

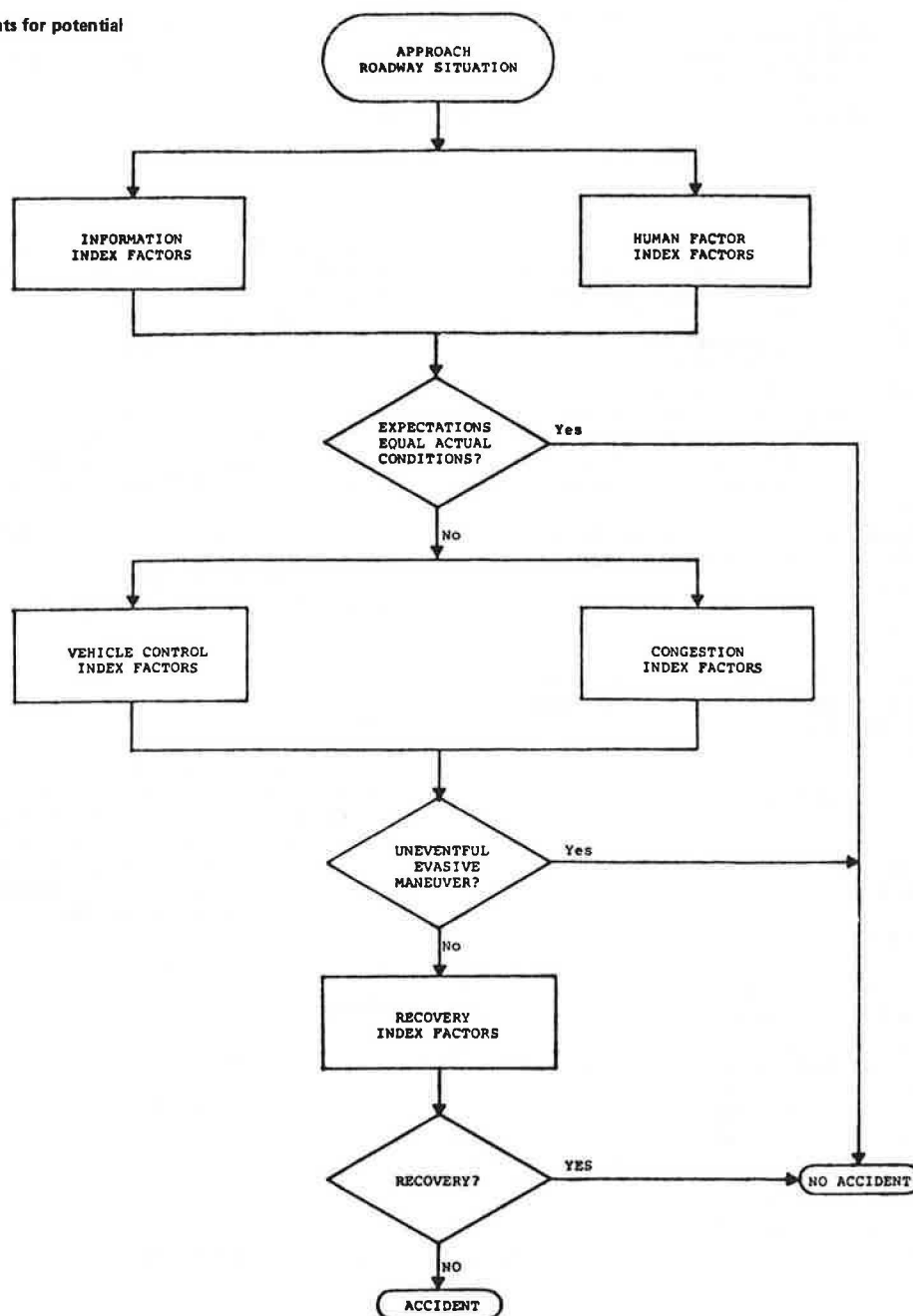


Table 1. Summary of selected surrogates by highway situation and type of highway safety analysis.

Highway Situation	Application in Highway Safety		
	Identification of Hazardous Locations	Evaluation of Countermeasures	Design Plan Review
Urban undivided tangent section	Access points/mile, turning volumes, speed changes/mile, fixed objects/mile	Speed changes/mile	Access points/mile, projected turning volumes
Rural undivided winding section	Curves/mile, lane width and shoulder width, physical evidence of driver error, speed changes/mile	Physical evidence of driver error, speed changes/mile	Curves/mile, lane width and shoulder width
Rural isolated curve	Speed reduction efficiency; curvature, grade, and distance since last curve; physical evidence of driver error; erratic maneuvers	Speed reduction efficiency, physical evidence of driver error, erratic maneuvers	Design speed differential; curvature, grade, and distance since last curve
Lane-drop location	Erratic maneuvers, merge gap availability, taper length, posted speed and sight distance	Erratic maneuvers, merge gap availability	Taper length, posted speed and sight distance
Narrow bridge	Ratio of bridge deck to pavement width, traffic mix, sight distance (time), physical evidence of driver error	Sight distance (time), physical evidence of driver error	Ratio of bridge deck to pavement width, traffic mix
Exit gore area	Deceleration lane length, sight distance, erratic maneuvers	Erratic maneuvers	Deceleration lane length, sight distance
Urban nonsignalized intersection	Traffic volume, approach speed and sight distance, traffic conflicts	Approach speed and sight distance, traffic conflicts	Projected traffic volume
Rural nonsignalized intersection	Traffic volume, approach speed and sight distance, traffic conflicts	Approach speed and sight distance, traffic conflict	Projected traffic volume
Rural undivided tangent section	Access points/mile, speed changes/mile, lane width, physical evidence of driver error	Speed changes/mile, physical evidence of driver error	Access points/mile, lane width
Rural signalized intersection	Traffic conflicts, traffic volume, sight distance, delay	Traffic conflicts, delay	Projected traffic volume, sight distance

ysis, (c) stepwise multiple regression analysis, and (d) independent-groups analysis. These tests were performed to obtain several types of quantitative information on the strengths of the relationships.

The analysis results provided varying degrees of support to the previously identified candidates. However, because the number of sites used in the analysis was relatively small, no candidate surrogate was eliminated from future consideration on the basis of these tests. Rather, the test results were used as another source of input (along with the literature review, workshop, and MIDAS analyses) in identifying those candidate variables that have a high probability of use as accident surrogates and therefore warrant further analysis.

CANDIDATE ACCIDENT SURROGATES

As a final step in the identification of candidate surrogates, each of the previously identified potential surrogates was evaluated according to five criteria, including

1. Relationship to accidents,
2. Clarity of definition,
3. Credibility,
4. Ease of data collection, and
5. Affectability.

Affectability is the likelihood that an improvement in the surrogate at a site will result in an improvement in the accident experience at that site. As an example, consider that the posted advisory speed at a horizontal curve is found to be a good indicator of the accident experience; i.e., higher accident rates become more likely as the posted advisory speed decreases. In the sense that this relationship is reasonably well established, posted advisory speed is a potential surrogate. However, it is clear that simply changing the advisory speed panel (to a higher value) will not result in an improvement in accident experience at a particular curve, because most likely this action

will increase accident frequency. Hence, even though the posted advisory speed might well be rated high on relationship to accidents, clarity of definition, credibility, and ease of data collection, it will be rejected as a surrogate for countermeasure evaluation on the basis of the affectability criterion.

Candidate surrogates resulting from the final screening process are shown in Table 1 by highway situation and type of safety analysis. Although each surrogate did not rate high on all of the criteria, each was considered at least passable on every criterion. These surrogates are considered worthy candidates for further study, development, and validation in that they exhibit a potential for producing a usable accident surrogate.

ACKNOWLEDGMENT

This paper reports some of the findings of a study performed by Goodell-Grivas, Inc., under the sponsorship of FHWA.

We acknowledge with thanks the contributions of Tapan K. Datta of Goodell-Grivas, Inc., William C. Taylor of the University of Michigan, and James I. Taylor of the University of Notre Dame.

The contents of this paper reflect our views, and we are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official policy of the U.S. Department of Transportation.

REFERENCE

1. Goodell-Grivas, Inc. Accident Surrogates for Use in Analyzing Highway Safety Hazards, 3 vols. FHWA, Repts. FHWA-RD-82-103 to 105, in preparation.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.

Skeleton Procedure for Evaluation of Highway Safety Improvements on a Road Network

D. MAHALEL

A systems approach to evaluation of safety improvements is described. The initial stage of the process is the identification of problem locations within the road network. This stage attempts to reduce the size of the problem and to make it solvable within a reasonable cost. The next stage consists of suggestions for improvement and an evaluation of their cost-effectiveness. This second stage is based on the definition of three standard design levels. The allocation process is defined as a mathematical programming problem, which ensures the optimal solution.

A method is presented for the allocation of a budget to a road network with the aim of minimizing the number of accidents. The emphasis is on the development of practical techniques based on an existing data base and relatively inexpensive implementation.

Sophisticated models based on extensive information and a large amount of data are often inoperable because of the nonavailability of relevant data. In addition, the allocation process itself is generally expensive and consumes a large portion of the available safety funds. The allocation procedure, therefore, is rarely applied. It is reasonable to assume furthermore that the procedure would not in any case be efficient because of high management and implementation costs.

The problem of the allocation of resources to ensure safety possesses some unique distinguishing features, as follows:

1. The size of the problem: The road network is composed of thousands of stretches of road and intersections that are potential sites for engineering improvement.

2. Multiplicity of alternatives: Because there are multiple reasons for road accidents, it is difficult to point to a specific solution. Therefore, there are a number of solutions for each location, distinguished by cost and effectiveness. The alternative solutions for each location are mutually exclusive, because the different locations are independent. The combination of mutually exclusive and independent solutions demands a sophisticated allocation procedure. For a particular budget, however, a specific combination of second-best solutions may be more efficient than the best alternative for each site.

3. Routine processes: Because of the dynamic nature of road use and environmental characteristics, a periodic review of the condition of the network is necessary. The socioeconomic changes in the vicinity of the highway cause changes in the exposure to road accidents, and thus there is need for relatively frequent routine allocations.

As a result of these three characteristics, the danger exists that the procedure for the allocation of funds will become too expensive; the allocation process itself may consume a large part of the budget and rarely will be carried out.

The following paragraphs describe the impact of engineering improvements on the prevention of road accidents. To reduce the size of the problem, a method is evolved that provides a preliminary definition of "black spots" (stretches of road on which the number of accidents is high relative to the daily traffic) and suggests a quick and efficient evaluation of the effectiveness of the engineering

projects. Finally, a budget-allocation model is presented.

ENGINEERING IMPROVEMENTS AS A TOOL FOR REDUCTION OF ROAD ACCIDENTS

Local engineering improvements are aimed at removing the black spots and bringing these locations to the general safety level of the entire infrastructure and in general at creating comfortable and uniform driving conditions throughout the network. The simplification of problems faced by drivers will help them to maneuver their vehicles in such a manner as to avoid accidents.

Driving on any stretch of road presents various problems, caused by the particular geometrical characteristics, traffic flow, traffic control system, weather, sight conditions, etc. Changes in the geometrical structure of the road, such as the existence of a curve, oblige drivers to maneuver their vehicles in a certain way in order to avoid an accident. The greater the radius of the curve, the less complicated the maneuvering is that is demanded and the greater are the chances of successful performance.

When the driver is faced with making a decision, a certain probability exists that an accident will be avoided; the probability is influenced by the complexity of the problem. The simpler the problem, the higher the probability--it may reasonably be assumed--that the driver will act according to an accident-preventive policy. Thus, Blumenthal (1) posited the event of an accident as a problem of faulty coordination between the level of performance of the driver and the performance demands of the road network.

Figure 1 (1) presents schematically the performance level of the driver and the performance demands of the road network as a function of time. The performance level of the driver varies because of such factors as fatigue, lack of attention, and illness. The demands of the network vary according to various levels of design, types of roadway, rates of traffic flow, etc. When the performance level of the driver is not compatible with the performance demands of the network, an accident occurs.

In this model, the significance of engineering improvements in the road infrastructure is reflected in an equivalent lowering of the performance demands. As a result, the gap grows between the performance level of the driver and the performance demands of the network, and the probability of road accidents lessens. In other words, engineering improvements in the road are designed to simplify the problems faced by the driver and to reduce the risk that the performance level of the driver will be less than the level required to meet the demands of the network.

The road network can be improved at various levels. On the one hand, it is possible to increase the number of motorways, interchanges, bypasses, etc. On the other hand, a lower level of improvements may be made, such as antiskid treatment, painting, and increasing the sight distance.

PRINCIPLES OF ALLOCATION METHOD

The method for allocating safety resources advanced in this paper is derived from that stage in the decisionmaking process at which the size of the safety budget for investment in the infrastructure is being decided. The question asked at this time concerns the location and content of the changes. Figure 2 demonstrates the basic structure of this method. Four steps are distinguished.

Step A: Safety Control of Road Network

The first step involves the preparation of a list of the black spots (short stretches of roadway and crossroads) where alternative engineering improvements may be suggested. This step is intended to decrease the dimension of the problem and to concentrate only on those elements that are thought to be worthwhile.

Step B: Alternative Projects for Improvement

From surveys undertaken at the sites of the black spots listed in step A, alternative projects are then devised for each site.

Step C: Definition of Cost and Benefit of Each Project

At this stage, an evaluation is made of the cost of the project and of the estimated safety benefit to be expected by a reduction in accidents.

Step D: Decision on Specific Projects to be Undertaken

Within the defined budget, a primary list of projects is prepared. The allocation of funds for these various projects is based on mathematical programming.

