

---

---

---

# 1194

TRANSPORTATION RESEARCH RECORD

---

## *Traffic Flow Theory and Highway Capacity*

---

TRANSPORTATION RESEARCH BOARD  
NATIONAL RESEARCH COUNCIL  
WASHINGTON, D.C. 1988

**Transportation Research Record 1194**

Price: \$28.00

mode

1 highway transportation

subject area

55 traffic flow, capacity, and measurements

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

Printed in the United States of America

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

Traffic flow theory and highway capacity.

p. cm.—(Transportation research record, ISSN 0361-1981 ; 1194)

ISBN 0-309-04765-X

1. Traffic flow—Congresses 2. Highway capacity—  
Congresses. I. National Research Council  
(U.S.). Transportation Research Board.

II. Series.

TE7.H5 no. 1194

[HE336.T7]

388 s—dc20

[388.3'14]

89-14570

CIP

**Sponsorship of Transportation Research Record 1194**

**GROUP 3—OPERATION, SAFETY, AND MAINTENANCE OF  
TRANSPORTATION FACILITIES**

*Chairman: James I. Taylor, University of Notre Dame*

**Facilities and Operations Section**

**Committee on Highway Capacity and Quality of Service**

*Chairman: Carlton C. Robinson, Highway Users Federation*  
*Secretary: Charles W. Dale, Federal Highway Administration*  
*Robert C. Blumenthal, Kenneth G. Courage, Thomas C. Ferrara,*  
*P. Kay Griffin, Wayne E. Hausler, V. F. Hurdle, Paul P. Jovanis,*  
*Wayne K. Kittelson, Joel P. Leisch, Herbert S. Levinson, Adolf D.*  
*May, Jr., Carroll J. Messer, Ronald C. Pfefer, William R. Reilly,*  
*Roger P. Roess, Nagui M. Roupail, Ronald C. Sonntag, Stan*  
*Tepley, Mark R. Virkler, Alexander Werner, Robert H. Wortman,*  
*John D. Zegger*

**Committee on Traffic Flow Theory and Characteristics**

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

David K. Witheford, Transportation Research Board staff

Sponsorship is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 1987.

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

# Transportation Research Record 1194

---

## Contents

<b>Volume-Based Model for Forecasting Truck Lane Use on the Rural Interstate</b> <i>D. Albright and C. Blewett</i>	1
<b>Comparative Analysis of Two Logics for Adaptive Control of Isolated Intersections</b> <i>Feng-Bor Lin</i>	6
<b>New Algorithm for Solving the Maximum Progression Bandwidth</b> <i>Huel-Sheng Tsay and Liang-Tay Lin</i> DISCUSSION, <i>Edmond C.-P. Chang</i> , 27 AUTHORS' CLOSURE, 28	15
<b>Examination of Shared Lane Operations</b> <i>J. A. Bonneson, C. J. Messer, and D. B. Fambro</i>	31
<b>Estimation of Independence of Vehicle Arrivals at Signalized Intersections: A Modelling Methodology</b> <i>Gang-Len Chang and James C. Williams</i>	42
<b>WEAVSIM: A Microscopic Simulation Model of Freeway Weaving Sections</b> <i>M. Zarean and Z. A. Nemeth</i>	48
<b>Effect of Weather on the Relationship Between Flow and Occupancy on Freeways</b> <i>Fred L. Hall and Deanna Barrow</i>	55
<b>Analysis of Platoon Dispersion with Respect to Traffic Volume</b> <i>Karsten G. Baass and Serge Lefebvre</i>	64
<b>Analysis of Traffic Flow at Complex Congested Arterials</b> <i>Panos G. Michalopoulos</i>	77

---