These four steps will now be discussed in detail.

SAFETY CONTROL OF ROAD NETWORK

The safety control of the road network is intended to identify the problem locations for which, at a later stage, the value of investment in safety means has to be examined. The immediate concentration on those elements in the road where safety benefits are most likely to be attained simplifies the problem of fund allocation, makes the entire procedure less costly, and enables routine, periodic inspection of the network. There is no doubt that a systematic, detailed scanning of the entire network, the compilation of a list of alternative programs for the network, and the allocation of funds over the entire network would ultimately lead to a more successful final allocation. In other words, an allocating system without safety control, i.e., without initial selection of black spots, would produce a greater reduction in accidents within a predetermined budget.

Figure 3 illustrates two hypothetical curves of the reduction in accidents as a function of the size of the budget. Curve A is derived from a combination of the investments without safety control, curve B with safety control. To the extent that the safety control is successful, the gap between the curves will decrease. The convexity of the curves demonstrates the decreasing marginal reduction in accidents with the increase in investment. This feature is caused by the fact that road accidents are not dispersed homogeneously throughout the road network.

Within the budget of C_1 , a reduction of n_1 accidents will be produced when safety control is

Figure 1. Hypothetical description of local failure of the system.

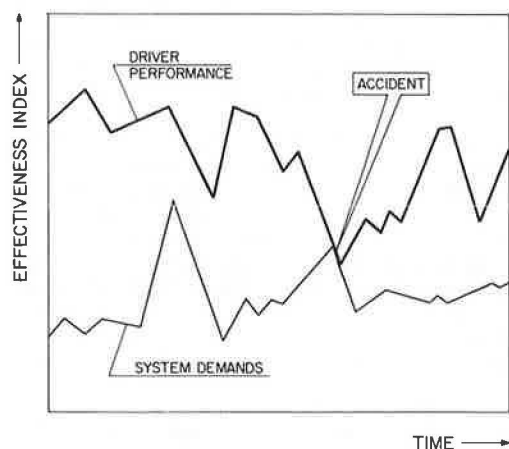


Figure 2. Schematic diagram of allocation process.

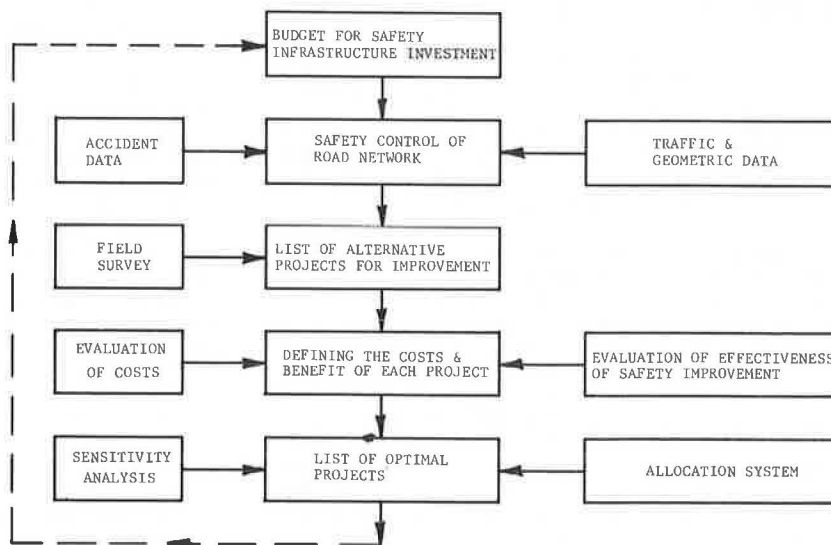
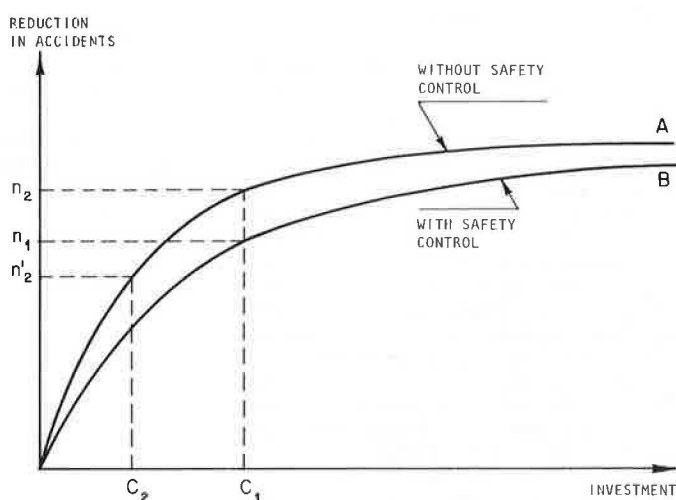


Figure 3. Accident-reduction curve with and without safety control.



carried out and n_2 accidents ($n_1 < n_2$) when no safety control is performed. As previously mentioned, however, the cost of compiling lists of alternative programs according to each system can be significant. If it is assumed that the additional cost of compiling a list of alternatives without safety control is $C_1 - C_2$ (Figure 3), this cost represents the planning expenses for large components of the network that will not be included in the final selection. If, therefore, C_1 represents a total budget with safety control, then C_2 ($C_2 < C_1$) will be designated as the actual investment without safety control. As a result, a budget of C_1 will help reduce n_1 accidents after safety control is applied and only n_2 accidents ($n_2 < n_1$) when safety control is not applied. This hypothetical example demonstrates the advantage of using safety control as a preliminary stage in the allocation process.

The exposure by safety control of those stretches and intersections where a high number of accidents occurs in relation to the number of opportunities for accidents does not explain the reasons for accidents at these black spots. This fact proves the lack of compatibility between the standards of the road and the traffic volume; in other words, there is an engineering deficiency in the road. The definition of the extent of noncompatibility between the demands of the road and the traffic volume cannot be made in absolute terms but only relative to the conditions existing throughout the road system during a specified period.

The identification of the black spots may be carried out in two stages:

1. Development of a model for estimating the potential for accidents and
2. Identification of those places where there is a gap between the actual number of accidents and the possible number of accidents as based on a probability model for accident occurrence at those locations.

The accident potential at location i can be described in the following manner:

$$\lambda_i = f(X_i, \Theta) \quad (1)$$

where

λ_i = potential or expectation of accidents on stretch of road i ,

$X_i = (X_{i1}, \dots, X_{im})$ = group of m independent variables describing stretch of road i , and

$\Theta = (\theta_1, \dots, \theta_m)$ = group of parameters of model.

The independent variables X_i are these: data on exposure (the number of vehicles per unit of time), data on the basic engineering characteristics (such as the number of lanes or stretch of road or intersection), and data on past accidents. Details of the models describing accident potential for the road network in Israel may be found in the last section.

The need to include the history of past accidents on the same stretch of road was treated with many reservations. The method of inclusion of the lag variable as an explanatory variable is accepted as an expression of dynamic models in economics. In this case, it became clear that the inclusion of the history of past accidents introduced an additional sensitivity into the model; that is, in addition to exposure and geometric structure, the model benefited from sensitivity to changes in the number of accidents from period to period. Empirically, the inclusion of the past accident history did not obscure the influence of exposure. Examination of the road section was then based on two lists of black spots, one without the past history and one with this explanatory variable. It became clear that the first list was a subgroup of the second.

Because the number of accidents on a stretch of road during a period of time is usually presumed to be Poisson distributed (2,3), it was impossible to use conventional regression methods in this case as a result of the dependence that existed between expectation and variance. It was necessary, therefore, to use the weighted least-squares method to achieve the best possible estimates that would be asymptotically normally distributed.

An intersection of stretch of road is declared a black spot when the following condition exists:

$$P[N(t) > Y_i | \lambda_i = f(X_i, \Theta)] = \sum_{n=Y_i}^{\infty} [\exp(-\lambda_i) \lambda_i^n / n!] < \alpha \quad (2)$$

where

α = level of significance,
 $N(t)$ = number of accidents during $(0, t)$, and
 Y_i = number of accidents on stretch of road i .

ALTERNATIVES AND EVALUATION OF EFFECTIVENESS

Once the list of black spots has been prepared, a field survey is carried out to inspect the existing engineering problems of the road. Because of the fact that road accidents are a result of several factors, it is often difficult to isolate the specific faults that cause an accident. On the other hand, it is often discovered that such a wide range of engineering problems exists that the number of possibilities for improvement is very large indeed. For example, at one specific location the following improvements were found to be useful: lighting, improving skid resistance, new painting, and cutting back the shrubbery. Although there were four suggested improvements, the number of possible combinations of alternatives is $2^4 = 16$, which includes the do-nothing alternative. Evaluating the extent of effectiveness of these various improvements presents a problem. At the current state of the art, there are no reliable estimates that are useful for measuring the effectiveness of single improvements. It is even more difficult to evaluate the effectiveness of various combinations, because an interdependence is found among the individual improvements.

In order to simplify the process of identifying specific faults and to bypass most of the problems in that connection, three standards of road planning may be defined: initial (preparatory) planning, medium (secondary) planning, and the completed solution of the problem. For each group of standards, the necessary elements, as well as the required set of inputs to attain that standard, should be set out in detail. Thus, the engineering team sent out into the field needs only to fill out the routine forms containing the necessary improvements to reach each standard and does not have to initiate solutions of its own.

The effectiveness of each planning level is measured in accordance with the safety conditions existing at other locations on the road network where the standards already exist. This simple method, besides solving problems of planning and of estimating effectiveness, also solves most of the problems of costing. Because this method constitutes a modular view of planning, standard units of cost can be fixed for each modular unit, and thus initial costing evaluations can be derived without detailed planning of the project.

BUDGET ALLOCATION

Because of the complexity of the allocation problem, the use of accepted engineering economics techniques, such as cost-benefit and internal rate of return, might produce only a nonoptimal solution. The reason is that at every black spot, there are a number of substitutable alternatives distinguished by cost and effectiveness. It could happen that within the existing budget, the most effective combination of solutions over the road network will be such that the most worthwhile projects for a number of locations will be the second-best alternative or even the third best. In other words, a method based on the choice of the best alternatives for each location will not necessarily ensure that the final combination of investments will be optimal. Thus, it is necessary to use a more sophisticated technique of mathematical programming that will promise the systematic scanning of all possible alternatives.

According to McFarland and others (4), an optimal solution can be reached by either integer or dynamic programming. In this study, it was decided to adopt an integer-programming solution. The advantages of using integer programming were the availability of

appropriate computer software and the special structure of the problem, which promised an almost-integer solution by using linear programming. The linear-programming (LP) solution enables us to conduct sensitivity tests.

Mathematically, the safety-fund allocation problem can be stated as follows:

$$\text{Min } \sum_{i=1}^N \sum_{j=0}^{N_i} b_{ij} X_{ij} \quad (3)$$

Subject to:

$$\sum_{i=1}^N \sum_{j=0}^{N_i} C_{ij} X_{ij} \leq C \quad (4)$$

$$\sum_{j=0}^{N_i} X_{ij} = 1, \quad i = 1, \dots, N \quad (5)$$

$$X_{ij} \geq 0 \text{ for all } i, j \quad (6)$$

$$X_{ij} = 0, 1 \text{ for all } i, j \quad (7)$$

where

- b_{ij} = number of accidents at location i after implementation of project j ,
- C_{ij} = cost of project i at location j ,
- C = size of budget,
- X_{ij} = decision variable regarding implementation of project j at i ,
- X_{i0} = do-nothing alternative,
- N = number of locations, and
- N_i = number of projects at i .

Constraints 4, 5, and 6 by themselves constitute an LP problem, whereas constraint 7 turns the problem into one of integer programming. The constraint matrix is shown in Figure 4. This structure of the problem promises the possibility of exclusion of constraint 7; despite that, the optimal solution to the LP problem (5-7) will be almost integer. According to Lasdon (5), any basic feasible solution of a linear program has the following property: at least $N - 1$ of the indices i have one X_{ij} positive (and hence unity).

In addition, it is possible to arrive at a solution by using algorithms of generalized upper bounding to reduce the magnitude of the problem. As a result, for some locations there will be no integer solution; instead, the solution will have a linear combination of two alternatives. In fact, at location i it is usually possible to define a revised project within a certain budget framework,

$$\sum_{j=1}^{n_i} C_{ij} X_{ij}$$

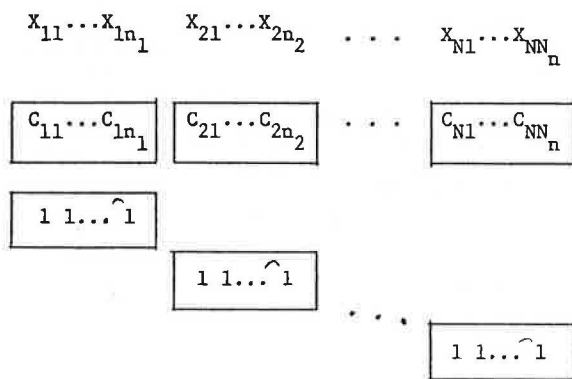
which will correspond to accidents represented by

$$\sum_{j=1}^{n_i} b_{ij} X_{ij}$$

Of course, it is always possible to add constraint 7 and to search for the integer solution; however, this has three drawbacks:

1. The problem becomes complex; the process may consume much computer time.
2. Usually, as will be shown in the example given below, not all the budget will be exploited in the integer solution, so that the general reduction in the number of accidents by means of the integer solution may be less than in the approximate LP solution.

Figure 4. Constraint matrix.



3. Most of the sensitivity tests that are possible by means of the LP solution are impossible by means of the integer solution.

CASE STUDY

The following limited example was worked out in order to test the practicability of the procedures suggested. It shows the advantage of solving the problems as an LP problem rather than searching for the integer solution. The case study also illustrates the superiority of both of those methods over the traditional methods used in engineering economics.

Fifteen locations were selected on the interurban network in Israel: 10 stretches of road, 4 black spots, and 1 intersection. At each location, different levels of improvement were defined. The number of accidents for the do-nothing alternative was taken as the expected number of accidents over the coming 20 years, on the assumption that no improvements would be carried out. The number of accidents for each level of improvement was estimated for the same 20-year period. If the life of an improvement was less than 20 years, it was assumed that at the end of its life the level of accidents would return to the do-nothing level.

Figure 5 shows the improvements in the number of accidents under different budget levels. Both an integer and a continuous solution for the LP problem are presented and compared. As expected, the LP solution presents higher benefits than does the integer solution. In most cases, a scheme for improvement can be defined that will be equivalent to a linear combination of two levels of improvement, as determined by the continuous LP solution. Figure 5 also presents the results obtained by the cost-effectiveness solution. As can be seen, allocation by the cost-effectiveness method is inferior to that by the mathematical-programming solution.

SUMMARY AND DISCUSSION

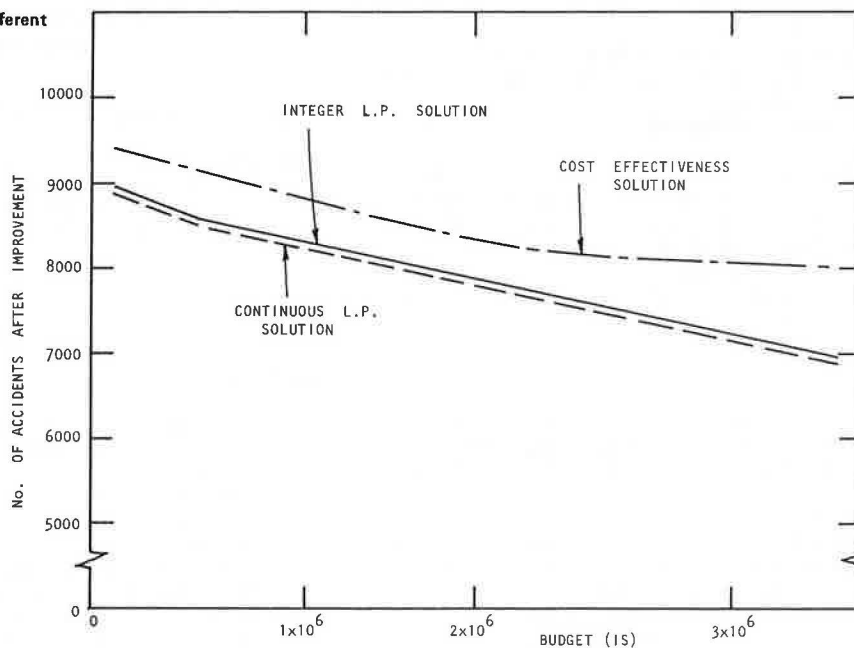
This paper has described a systems approach to the solution of the problem of allocation of safety resources for black spots. Special attention was given to a level of applicability within limited means of management costs, background data, and labor time to reach a decision based on the allocation of the safety budget. From this point of view, the model should be classified as a skeleton model, because it is based on practical information and provides input for a more-detailed planning process in the future.

The extent of the problem arising from the size of the road network initially demands concentration on problem locations within the network. This process is performed by constructing a model of accident potential and by isolating stretches of road and intersections where significant differences exist between the accident potential and the actual number of past accidents.

The next stage in the process is the suggestion of improvements and an evaluation of their cost and effectiveness. In order to perform this stage efficiently, three levels of design standards were defined. The field team was asked only to designate the necessary improvements to bring the black spots to the level of each of these standards. The effectiveness of these improvements is measured by comparing the safety conditions with conditions existing at other locations where these standards already exist.

The input of the allocation problem is a group of

Figure 5. Allocation results of three different methods.



black spots; at each spot, in addition to the do-nothing alternative, there exist at least three other alternatives (three design levels), which are mutually exclusive. Thus, the range of alternative solutions consists of a mixture of independent, mutually exclusive projects. The optimal solution can be found by solving it as a problem of integer programming. The structure of the constraint matrix creates a situation in which the solution of the problem by linear programming is almost integer; for practical purposes, this solution is usually sufficient.

The advantage of the allocation method advanced here shows itself when the allocation budgets are relatively small. This is due to the ability of the integer-programming method to systematically scan all possible solutions and to locate the best one. Some of the projects selected for the optimal solution are, however, second best. As the total budget grows, the advantages of this method decrease in comparison with the classical methods.

ESTIMATE OF ACCIDENT POTENTIAL

Models defining the accident potential of road sections and intersections were estimated for the road network in Israel. The estimating process and the final structure of these models are described in this section.

Road Sections

On road sections, the independent variables used in this study were traffic flows, number of accidents at a location in the period prior to that for which the expectation was being estimated, and the number of carriageways.

The model adopted for further study was of the following form:

$$E(Y_i) = \theta_0 + \theta_1 X_{i1} + \theta_2 X_{i2} + \theta_3 X_{i3} \quad i = 1, 2, \dots \quad (8)$$

where

- $E(Y_i)$ = expected number of accidents on section i during 1970-1972,
- X_{i1} = average daily traffic flow on section i ,
- X_{i2} = past number of accidents on section i ,
- X_{i3} = dummy variable denoting whether section i is single or dual carriageway, and

$\theta_0, \theta_1, \theta_2, \theta_3$ = parameters of regression equation.

Information on the number of carriageways was included in the model by means of a dummy variable (X_{i3}); values of 0 were taken for single carriageways and 1 for dual carriageways. This variable was later found to be not statistically different for the two types of roads, and therefore it was omitted in the final model.

Intersections

The relationship between accidents and a number of variables was also studied at both urban and interurban intersections. The independent variables studied were (a) an index of traffic flows, (b) the number of accidents in the previous period, (c) the length of time the intersection had been signalized, (d) the number of conflict points, and (e) the town in which the intersection was located.

At the 202 urban intersections studied, the correlation between flow index and accidents was

$R = 0.74$ and between past accidents (1967-1970) and accidents in the previous period (1971-1972), $R = 0.87$. No consistent relationship was found between accidents and signalization or number of conflict points, so these variables were therefore dropped from the final procedure. The dummy variable for the different towns also contributed very little and was also dropped. Because of different traffic characteristics, separate models were estimated for urban and interurban intersections.

The model finally adopted was of the following form:

$$E(Y_i) = \theta_0 + \theta_1 X_{i1} + \theta_2 X_{i2} \quad (9)$$

where

$E(Y_i)$ = number of injury accidents in 1971-1972,

X_{i1} = number of injury accidents in 1967-1970,

X_{i2} = index of traffic flows, and

$\theta_0, \theta_1, \theta_2$ = constants.

Estimate of Model Parameters

The final models adopted were estimated by means of a multiple linear regression. The regression coefficients were calculated by a weighted least-squares method, as developed by Jorgensen (6) and Weber (7). In the case that a dependent variable was Poisson distributed, the expectation would be equal to the following variance:

$$\lambda_i = E(Y_i) = \text{Var}(Y_i) \quad (10)$$

In such a case, the least-squares estimates obtained by the normal procedure were unbiased but not consistent.

The final formulas fitted to the different locations were as follows:

1. Road sections: In the final analysis, 1630 road sections were included, and the equation fitted was of the following form:

$$E(Y_i) = 0.023 + 4.225 \times 10^{-4} X_{i1} + 0.499 X_{i2} \quad (11)$$

The multiple correlation coefficient was $R = 0.82$.

2. Urban intersections: The final analysis included 202 urban intersections, and the equation fitted was

$$E(Y_i) = 0.507 + 0.514 X_{i1} + 2.724 \times 10^{-5} X_{i2} \quad (12)$$

The multiple correlation coefficient was $R = 0.81$.

3. Interurban intersections: There were 40 interurban intersections in the final analysis, and the equation fitted was

$$E(Y_i) = 2.546 + 0.514 X_{i1} + 1.330 \times 10^{-5} X_{i2} \quad (13)$$

The multiple correlation coefficient was $R = 0.83$.

REFERENCES

1. M. Blumenthal. Dimensions of Traffic Safety Problem. Presented at Automotive Engineering Congress, Society of Automotive Engineers, Detroit, MI, 1967.
2. M. Greenwood and G.U. Yule. An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. Journal of the Royal Statistical Society, Vol. 83, 1920.

3. P.A. Hall. The Identification of High Accident Location. Presented at 11th International Study Week in Traffic Engineering and Safety, Brussels, Belgium, 1972.
4. W.F. McFarland, L.I. Griffen, J.B. Rollins, W.R. Stockton, D.T. Phillips, and C.L. Dudek. Assessment of Techniques for Cost-Effectiveness of Highway Accident Countermeasures. FHWA, Rept. FHWA-RD-79-52, Jan. 1979.
5. L.S. Lasdon. Optimization Theory for Large Systems. Macmillan, London, 1970.
6. D.W. Jorgensen. Multiple Regression Analysis of the Poisson Process. Journal of the American Statistical Association, Vol. 56, 1961.
7. D.C. Weber. Accident Rate Potential: An Application of Multiple Regression Analysis of a Poisson Process. Journal of the American Statistical Association, Vol. 66, 1971.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.

Evaluating Need for Accident-Reduction Experiments

WILLIAM D. BERG AND CAMIL FUCHS

New traffic control devices or new applications of existing devices are frequently proposed as a means of facilitating the driving guidance and control process and thereby improving traffic safety. Before such changes can be approved at the national level, some research must be undertaken to evaluate the potential safety effectiveness of the new device. Safety effectiveness can be measured directly in terms of a reduction in accident rate or indirectly in terms of a change in an alternative measure of effectiveness. A requirement that accident data be collected before a new traffic control device standard or guideline is approved may itself be impractical and/or not cost-effective. A four-step methodology is presented for quantitatively addressing the need to undertake accident-reduction research experimentation. Statistical analysis and sampling requirements are developed first. This is followed by a determination of the minimum accident-rate reduction that would economically justify nationwide deployment of the new traffic control device treatment. The cost-effectiveness of alternative experimental designs is then evaluated. The final step is a trade-off analysis of the value of information to be derived versus the cost of obtaining the information. A case study application of the methodology is also presented.

New traffic control devices or new applications of existing devices are frequently proposed as a means of facilitating the driving guidance and control process and thereby improving traffic safety. Before such changes can be approved at the national level, some research must be undertaken to evaluate the potential safety effectiveness of the new device. In addition, the costs required to implement the new traffic control device treatment, under either an as-needed or an immediate-replacement policy, must be evaluated.

Safety effectiveness can be measured directly in terms of a reduction in accident rate or indirectly in terms of a change in an alternative measure of effectiveness. Examples of the latter include vehicle speed profiles, variance in lateral placement of vehicles within a roadway lane, driver head and/or eye movements, and various types of traffic conflicts as defined by procedures for traffic-conflicts analysis (1-3). Regardless of whether accident data or alternative measures of effectiveness are used, the principal issue is how much information is necessary to make a reasonably confident decision about potential safety cost-effectiveness.

A requirement that accident data be collected and evaluated before a new traffic control device standard or guideline is approved may itself be impractical and/or not cost-effective. If this is the case, then a decision about approval of the new traffic control device must be based on an evaluation of alternative measures of effectiveness. This would require an assumption about the true relationship between accident rate and the alternative measure. Because this is usually a qualitative judge-

ment, there can be substantial differences of opinion about potential safety effectiveness and therefore a lack of necessary support for what may actually be a very cost-effective standard or guideline.

The purpose of this paper is to present a methodology for analytically addressing these issues. The methodology was developed during research on an experimental design for evaluating the safety benefits of railroad advance-warning signs (4). The results from that case study will be used to demonstrate the application of the methodology.

EVALUATION METHODOLOGY

The evaluation methodology is designed to address the following basic questions relative to proposed research experiments of the accident-reduction potential of a new traffic control device standard or guideline:

1. What are the sampling requirements based on a treatment-control comparison?
2. What is the critical, or minimum, accident-rate reduction that the experimental design should be capable of detecting?
3. What is the cost-effectiveness of alternative experimental designs?
4. What is the value of the information to be derived from the experiment?

The evaluation methodology is therefore presented as a four-step process.

STATISTICAL ANALYSIS AND SAMPLING REQUIREMENTS

It is assumed that a treatment-control comparison is to be used in the analysis of the experimental data, although this can be supplemented with a before-and-after analysis if desired. With a treatment-control type of design, one group of sites is selected to receive or be treated with the proposed new control device application. A second group of sites would be selected as a control or base against which measured changes in accident rate at the treatment sites can be compared.

The sampling scheme is composed of two parts. First, the selected population of study sites would be divided into k homogeneous sets, each composed of n similar sites (where $n = 1, 2, 3, 4, 5$, or larger). Then, from each of the k sets, one of the

n sites would be randomly selected to be in the treatment group. All of the other n - 1 sites would be included in the control group.

There are two advantages to this sampling scheme. First, the direct comparison between the treatment site and the respective sites in the control group for each set ensures that the comparison is performed among sites as homogeneous as possible, thereby permitting random fluctuations to be minimized. Second, the k sets of homogeneous sites can later be consolidated into a smaller number of sets, or scenarios. Although each resulting scenario would not be as homogeneous as one of the original sets, it would still be possible to determine whether the change in traffic control device had a statistically significant effect on accident experience. The various scenarios would reflect a cross-classification of the highway population in terms of appropriate design and operational characteristics.

The comparison of accident rates between treatment and control sites can be made on the basis of the overall accident rates for the two groups of sites or in terms of subsets of sites having similar characteristics. In both cases it is assumed that accidents are rare events and are therefore Poisson distributed. For the overall comparison of treatment and control sites, it can be shown that by using the normal approximation to the Poisson distribution and applying the correction for continuity, the statistic for testing the null hypothesis of no effect is

$$Z = \frac{(1/k) \sum_{j=1}^k \{Y_j - (1/2)\} - \{\bar{X}_j + [1/2(n-1)]\}}{\sqrt{\sum_{j=1}^k [Y_j + (n-1)\bar{X}_j] / (n-1)k^2}} \quad (1)$$

in which Y_j is the number of accidents at the treatment site in set j and \bar{X}_j is the mean number of accidents for the control sites in set j (5). The null hypothesis will be rejected at the 5 percent significance level if $Z < -1.64$; the conclusion is that the new traffic control device has a significant effect in reducing accidents.

Rather than an analysis of all the sample sites as a group, it may be of interest to examine subsets of sites, each of which exhibits similar characteristics. The subsets can be identified as scenarios, $s = 1, 2, \dots, m$, that represent combinations of the k sets of sites described previously. If it is assumed that the summations of accidents for the treatment and control sites within each scenario are Poisson distributed and if the correction for continuity is applied, it can be shown that the statistic for testing the null hypothesis to be of no effect is

$$Z_s = \frac{\{\bar{Y}_{.s} - (1/2n_{s1})\} - \{\bar{X}_{.s} + (1/2n_{s2})\}}{\sqrt{\{[(n_{s1}\bar{Y}_{.s} + n_{s2}\bar{X}_{.s}) / (n_{s1} + n_{s2})]^{1/2} [(1/n_{s1}) + (1/n_{s2})]^{1/2}\}}} \quad (2)$$

in which $\bar{Y}_{.s}$ is the mean number of accidents per site for the treatment group in scenario s, $\bar{X}_{.s}$ is the mean number of accidents per site for the control group in scenario s, and n_{sg} is the number of sites in scenario s, where $g = 1$ for the treatment group and $g = 2$ for the control group (5). Under the null hypothesis, Z_s^2 has a chi-square distribution with one degree of freedom, and $\sum_{g=1}^m Z_s^2$ has a chi-square distribution with m degrees of freedom.

The null hypothesis that the accident rates for the treatment and control sites over all the scenarios are equal will be rejected if the value of $\sum_{g=1}^m Z_s^2$ exceeds the critical value in the chi-square table. Furthermore, those scenarios for which there is a statistically significant dif-

ference between the accident rates for the treatment and the control groups can be identified by examining the respective values of Z_s^2 . The sign of Z_s will indicate whether the treatment of the control group has the lower accident rate. If the sign is negative, it can be said that the new traffic control treatment provides a statistically significant reduction in the expected accident rate.

The partitioning of accident rates by treatment and control groups and by scenario also offers the opportunity to apply a two-way analysis of variance (ANOVA). This procedure as well as that for a before-and-after comparison are described elsewhere (5-8).

The sample size required for the treatment-control study depends on the following parameters:

1. The desired power of the test (β),
2. The value of the overall mean accident rate ($\bar{\lambda}$), and
3. The expected percentage of reduction in accident rate (Δ).

The power is the probability of correctly detecting a change in accident rate, if there is a change (8). For a fixed mean accident rate ($\bar{\lambda}$) and a fixed percentage of accident-rate reduction (Δ), the required sample size will vary directly with the desired power. Furthermore, there is a greater likelihood in detecting a change in accident rate when the rate of accidents is high than when accidents are a rare event. Finally, it is clear that a larger change is more likely to be detected than a smaller one.

The sample-size relationship presented below was derived by assuming that the principal statistical analysis would be the overall comparison of all treatment and control sites. A larger sample size would be required to achieve the same power for the analysis of scenarios or for before-and-after analyses. The sampling scheme is designed to create k homogeneous sets of n sites, where one site will be randomly selected as a treatment site and the remaining n - 1 sites will serve as control sites. If we assume that accidents are Poisson distributed and use the normal approximation to the Poisson distribution, it can be shown that the total number of homogeneous sets of n sites needed to test the null hypothesis that Δ equals zero at the 5 percent significance level is

$$k = [\Phi^{-1}(\beta) + 1.64]^2 [n/(n-1)] (1/\bar{\lambda}\epsilon^2) \quad (3)$$

in which $\Phi^{-1}(\beta)$ is the inverse of the standard normal distribution at point β , $\bar{\lambda}$ is the mean accident rate over all sets, and ϵ is the expected change in accident rate (Δ) expressed as a fraction (5). The value of $\Phi^{-1}(\beta)$ can be easily determined from a table of the cumulative standard normal distribution as the value of the standard variate that yields a cumulative probability of β . The total number of sets (k) of one treatment and n - 1 control sites is in effect the desired sample size.

CRITICAL ACCIDENT-RATE REDUCTION

Because highway traffic accidents at a given study site are rare events, the overall mean accident rate ($\bar{\lambda}$) can be very small. This can create the need for very large required sample sizes. It is therefore possible that no experimental design would be statistically sensitive to small changes in accident rate, be feasible in terms of site availability, and be economically practical to conduct.

The smallest relative change in accident rate

that would need to be detected can be defined as the critical accident-rate reduction. Quantitatively, this is the minimum relative reduction in accident rate that would economically justify nationwide deployment of the new traffic control device treatment. To determine the critical accident-rate reduction, an economic decision model must first be specified (9,10). By using a net-present-value (NPV) criterion, the model is expressed as

$$NPV = PVB - PVC \quad (4)$$

in which the present value of benefits (PVB) is the present dollar value of a time stream of benefits and the present value of costs (PVC) is the present dollar value of costs over the same time period. If the NPV of an investment situation is greater than zero, then that investment is considered to be economically feasible. When mutually exclusive alternatives, each with a positive NPV, are compared, that alternative with the highest NPV is preferred.

PVB is defined as the present dollar value of future accident-rate reductions attributable to the nationwide deployment of the new traffic control device treatment. PVC is defined as the present dollar value of the costs of deploying and maintaining the new treatment on a nationwide basis. In addition to a policy of immediate deployment of the new treatment, it may be appropriate to consider an as-needed replacement policy that would permit gradual implementation over a period of years.

The actual formulation of the PVB and PVC functions will depend on the nature of the proposed traffic control device treatment. In general, PVB for an immediate-deployment policy can be expressed as

$$PVB = (AAR)(\Delta)(AC)(N)(SPW_{i,n}) \quad (5)$$

where

AAR = present average annual accident rate per site,

Δ = percentage of reduction in accident rate (AAR) due to increased effectiveness of new traffic control device,

AC = average dollar cost of an accident,

N = number of sites at which treatment is to be deployed, and

$SPW_{i,n}$ = series present-worth factor for discount rate of i percent and analysis period of n years.

PVC can be expressed as

$$PVC = 2N [(\Delta C + LC + MC) + (\Delta C/m^2)(GPW_{i,m}) + (\Delta C/m)(SPW_{i,n-m})(PW_{i,m})] \quad (6)$$

where

ΔC = dollar materials cost difference between current and proposed new traffic control device treatment,

LC = dollar labor cost for deploying each new treatment,

MC = dollar mileage cost per treatment for installation crew,

m = average life of new treatment (years),

$GPW_{i,m}$ = uniform gradient present-worth factor for discount rate of i percent over $n-m$ years, and

$PW_{i,m}$ = present-worth factor for discount rate of i percent over m years.

The smallest relative change in accident rate

that the experimental design should be capable of statistically detecting is determined by applying the economic decision model and determining the relative accident-rate reduction (Δ) that would yield an NPV of zero. Any reduction in accident rate smaller than this value would be insufficient to economically justify deployment of the proposed new traffic control treatment. Any accident-rate reduction larger than this value would mean that deployment of the new treatment could be economically justified. Thus it would be important to be able to detect accident-rate reductions as small as the critical value found at the break-even point.

EXPERIMENTAL DESIGN COST-EFFECTIVENESS

As discussed previously, the required sample size for a field evaluation of accident-reduction potential consists of k treatment sites and k ($n - 1$) control sites. The magnitude of k is given by Equation 3 and can be seen to

1. Increase with the desired power (β),
2. Decrease with the level of the overall mean accident rate (λ),
3. Decrease with the square of the change in accident rate (ϵ), and
4. Decrease slightly with the size of the sets (n).

The value selected for each of these parameters will govern both the effectiveness of the experimental design in terms of its ability to detect actual changes in accident rate and the cost of conducting the experiment and analyzing the data. Thus, several alternative experimental designs can be specified and then evaluated for their potential cost-effectiveness.

The selection of the power of the test (β) is a subjective decision. Increasing its value reduces the likelihood of not statistically detecting a change in accident rate, if in fact one occurs. From a safety standpoint, it is clearly important to make the value of this parameter as large as practical. A range of values, for example, between 50 and 90 percent, could be used to specify alternative experimental designs and sampling requirements.

The overall mean accident rate (λ) is a function of the roadway situation under study. Generally, typical accident-rate data would be available from accident records systems or safety publications. Whatever the traffic control device treatment, the annual accident rate can be expected to be relatively low. This has the effect of requiring large, and possibly very large, sample sizes. Therefore, it may be necessary to consider a multi-year rather than a one-year study period. For example, the overall mean accident rate (λ) for a three-year study is three times as large as the average annual accident rate. This would have the effect of reducing the required number of treatment and control sites by a factor of 3.

The selection of a change in accident rate (ϵ) that the experimental design should be capable of detecting can be approached in two ways. First, a subjectively established range of minimum values can be used. Alternatively, the critical accident-rate reduction based on benefit-cost considerations can be employed. In either case, the smaller the value, the larger the required sample size.

The final parameter that can be varied is the size of the treatment-control sets (n). As n increases, the required number of treatment sites is reduced but the total number of treatment plus control sites is increased. The most desirable ratio depends on the relative cost of preparing the treat-

ment sites versus the cost of data collection at all treatment and control sites.

By varying each of the above four parameters, a set of alternative sampling plans can be specified in terms of the required number of treatment sites (k) and control sites [$k(n-1)$]. The cost-effectiveness of each alternative experimental design can then be examined in terms of the cost of conducting the study, the availability of treatment and control sites, and the smallest relative change in accident rate that can be detected with an acceptable power of test. Boundary conditions may constrain the range of feasible experimental designs. For example, a required sample size may exceed the number of available sites. Alternatively, a maximum budget level may limit the number of sites that can be selected and thus the minimum relative change in accident rate that can be detected during a reasonable data-collection period. In general, the cost of conducting an accident study will increase as the relative change in accident rate to be detected is decreased, the power of the test is increased, and the data-collection period is decreased.

VALUE OF INFORMATION

The ultimate question is which, if any, of the alternative experimental designs should be undertaken. For a given study cost, the most effective design is that which has the potential of detecting the smallest relative change in accident rate with an acceptable power of test. As the study budget is increased, the effectiveness of the experimental design will generally increase, but at a diminishing rate. Therefore, the selection of an experimental design must consider that trade-off between the value of the information to be derived and the estimated cost of obtaining that information.

This trade-off can be examined in two ways. First, the smallest accident-rate reduction that is likely to be statistically detectable can be compared with the lowest rate that would economically justify the nationwide application of the new traffic control treatment. Second, the cost of undertaking the accident study can be compared with the cost of deploying the new treatment nationwide. These comparisons must then be interpreted with respect to both experimental and economic practicality.

If the sample sizes necessary to detect the critical accident-rate reduction exceed the population of available sites, then the study would clearly be impractical because no experimental design could be expected to detect all possible accident-rate reductions that would economically justify deployment of the new traffic control treatment. Similarly, if the estimated cost of conducting the most cost-effective study designs were to equal or exceed the approximate total cost of nationwide deployment of the new traffic control treatment, then the study would be economically impractical. Finally, if the estimated study cost necessary to detect the critical accident-rate reduction was simply considered to be too expensive, then the study would also be considered impractical.

If the final decision is that a practical, cost-effective experimental design is available, then the study should be initiated. However, if no such study design can be found, then three options exist. First, no further action of any type would be taken. This choice should only be favored if the proposed traffic control treatment is subjectively judged to have little merit. Second, experimental research could be conducted by using alternative safety measures of effectiveness instead of accident-rate data. Finally, the proposed new traffic

control treatment could be approved for use without further experimental research. This option should only be favored if the critical accident-rate reduction is very low and the new traffic control treatment is judged to offer positive (albeit unmeasured) safety benefits.

APPLICATION OF METHODOLOGY

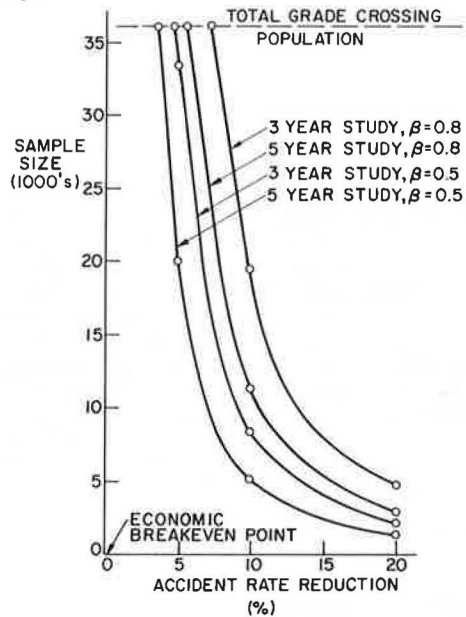
As described in detail elsewhere, the above methodology was used in a study to develop an experimental design for evaluating the safety benefits of a new railroad advance-warning sign (4,5). The sample population of interest consisted of the total nationwide population of 36 104 railroad-highway grade crossings that have reflectorized crossbucks and standard advance-warning signs. The overall mean accident rate ($\bar{\lambda}$) was 0.049 accident per crossing per year. Three-year and five-year expected accident rates were then calculated to reflect the expected mean accident rates for a three-year and a five-year study. The relative change in accident rate to be detected was varied from 5 to 20 percent, and the ratio of treatment to control sites was varied from 1:1 to 1:4. A 50 percent and an 80 percent power of test were also specified.

By using Equation 3, it was found that to be 50 percent confident of detecting an actual 5 percent reduction in accident rate over a three-year study period, it would be necessary to select almost 15 000 treatment and 15 000 control sites. The required number of treatment sites could be reduced to approximately 11 000 if desired, but the number of control sites would then have to be increased to a little more than 22 000. Thus by decreasing the ratio of treatment to control sites from 1:1 to 1:2, the total number of required sites would increase from approximately 30 000 to 33 000. Of greater significance, however, was the fact that both sample sizes nearly equaled the total population of 36 104 grade crossings. If the ratio of treatment to control crossings were to be reduced any further, the total required sample size would exceed the population size. Increasing the power of the test to 80 percent and assuming a relatively long five-year data-collection period resulted in the same finding.

The critical accident-rate reduction was then calculated and found to vary over a range of 0.01-0.13 percent, depending on the sign-deployment policy and the unit cost data used. This was equivalent to a reduction of about one grade crossing accident in five to six years over the total population of 36 104 grade crossings. These results clearly suggested that it might be both experimentally and economically impractical to attempt to determine whether the actual safety effectiveness of the new advance-warning sign would justify its deployment on a nationwide basis.

The trade-off among the smallest detectable accident-rate reduction, the required sample size, and the accident-rate reduction associated with the economic break-even point for justifying deployment of the new advance-warning sign was examined by preparing the graph shown in Figure 1. It is clear that none of the alternative experimental designs could be expected to provide the information necessary to establish whether the potential safety benefits of the new sign would exceed the total cost of nationwide deployment. When the six most cost-effective experimental design alternatives were compared with the estimated cost of nationwide deployment of the new sign, it was found that the cost of four of the alternative experimental designs would significantly exceed that total initial cost of deploying the new sign on an as-needed basis over a seven-year period. Moreover, these study costs fell within approx-

Figure 1. Sample-size requirements: total nationwide population of rural and urban grade crossings with reflectorized crossbucks and advance-warning signs.



imately 30-75 percent of the total initial cost of an immediate nationwide sign-replacement policy. It was therefore concluded that the proposed research study would be both experimentally and economically impractical and should therefore not be undertaken.

CONCLUSIONS

It is believed that the methodology described above is both generalizable and practical. It can provide a quantitative basis for decisionmaking where strong differences in personal opinion may exist regarding the need for accident data as a precondition for approving a new traffic control device treatment. Application of the methodology in these situations can assist in making rational choices, avoiding needless experimentation, and facilitating early decisions and timely realization of the benefits of meritorious new traffic control device treatments.

ACKNOWLEDGMENT

This paper is based on a study sponsored by the Office of Research of the Federal Highway Administration. The contents reflect our views, and we are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official view or policy of the U.S. Department of Transportation.

REFERENCES

1. W.A. Stimpson, W.A. Kittelson, and W.D. Berg. Methods for Field Evaluation of Roadway-Delineation Treatments. TRB, Transportation Research Record 630, 1977, pp. 25-32.
2. J.S. Koziol and P.H. Mengert. Railroad Grade Crossing Passive Signing Study. Transportation Systems Center, U.S. Department of Transportation, Cambridge, MA, Final Rept. FHWA-RD-78-34, Aug. 1978.
3. W.D. Glauz and D.J. Higletz. Application of Traffic Conflicts Analysis at Intersections. NCHRP, Rept. 219, Feb. 1980.
4. W.D. Berg, C. Fuchs, and J. Coleman. Evaluating the Safety Benefits of Railroad Advance-Warning Signs. TRB, Transportation Research Record 773, 1980, pp. 1-6.
5. W.D. Berg. Experimental Design for Evaluating the Safety Benefits of Railroad Advance Warning Signs. FHWA, Rept. RHWA-RD-79-78, April 1979.
6. F. Mosteller, and J.W. Tukey. Data Analysis and Regression. Addison-Wesley, New York, 1977.
7. F.I. Anscombe and J.W. Tukey. Examination and Analysis of Residuals. Technometrics, Vol. 5, 1963, pp. 141-160.
8. K.A. Brownlee. Statistical Theory and Methodology in Science and Engineering. Wiley, New York, 1960.
9. R. Winfrey. Economic Analysis for Highways. International Textbook Company, Scranton, PA, 1969.
10. A Manual on User Benefit Analysis of Highway and Bus-Transit Improvements. American Association of State Highway and Transportation Officials, Washington, DC, 1977.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.

Common Bias in Before-and-After Accident Comparisons and Its Elimination

EZRA HAUER AND BHAGWANT PERSAUD

When treatment is applied to road sections, intersections, drivers, or vehicles that have had a poor accident record in the past, a simple before-and-after comparison of accidents will usually make useless treatments appear effective and overestimate the effect of useful treatments. Because much of what we know about the effect of various safety countermeasures comes from such studies, it is important to demonstrate that this bias can be very large, to show that it can be relatively easily purged from before-and-after comparisons, and to examine whether the method works. These are the three central aims of this paper. To render the statements credible, several data sets are used. The data come from Canada (Ontario), Sweden, the United Kingdom, Israel, and the United States (North Carolina and California) and relate to road sections, intersections, traffic circles, driver violations, and driver accidents. These data sets are used to demonstrate the magnitude of the bias, to illustrate the technique for its elimination, and to examine the success of this debiasing procedure.

It is common practice to apply treatment to those elements of the transport system (intersections, drivers, vehicles) that have had a poor accident record. This makes good sense. The effect of such treatments is often assessed by comparing the accident histories before and after treatment. The cumulative experience gained from several such before-and-after comparisons tends to become the current lore about the effectiveness of the treatment.

We will argue below that when systems are selected for treatment because of their poor accident experience, the simple before-and-after comparison leads to consistently biased conclusions; it makes the treatment appear to be more effective than it actually is. In the long run, the cumulative result of consistent overestimation of treatment effectiveness can lead to misallocation of resources.

The simple before-and-after comparison is and will continue to be an important source of information about the effect of corrective treatments. Therefore, it is important that the existence and size of the bias be recognized and that it be eliminated from the results of before-and-after comparisons.

Accordingly, this paper is divided into three principal parts. The first will show that the bias by selection is not a figment of the theorists' imagination. On the contrary, it is an all-pervasive and empirically substantiated phenomenon. We will also show that the size of this bias is often comparable with what one might expect the result of an effective corrective treatment to be. The second part of the paper will present procedures for the removal of the bias from results of before-and-after studies. In the third part we will examine how these procedures perform when applied to real data.

REGRESSION TO THE MEAN

More than a century ago, Sir Francis Galton (1) observed that the offspring of tall parents were, on the average, shorter than their progenitors. This phenomenon became known as the regression to the mean. As this phrase suggests, when a random deviation from the mean occurs (upward or downward), one should expect the next "trait" to be a return (regression) to the mean. One can show other examples of the same phenomenon that are closer to our common experience. We list below the scores from a professional golf tournament. A large number of golf players are let loose on the course. At the end of

the first two days of play, those with the lowest scores are said to have "made the cut" and are allowed to proceed to the final two days of the tournament. (In golf, the score is a count of strokes needed to place a small white ball into a fixed number of holes in the ground. Thus, the smaller the score, the better one played.) The scores for the first and subsequent pairs of days of the 50 golfers who have made the cut are listed below. Thus, one golfer who scored 134 during the first two days scored 149 during the next two days; the two golfers who scored 138 on the first two days had an average score of 144.5 during the next two days, etc.

No. of Golfers in Group	Total Score for Avg Golfer in Group	
	Days 1 and 2	Days 3 and 4
1	134	149.0
2	138	144.5
1	139	143.0
3	140	144.3
7	141	143.4
5	142	145.0
4	143	148.7
10	144	146.1
9	145	147.0
8	146	146.9

It is quite apparent that those who got in a few lucky shots on days 1 and 2 returned to their average game on days 3 and 4.

What are the salient features common to the two examples that link golf and heredity and how are these relevant to before-and-after comparisons in safety?

In both cases, people were selected for subsequent monitoring on the basis of some trait or score that was found to be higher (or lower) than the average score for the population from which the selection is made. When later the same trait or score was observed again, there was a tendency for it to be closer to the population average.

This does not mean that selection procedures that identify groups on the basis of high (or low) scores will not identify groups whose future scores will be higher (or lower) than the average scores of the population from which they are selected. In fact, such selection procedures will often identify groups whose future scores will continue to be higher (or lower) than average. The crucial issue is the fact that the averages of the original selection scores will almost invariably be higher (or lower) than the subsequent scores of the selected groups.

The parallel in safety is simple. Intersections, drivers, vehicles, etc., are often selected for treatment or improvement if they have an unusually high number of accidents or a high accident rate. Due to the same regression-to-the-mean phenomenon, one should expect these systems to have less accidents subsequently even if nothing is done to them. Yet in a simple before-and-after study, such an observed reduction is normally interpreted as an indication that the countermeasure has been effective.

Skeptics might doubt that what holds for golf scores or stature applies also to the occurrence of

Table 1. Regression to the mean: Ontario data.

No. of Sections in Group	No. of Accidents for Avg Section in Group		Change (%)
	First Year	Second Year	
12 859	0	0.404	Increase
4 457	1	0.832	-16.8
1 884	2	1.301	-35.0
791	3	1.841	-38.6
374	4	2.361	-41.0
160	5	3.206	-35.9
95	6	3.695	-38.4
62	7	4.968	-29.0
33	8	4.818	-39.8
14	9	6.930	-23.0
33	>10 ^a	10.39	-22.0

^a Avg = 13.33.

accidents. To convince them that the phenomenon is real and important, we will use accident data taken from several countries and that describe a variety of circumstances.

System of Road Sections: Ontario, Canada

The King's Highways (excluding freeways) are divided into 20 762 sections 1 km in length. As shown in Table 1, 12 859 of these had no nonintersection accidents in the first year of a two-year period, 4457 had one such accident, etc. The average number of accidents per kilometer during the first year was 0.707 and during the second year, 0.746.

On many of the 12 859 road sections that had no accidents in the first year (first line in Table 1), some accidents occurred during the second year. That this should be so is in accord with intuition. It will facilitate the understanding of the regression-to-the-mean phenomenon if we attempt to make the basis of this intuitive understanding more explicit.

Every road section that is open to traffic will register an accident from time to time. Thus, the long-term mean number of accidents per year for any road section, however small, is always larger than zero. A year in which no accidents occurred on a road section is therefore an event that by chance is below the long-term mean value. During the next year there could again be zero accidents. However, there could also be 1, 2, 3, ... accidents. The average number of accidents in the long term is therefore larger than zero.

This is the essence of the regression to the mean. When a random down fluctuation occurs, one should expect a return to the mean; when a random up fluctuation occurs, one should also expect a return to the mean.

Inspection of Table 1 reveals that this is in fact what happens. Road sections with no accidents in the first year have, on the average, 0.404 accident in the second year. Thus, 0.404 is an estimate of the long-term mean number of accidents for this group of road sections. Similarly, the 4457 road sections that all recorded one accident in the first year regress to the mean and have, on the average, 0.832 accident in the second year. Road sections with two accidents in the first year regress to 1.301 accidents in the second, etc. Naturally, road sections that had no accidents in the first year have a lower long-term mean (0.404) than road sections with, say, eight accidents in the first year (4.818).

The last column of Table 1 shows that the phenomenon is consistent, real, and nothing short of dramatic. To be sure, there are several confounding

factors at play. Treatment in one form or the other may have occurred, road and environmental conditions may have been better during the second year, law enforcement may have improved. However, none of these potentially confounding factors can begin to explain a substantial part of the observed convergence on the mean.

Imagine now what would have happened had the Ministry hired a shaman to pray for accident reduction on road sections that had in the first year seven or more accidents. The apparent effectiveness of this "treatment" is, from the figures in Table 1, about 30 percent.

In summary, road sections that in one period had more than the average number of accidents will usually have less accidents in the next period even if no corrective treatment is applied. This is the essence of our concern. If one is to have a realistic assessment of the effect of some corrective treatment on the basis of a before-and-after comparison, the bias due to the regression to the mean has to be accounted for.

It is natural to ask at this point whether this bias can be substantially reduced (or even rendered negligible) if systems are selected for treatment on the basis of accident data that have been accumulated over a long period of time. The hope is that such data closely represent the mean rather than a random up or down fluctuation. The empirical evidence can be examined on the basis of the following data set.

System of Road Sections: Israel

A seven-year history of nonintersection accidents was obtained for each of 828 km of rural roads in Israel. This data set enabled us to examine the relationship between the magnitude of the regression to the mean and the duration of accident history. To do so, we regard the last year (year 7) as the "after" period; years 6, 5+6, 4+5+6, etc., are considered "before" periods, which are 1, 2, 3, ... years in duration.

Table 2 summarizes the empirical evidence. Several interesting observations follow.

First, the regression to the mean for road sections does not recognize national boundaries. It is as evident for Israeli as for Canadian road sections. Thus, prayers for accident reduction on road sections that in one year had three or more accidents (Table 2, bottom) might seem to bring about a 55 percent reduction in the number of accidents.

Second, as the duration of the before period increases (column 1 of Table 2), the relative size of the regression to the mean diminishes (column 5). This is as anticipated. Note, however, that even with a before history of six years, the size of the phenomenon is far from negligible.

At the peril of boring the converted and belaboring unnecessarily a point that is already clear, we will present further empirical evidence. The intent is to demonstrate by real cases that what holds for human stature, golf scores, and road sections also holds for traffic circles, intersections, driver accidents, and traffic law violations. We hope that the breadth of geographical coverage will show universality and that the variety of circumstances will underscore the all-pervasive nature of this phenomenon.

The first two accident-related data sets dealt with road sections. The next two illustrations focus on intersections.

Approaches to Traffic Circles: United Kingdom

Transverse yellow bar markings with gradually dimin-

Table 2. Effect of duration of before period: Israeli data.

Duration of Before Period (years)	No. of Road Sections	No. of Accidents per Year		Change (%)
		Before	After	
Road Sections with One or More Accidents per Year				
1	337	530	317	-40
2	258	393	277	-30
3	231	321	250	-28
4	191	292	230	-21
5	178	272	224	-18
6	170	258	222	-14
Road Sections with Two or More Accidents per Year				
1	126	319	164	-49
2	64	159	91	-43
3	47	111	78	-30
4	45	107	75	-30
5	37	92	70	-24
6	28	72	55	-24
Road Sections with Three or More Accidents per Year				
1	45	157	71	-55
2	16	53	31	-41
3	5	17	15	-12
4	4	15	10	-33
5	7	25	21	-16
6	7	24	19	-21

Table 3. Regression to the mean: U.K. roundabout data.

No. of Accidents per Site in Year 1	No. of Sites in Group	Aggregate Accidents for Each Group		Change (%)
		Year 1	Year 2	
10+	5	74	54	-37
9	2	18	10	-44
8	7	56	49	-13
7	2	14	6	-57
6	9	54	48	-15
5	9	45	35	-22
4	7	28	29	+4
3	11	33	44	+33
2	15	30	46	+53
1	9	9	13	+44
0	6	0	7	Increase
	82	361	341	

ishing spacing are known to reduce speed. Results of a study on the effect of such markings on accidents at approaches to traffic circles (roundabouts) have been published (2). These will be used later to show that when before-and-after comparisons are cleansed of the bias by selection, proper conclusions can be reached even when it is not practical to include corresponding control sites in the study. At present, however, we will use the data only to show that had the researchers not taken proper precautions, the regression to the mean would have destroyed the validity of their results.

For the 82 roundabout approaches included in the study, the number of injury accidents during each of two years prior to treatment is shown in Table 3.

Since there was no treatment or any other important change between the two years, the observed effect is largely due to the regression to the mean. The mean annual number of accidents per approach is 4.5. Sites that had more than four accidents in year 1 tend to experience a reduction in year 2; approaches with less than the mean in year 1 had an increase.

The danger is obvious. Had the researchers decided to treat with yellow bar markings approaches

Table 4. Regression to the mean: reported and injury accidents at intersections in Sweden.

No. of Intersections in Group	No. of Accidents per Intersection During Before Period	Avg No. of Accidents per Intersection During Equivalent After Period	Change (%)
All Reported Accidents			
1500	0 (0)	0.37	Increase
657	1 (0.85)	0.77	(-9)
244	2 (1.70)	1.04	(-39)
101	3 (2.56)	1.79	(-30)
53	4 (3.41)	1.94	(-43)
39	5 (4.26)	2.69	(-37)
17	6 (4.11)	3.05	(-40)
26	8.92 (7.60) ^a	5.46	(-28)
Injury Accidents			
2039	0 (0)	0.19	Increase
441	1 (0.85)	0.42	(-51)
119	2 (1.70)	0.71	(-59)
24	3 (2.56)	1.33	(-48)
14	4.143 (3.53) ^b	1.50	(-57)

Note: Figures in parentheses are exposure adjusted.

^aThese figures are for sites with seven or more accidents in 1972-1975.

^bThese figures are for sites with four or more accidents during 1972-1975.

that have many accidents (more than four per year), they would have observed a reduction even if the treatment was useless. With a useful treatment, the observed reduction would be an overestimate of treatment effectiveness.

The consistency with which the phenomenon appears in such a limited data set is quite remarkable. A more extensive study is examined below.

System of Intersections: Sweden

This illustration is based on 2637 unsignalized rural road junctions that were unaltered during the period 1972-1978 (3). The four-year stretch 1972-1975 is regarded as the before period and the three years 1976-1978 as the after period. The top part of Table 4 is based on all police-reported accidents, whereas the bottom half is based only on personal-injury accidents.

In Table 4, the number of before-period accidents has been adjusted to reflect differences in length of the before and after periods and in exposure. Unlike the road sections examined earlier, only intersections that remained physically unaltered are included in this data set. Therefore, the change in the number of accidents between the before and after periods is attributable almost exclusively to the regression to the mean.

It is interesting to note that the change in injury accidents is larger than that for all reported accidents. The exact reason will become clear later. However, the main element of the explanation is already evident. It is the difference in the mean towards which the number of accidents regresses. Since there are less injury accidents than all reported accidents, the size of the regression to the mean for injury accidents is larger than that for all reported accidents.

The four data sets examined so far dealt with accidents occurring on elements of the road system. The regression-to-the-mean effect that has been shown to exist in these illustrations must be taken into account when safety benefits of corrective treatments to road sections and intersections are estimated.

Corrective treatment is often aimed at the road user (rather than at the road system). Thus, for example, those convicted repeatedly for impaired

driving may be sent for a week-long course, those who accumulated a certain number of demerit points may receive a warning letter, etc. As before, when the selection for treatment is on the basis of a bad record, the regression to the mean may be at work. Accordingly, one should expect an improvement in the record even when no treatment is administered. That this is in fact the case will be demonstrated in the following examples.

System of Drivers: United States (North Carolina)

In an investigation about the predictability of accidents and violations on the basis of past performance, the records of some 2.5 million drivers in North Carolina were examined (4).

In the tabulation below, drivers are placed in seven groups according to the number of traffic law violations during the first two-year period. For each group, the average number of violations in the subsequent two-year period is also given. The existence and size of the regression-to-the-mean effect are evident from the examination of the last column. (The total number of drivers in the group included 674 drivers with seven or more violations in the first period.)

No. of Drivers in Group	No. of Violations for Avg Driver in Group		
	First	Second	Change (%)
	Two-Year Period	Two-Year Period	
2 096 935	0	0.191	Increase
298 645	1	0.457	-54.3
73 216	2	0.763	-61.9
21 907	3	1.024	-65.9
7 224	4	1.266	-68.4
2 597	5	1.459	-70.8
1 042	6	1.500	-75.0
2 502 240	0.225	0.252	+12.0

As in earlier illustrations, the regression to the mean is not the only cause of the noted change. First, the average number of violations per driver has changed somewhat (from 0.225 to 0.252). Second, many violations are linked to accidents the occurrence of which may have a substantive effect on one's driving. Third, the level of enforcement, driver maturation, possible corrective treatments, etc., all may exert some influence. Nevertheless, as will be demonstrated later, the main reason for the drop from, say, 6 to 1.5 is the fact that the group includes drivers who, on the average, have 1.5 violations per year and who due to the laws of chance happened to record six in the first two years. Similarly, the estimated long-term mean number of violations in two years for the group of drivers who during the first two-year period had no violations is 0.191. The regression is always to the long-term mean specific to the selected group.

In the table below, the same North Carolina drivers are placed into seven groups according to the number of accidents they had in the first two-year period. For each group, the average number of accidents in the subsequent two-year period is listed. The Change column gives an indication of the regression to the mean. Some of this change may be due to reduced exposure and change in driving style, which is linked to the accident trauma experienced in the first period. How much of the change is attributable to such causes and what part is due to regression to the mean will be explored later. (The total number of drivers included 10 with seven or more accidents in the first period.)

No. of Drivers in Group	No. of Accidents for Avg Driver in Group		Change (%)
	First	Second	
	Two-Year Period	Two-Year Period	
2 234 577	0	0.117	Increase
235 080	1	0.216	-78.4
27 919	2	0.348	-82.6
3 953	3	0.499	-83.4
584	4	0.703	-82.4
99	5	0.848	-83.0
18	6	0.944	-84.3
2 502 240	0.122	0.130	+6.1

These tabulations lead to the following observations:

First, drivers with one or more violations or accidents in the first period are seen to have, on the average, 60-80 percent less violations or accidents in the second period, in spite of the increase in the number of violations and accidents from the first to the second period. Note that the magnitude of the change is much larger for drivers than that for the elements of the road system.

Second, as observed earlier in the case of the intersections in Sweden, the smaller the mean towards which observations regress, the larger the change.

System of Drivers: United States (California)

This last illustration endeavors to reinforce two points made earlier: (a) regression to the mean is not bound by geography--accident records of drivers in California show the same effect as the records of North Carolina drivers; (b) the magnitude of the effect remains large even when the before period is three years long.

In the table below, some 93 000 randomly selected drivers who have driven in California for at least six years have been placed into four groups according to the number of accidents they had during the three-year period 1969-1971. (Data were obtained from R. Peck, Department of Motor Vehicles, State of California, March 1982.) For each group, the average number of accidents in the subsequent three-year period (1972-1974) is listed. The Change column gives an indication of the regression to the mean. (The total number of accidents for 1969-1971 was not reported for the last group of drivers.)

No. of Drivers in Group	No. of Accidents for Avg Driver in Group		Change (%)
	1969-1971	1972-1974	
79 327	0	0.152	Increase
11 897	1	0.238	-76.2
1 525	2	0.374	-81.3
250	3+	0.548	--

Entries in the previous two tables are remarkably similar in spite of the difference in geography, driver population, and time frame. One is again led to conclude that the phenomenon is universal. Even though three years of data are used for the before period in the California study, the effect remains very large.

Having shown the existence of the regression to the mean in a diverse set of circumstances, we will regard it as empirically substantiated. From here, we take it as "common ground" that if a system (road section, intersection, driver, vehicle, etc.) is observed to have a higher-than-mean number of accidents in one period, it should be expected to have less accidents in the next period if no corrective treatment is applied to it.

It follows that if corrective treatment is ap-

plied to systems that had a poor accident record, it is impossible to estimate the safety benefits of the treatment from its accident record after treatment without first accounting for the effect of the regression to the mean. How to do so is explained next.

OBTAINING UNBIASED ESTIMATES IN UNCONTROLLED BEFORE-AND-AFTER COMPARISONS

To assess the safety effect that some corrective treatment had on a system to which it has been applied, one compares the safety performance of the system after the treatment against what one would expect the safety performance of the system to have been during the same period of time had the treatment not been applied. This is the logical premise on which all attempts to estimate the safety effect of countermeasures are founded, no matter how simple or sophisticated the associated experimental design.

In light of this logical premise, it is important to note that a simple before-and-after comparison is equivalent to stating, "I assume that the safety performance before treatment is a good estimate of what the safety performance would have been during the after period had treatment not been applied."

The first part of this paper has been devoted to the demonstration that this assumption runs counter to abundant empirical evidence and that when systems are selected for corrective treatment because of their poor safety record, one must expect them to have an improved accident record in the subsequent period if no treatment is applied. The central task of this section is to provide a credible answer to the question, What accident record should one expect during the after period if no treatment is administered?

To cast the problem in more precise terms, let b denote the number of accidents occurring on a system during some period before treatment and α denote the number of accidents expected to occur on the same system during an after period of the same duration had treatment not been applied. Our task is to estimate α .

When α is estimated, what evidence should count? A general answer to this simple question is not easy to provide. However, when the system is selected for treatment out of a larger group of candidates, there are at least two pieces of information that must affect the estimate:

1. The safety performance of the selected system during the before period (b) and
2. The safety performance during the before period of all other candidate systems.

The relevance of the first piece of information is self-evident. The second piece of information is relevant because the safety record of the selected system was compared with the safety records of other candidates when the selection was made.

The estimation of α turns out to be astonishingly simple.

Estimation Rule 1

The estimate of the number of accidents expected to occur in the after period on systems that during the before period had k accidents is the number of accidents occurring on systems that had $(k + 1)$ accidents in the before period.

This rule is best illuminated by a simple example. Consider the data in Table 1 and pose the following question: How many accidents should one expect to occur during the second period on the 62 Ontario road sections that in the first period had

seven accidents? [We know, of course, from Table 1 that the actual number of accidents on these road sections during the second period was $308 = (62 \times 4.968)$. But for the moment we imagine that we are just at the end of the first period and have to provide an estimate.]

Following estimation rule 1, we find $\alpha = 8 \times 33 = 264$ accidents. It is tempting to immediately compare this estimate with what actually transpired. Such temptation should be resisted until a systematic juxtaposition of estimates and the actual number of accidents is carried out later in this section. Suffice to note that were one to assume (as is normal practice) that road sections with seven accidents in the before period are likely to have the same number of accidents (seven) in the after period, a gross overestimate would be obtained.

Estimation rule 1 can be recast into an alternative (equivalent) form.

Estimation Rule 2

The estimate of the number of accidents expected to occur in the after period on a system that during the before period had k accidents is the number of accidents occurring during the before period on systems that had $(k + 1)$ accidents divided by the number of systems on which k accidents occurred during the before period.

To continue the previous illustrative example, an Ontario road section that during the first period had seven accidents should be expected to have $(8 \times 33)/62 = 4.3$ accidents during the second period.

It should be noted that when the two periods differ in duration or in exposure, a correction should be applied to the estimate of α in the customary manner.

In some cases it is convenient to use a third variant of estimation rule 1.

Estimation Rule 3

The estimate of the number of accidents expected to occur during the after period on systems that during the before period had k or more accidents is the number of accidents occurring on systems that had $(k + 1)$ or more accidents during the before period.

To return to the data in Table 1, Ontario road sections that during the first period had seven or more accidents should be expected to have during the second period $830 = (8 \times 33 + 9 \times 14 + 13.33 \times 33)$ accidents. As can be easily ascertained, the actual number of accidents on these road sections was 907.

The above estimation rules have been obtained in an entirely rigorous way, as shown in the next section. The proof rests on the universally accepted but empirically unproven assumption that accident occurrence on each system obeys the Poisson probability law. Thus, while each road section or driver has its own characteristic mean (number of accidents per unit time), the probability of a specific realization is specified by the Poisson distribution.

Since the proof of the pudding is in the eating, the last section is a patient juxtaposition of unbiased estimates and the empirical evidence contained in the data sets already introduced in the first part of this paper.

Determination of Estimation Rules

Out of n systems, those that during time period 1 recorded k or more accidents are chosen. We wish to find the number of accidents one should expect to occur on the chosen systems during period 2 of the same duration. Let $B(k)$ be the expected number of

accidents on the chosen systems in period 1 and $A(k)$ be the expected number of accidents on the chosen systems in period 2.

Let m_i , $i = 1, 2, \dots, n$, be the expected number of accidents on candidate system i during periods 1 and 2 and let p_{ij} denote the probability that system i will have j accidents. With this notation,

$$P(\text{system } i \text{ is chosen}) = \sum_{j=k}^{\infty} p_{ij} \quad (1)$$

and the contribution of system i to $B(k)$ is

$$\sum_{j=k}^{\infty} j p_{ij}$$

If we sum over all systems,

$$B(k) = \sum_{i=1}^n \sum_{j=k}^{\infty} j p_{ij} \quad (2)$$

If we assume that the number of accidents on a system obeys the Poisson probability law,

$$p_{ij} = \exp(-m_i) m_i^j / j!$$

for system i and

$$j p_{ij} = m_i p_{i,j-1} \quad (3)$$

If we substitute Equation 3 into Equation 2,

$$\begin{aligned} B(k) &= \sum_{i=1}^n m_i \sum_{j=k}^{\infty} p_{i,j-1} \\ &= \sum_{i=1}^n m_i \sum_{j=k-1}^{\infty} p_{ij} \end{aligned} \quad (4)$$

On the other hand, the expected number of accidents on system i during period 2 is m_i . By using Equation 1,

$$A(k) = \sum_{i=1}^n m_i \sum_{j=k}^{\infty} p_{ij} \quad (5)$$

If we compare Equations 4 and 5,

$$A(k) = B(k+1) \quad (6)$$

Equation 6 leads directly to estimation rule 3.

Note that to obtain Equation 6, one assumes only that accident occurrence is governed by the Poisson probability law. The mean number of accidents (m_i) differs from system to system.

By using the result in Equation 6, we can write

$$A(k) - A(k+1) = B(k+1) - B(k+2) \quad (7)$$

The expression on the left-hand side is the expected value of the difference (number of accidents in period 2 on systems that in period 1 had k or more accidents minus number of accidents in period 2 on systems that in period 1 had $k+1$ or more accidents). This is, of course, the expected number of accidents in period 2 on systems that in period 1 had k accidents and will be denoted by $\alpha(k)$.

Similarly, the expression on the right-hand side in Equation 7 is the expected number of accidents occurring during period 1 on systems with $k+1$ accidents and will be denoted by $\beta(k+1)$. Thus,

$$\alpha(k) = \beta(k+1) \quad (8)$$

Equation 8 leads directly to estimation rule 1.

EMPIRICAL EXAMINATION OF WHETHER ESTIMATION RULES YIELD UNBIASED ESTIMATES

The main content of this section is an examination of the estimated number of after-period accidents (or violations) compared with the number of after-period accidents (violations) actually recorded. The hope is that they are in good agreement.

Were one to plot estimates along one axis of a rectangular coordinate grid and the recorded number of accidents on the other axis, an ideal agreement would place all such points on the diagonal. There are at least three reasons why such an ideal relationship should not be expected.

First, as is evident from reading the estimation rules, what plays the role of an estimate is a (Poisson-distributed) random variable. As such it is, of course, subject to considerable variability. Second, the recorded number of accidents to which the estimate is compared is similarly subject to large random variation. Thus, even if the expected location of the point was on the diagonal, the random variation in its ordinate and abscissa will cause it to deviate from it. Third, as has been pointed out several times earlier, the before period can never be regarded as equivalent to the after period even if precautions are taken to consider only untreated systems and correction for exposure is applied. This source of uncertainty will cause even the (unknown) expected location of the points to be off the diagonal. Thus, a realistic assessment of what constitutes good agreement is that the points are fairly close to the diagonal and a regression line through the points has a slope close to unity.

Since the estimation rules are essentially equivalent, it would be redundant to apply all three to each data set. We begin by applying all three rules to the Ontario data and discuss the application in some detail. This will allow us to deal with all other data sets more summarily. Finally, results of a controlled evaluation are compared with what is obtainable from an uncontrolled study from which the bias is eliminated.

Application to Ontario Road Sections

The numbers in Table 5 summarize the application of all three estimation rules to the data on Ontario road sections that were introduced earlier and described in Table 1. The left part of Table 5 deals with the application of rules 1 and 2, the right part with estimation rule 3.

Column 2 gives the number of road sections that during the before period had $k = 0, 1, 2, \dots$ accidents. These values are copied from Table 1. Column 3 gives the total number of accidents occurring on these sections during the before period. Thus, for example, on the 374 sections that during the before period had 4 accidents, the total number of accidents was $374 \times 4 = 1496$.

By estimation rule 1, one should expect that during the after period, the total number of accidents on road sections that during the before period had 3 ($= 4 - 1$) accidents will be 1496. This is the entry in row 4 of column 5. Note that the estimate in column 5 is simply the entry in column 3 lifted by one row. The number of accidents actually recorded during the after period is given in column 7. It is the comparison of the entries in columns 5 and 7 by which the performance of estimation rule 1 is to be judged. Thus, while on the basis of the accidents in the before period one would have estimated that 1496 accidents will occur on the 791 road sections, the number of accidents actually recorded was 1456. Were one to follow the common practice and assume

Table 5. Application of estimation rules to Ontario data.

Estimation Rules 1 and 2 ^a								Estimation Rule 3 ^b			
No. of Accidents (k)	No. of Sections with Exactly k Accidents	Total No. of Accidents						No. of Sections with k or More Accidents During Before Period	Aggregate No. of Accidents at Identified Sites		
		Before Period (first year)		After Period (second year)					One Year After		
				Estimated		Recorded					
		Rule 1	Rule 2	Rule 1	Rule 2	Rule 1	Rule 2	One Year Before	Estimated	Recorded	
0	12 859	0	0	4457	0.354	5199	0.404	20 762	14 648	14 648	15 467
1	4 457	4457	1	3768	0.845	3706	0.832	7 903	14 648	10 191	10 268
2	1 884	3768	2	2373	1.260	2452	1.301	3 446	10 191	6 503	6 562
3	791	2373	3	1496	1.891	1456	1.841	1 562	6 503	4 130	4 110
4	374	1496	4	800	2.139	883	2.361	771	4 130	2 634	2 654
5	160	800	5	570	3.563	513	3.206	397	2 634	1 834	1 771
6	95	570	6	434	4.568	351	3.695	237	1 834	1 264	1 258
7	62	434	7	264	4.258	308	4.968	142	1 264	830	907
8	33	264	8	126	3.818	159	4.818	80	830	566	599
9	14	126	9	80	5.714	97	6.930	47	566	440	440
10	8	80	10			74		33	440	360	343
≥11	25	360				269		25	360		

^aFor sites with exactly k accidents before.^bFor sites with k or more accidents before.

that what happened before is an indication of what should be expected after, one would expect 2373 accidents to occur on these road sections, an obvious overestimate. Furthermore, were some treatment applied to such sections, one might erroneously conclude that the apparent reduction from 2373 to 1456 can be attributed to the treatment when in fact it is but an artifact of the regression to the mean.

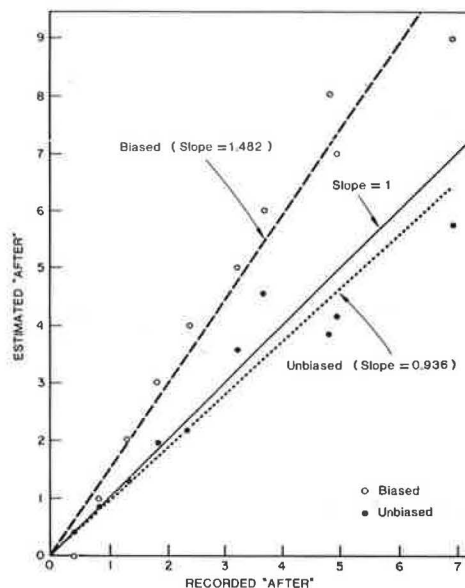
Columns 4, 6, and 8 give the entries relating to the application of estimation rule 2. Thus, column 4 gives the number of accidents during the before period for an average road section in the group. Entries in column 6 are the estimates of the number of accidents expected to occur on such a road section during the after period. To apply estimation rule 2, one has to divide the entry in column 5 by the entry in column 2. For example, a section for which $k = 3$ should be expected to have, during the after period, $1496/791 = 1.891$ accidents. The performance of estimation rule 2 is to be judged by the comparison of columns 6 and 8. Note that entries of column 8 are taken from Table 1.

It is easy to visualize the performance of the estimation rules with the aid of the graphical representation in Figure 1. Each pair of numbers from columns 6 and 8 is represented by a solid circle. Perfect agreement between the unbiased estimate and the recorded number of accidents would place the solid circle on the diagonal. It is easy to see that the solid circles do hug the diagonal.

The performance of the biased estimates is represented by open circles based on the entries of columns 4 and 8. The vertical distance between the open circle and the diagonal indicates the magnitude of the bias due to the regression to the mean. A line through the origin has been fitted to each set of points and its slope is indicated in the figure.

Estimation rules 1 and 2 apply to sections with k accidents; rule 3 applies to systems with k or more accidents. Accordingly, one has to first put the data from columns 2 and 3 into the cumulative form shown in columns 9 and 10. Starting from the bottom of column 2, $25 + 8 = 33$ is the number of road sections with 10 or more accidents during the before period as listed in column 9. Similarly, $360 + 80 = 440$ is the number of accidents occurring on these road sections during the before period and listed in column 10. By estimation rule 3, one should expect that in the after period the number of accidents on road sections that during the before period had, say, nine or more accidents, will be the same as the

Figure 1. Application of rule 2 to Ontario data.



number of before-period accidents on road sections with 10 or more accidents, i.e., 440. This is the entry in column 11. As before, to obtain estimates, raise the entries in column 10 by one row. The performance of estimation rule 3 is judged on the basis of the comparison of the estimate in column 11 and the recorded number of after-period accidents in column 12. The entries of column 12 are obtained by the cumulation from below of the numbers in column 7. The performance of rule 3 in the case of Ontario road sections is shown in Figure 2.

Application to Other Data Sets

Estimation rules 2 and 3 have been applied to all data sets introduced earlier. We forego here the presentation of detailed comparisons as in Table 5 and describe the performance of biased and unbiased estimates in graphical form only. Figures 3-12 relate the number of accidents recorded during the after period to what one would expect on the basis of the number of accidents occurring during the be-

fore period. Solid circles depict the correspondence obtained by using the unbiased estimates and open circles the correspondence of the biased estimates. The slope of the best-fitting line to the solid circles and to the open circles is shown on each rule-2 graph. Figures 10 and 12 are plotted on

a log-log scale because of the range of numbers.

Some important observations can be made on the basis of these figures:

1. As is indicated by the slopes of the best-fitting lines, estimates obtained by using the recommended rules appear to be largely free of bias. Thus, in spite of the unavoidable limitation of comparing the safety of a system in two different periods of time, the theoretical considerations that led to the formulation of estimation rules seem to be supported by empirical data.

2. When estimates are based on a small number of accidents, they are unreliable, as should be expected. To illustrate, consider, for example, Figure 5, which is based on data in Table 3. Since each point is based only on a few accidents, the difference between the biased and unbiased estimates is entirely obscured by the random variations. However, when accidents are accumulated as for the application of estimation rule 3 (Figure 6), the same data clearly show the superiority of unbiased estimation.

Driver violations and accidents are overestimated by the unbiased method as well. This is not surprising due to the maturation of the driver population. Drivers who have had accidents or violations during the before period might be expected to have fewer accidents or violations subsequently, not only because they are more careful, but possibly because

Figure 2. Application of rule 3 to Ontario data.

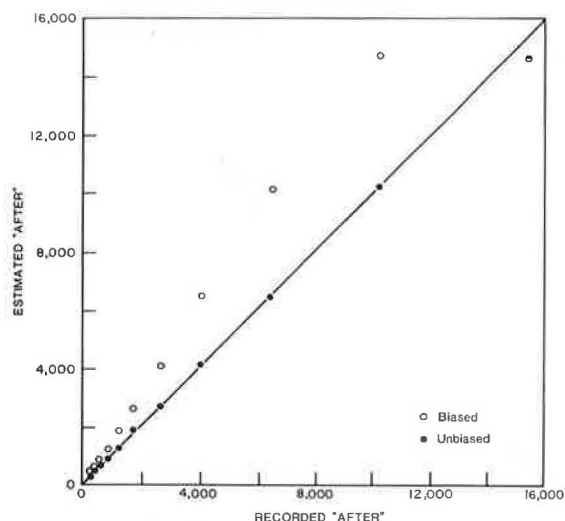


Figure 3. Application of rule 2 to Israeli data.

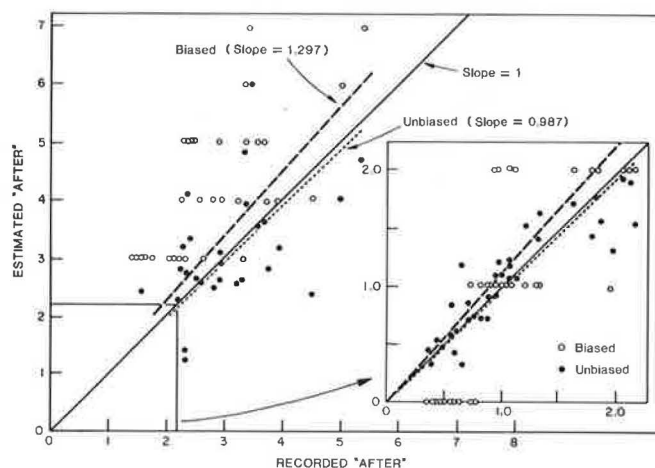
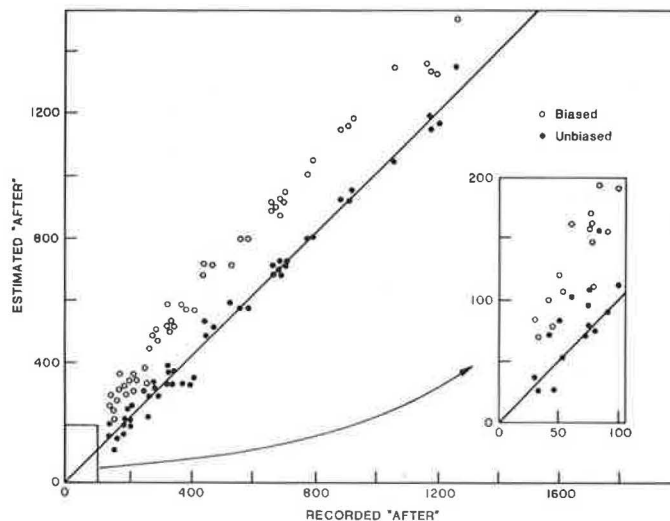


Figure 4. Application of rule 3 to Israeli data.



they tend to drive less due to hospitalization, license revocation, and similar reasons.

Comparing Results of Controlled Study to Study Without Control

The best way to avoid problems due to regression to

the mean is to match to each system treated another system with the same accident experience that is left untreated and draw conclusions from their performance during the after period. However, when this is not practical, the suggested estimation rules should be applied.

The study on the safety effect of yellow bar

Figure 5. Application of rule 2 to U.K. roundabout data.

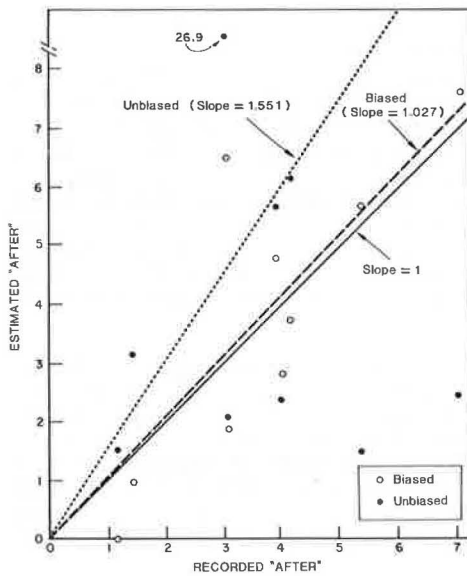


Figure 6. Application of rule 3 to U.K. roundabout data.

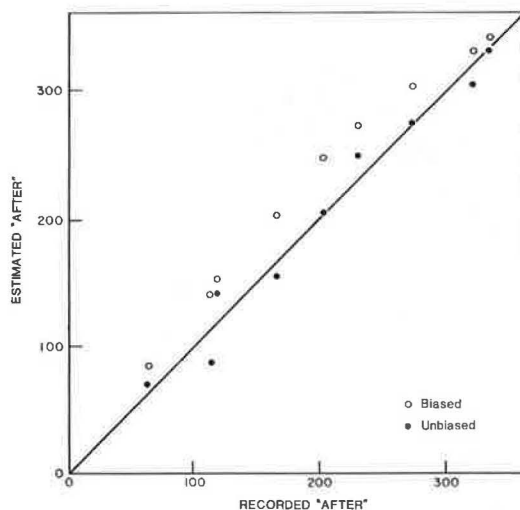


Figure 7. Application of rule 2 to Swedish intersection data.

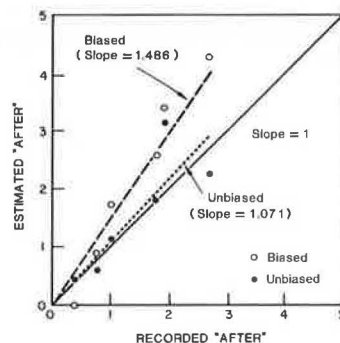


Figure 8. Application of rule 3 to Swedish intersection data.

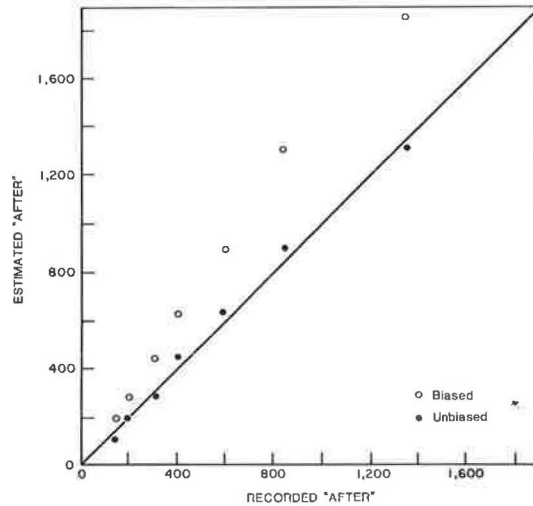


Figure 9. Application of rule 2 to North Carolina driver violations.

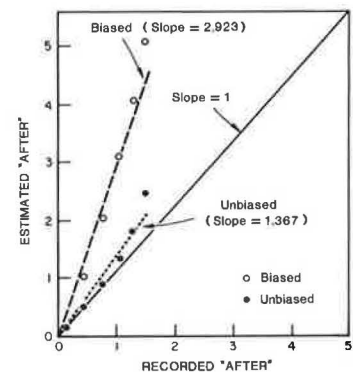


Figure 10. Application of rule 3 to North Carolina driver violations.

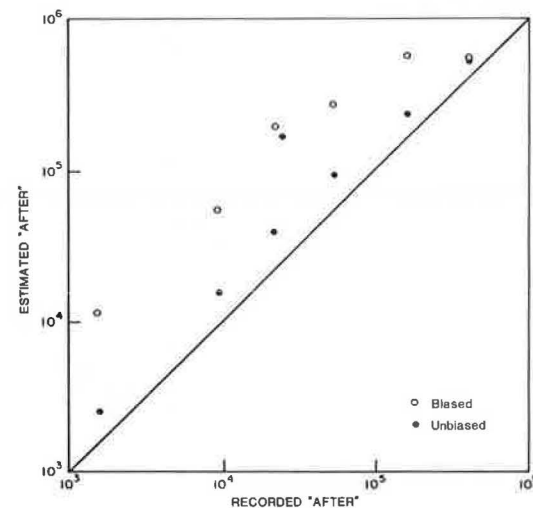


Figure 11. Application of rule 2 to driver-accident data.

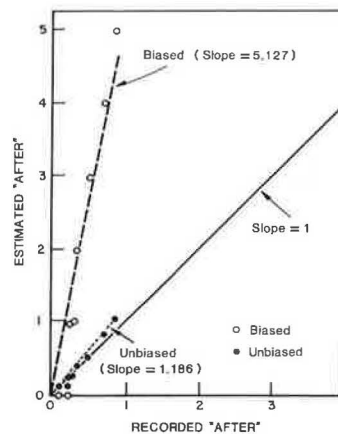
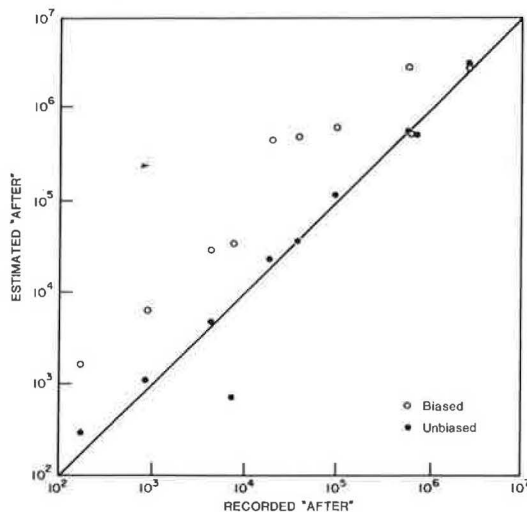


Figure 12. Application of rule 3 to driver-accident data.



markings was of the controlled type, and therefore the reported 57 percent reduction in susceptible accidents is free from bias. It is interesting to examine what the estimate of effectiveness obtained by the use of estimation rule 3 would be had the researchers not enjoyed the luxury of being able to identify appropriate control sites.

By using only data about treated sites and applying estimation rule 3, we obtain the results shown below:

No. of Accidents in Two-Year Period Before Treatment	No. of Sites in Group	Apparent Treatment Effective- ness (%)	Unbiased Estimate of Treat- ment Ef- fective- ness (%)
>4	7	74	63
>3	15	76	63
>2	27	66	52
>1	36	64	60

Column 3 is based on the simple (biased) before-and-after comparison; column 4 gives the unbiased estimate of treatment effectiveness. Two points deserve mentioning. First, as has been argued earlier, the apparent treatment effectiveness is an overestimate. Second, the unbiased estimates are quite consistent with the finding of 57 percent re-

duction based on the more elaborate controlled experiment.

SUMMARY AND DISCUSSION

We have shown that when systems are selected for remedial treatment because of their poor accident history, simple before-and-after comparisons tend to overestimate the safety effectiveness of the treatment due to the regression-to-the-mean phenomenon. Several real-world data sets have been used to show that this phenomenon is real and important. A method for purging the bias from simple before-and-after comparisons has been suggested and summarized by three equivalent estimation rules. The performance of these rules for the data sets introduced earlier has been examined.

The estimation rules are based on analytical considerations described fully elsewhere (5). The same results were published earlier by Robbins (6).

Within the limits of accuracy influenced by the random variations of the data and the fact that the before conditions are never identical to the after conditions, the validity of the estimation rules is supported by the findings. It should be noted that when estimation is based on a small number of accidents, the accuracy is correspondingly limited.

To improve estimation accuracy when conclusions must be based on the comparison of a relatively small number of accidents, an empirical Bayes approach has been suggested (7,8). The performance of the two alternative methods of estimation has been examined in a different context (9). There it has been concluded that for large sample sizes and under certain other conditions, the estimation-rule approach is preferable.

A similar examination in a safety context has not yet been performed.

ACKNOWLEDGMENT

This publication was supported by the Insurance Institute for Highway Safety. The opinions, findings, and conclusions expressed are ours and do not necessarily reflect the views of the Institute.

REFERENCES

1. Sir Francis Galton. Typical Laws of Heredity. Proceedings of the Royal Institute, Vol. 8, Feb. 9, 1877, pp. 282-301.
2. R.D. Helliard-Symons. Yellow Bar Experimental Carriageway Markings: Accident Study. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, TRRL Rept. 1010, 1981.
3. U. Br de and J. Larsson. Regression-to-the-Mean Effect: Some Empirical Examples Concerning Accidents at Road Junctions. Proc., OECD Seminar on Short-Term and Area-Wide Evaluation of Safety Measures, Organization for Economic Cooperation and Development, Netherlands, April 1982.
4. E. Hauer. Bias-by-Selection: Overestimation of the Effectiveness of Safety Countermeasures Caused by the Process of Selection for Treatment. Accident Analysis and Prevention, Vol. 12, No. 2, June 1980.
5. H. Robbins. Prediction and Estimation for the Compound Poisson Distribution. Proceedings of the National Academy of Sciences, Vol. 74, No. 7, July 1977, pp. 2670-2671.
6. D.F. Jarrett, C. Abbess, and C.C. Wright. Bayesian Methods Applied to Road Accident Black-spot Studies: Some Recent Progress. Proc., OECD Seminar on Short-Term and Area-Wide Evaluation of Safety Measures, Organization for Eco-

conomic Cooperation and Development, Netherlands, April 1982.

7. C. Abbess, D. Jarrett, and C.C. Wright. Accidents at Blackspots: Estimating the Effectiveness of Remedial Treatment, with Special Reference to the "Regression-to-the-Mean Effect". Traffic Engineering and Control, Vol. 22, No. 10, 1981, pp. 535-542.

8. D.G. Morrison and D.C. Schmittlein. Predicting Future Random Events Based on Past Performance, Management Science, Vol. 27, No. 9, Sept. 1981.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.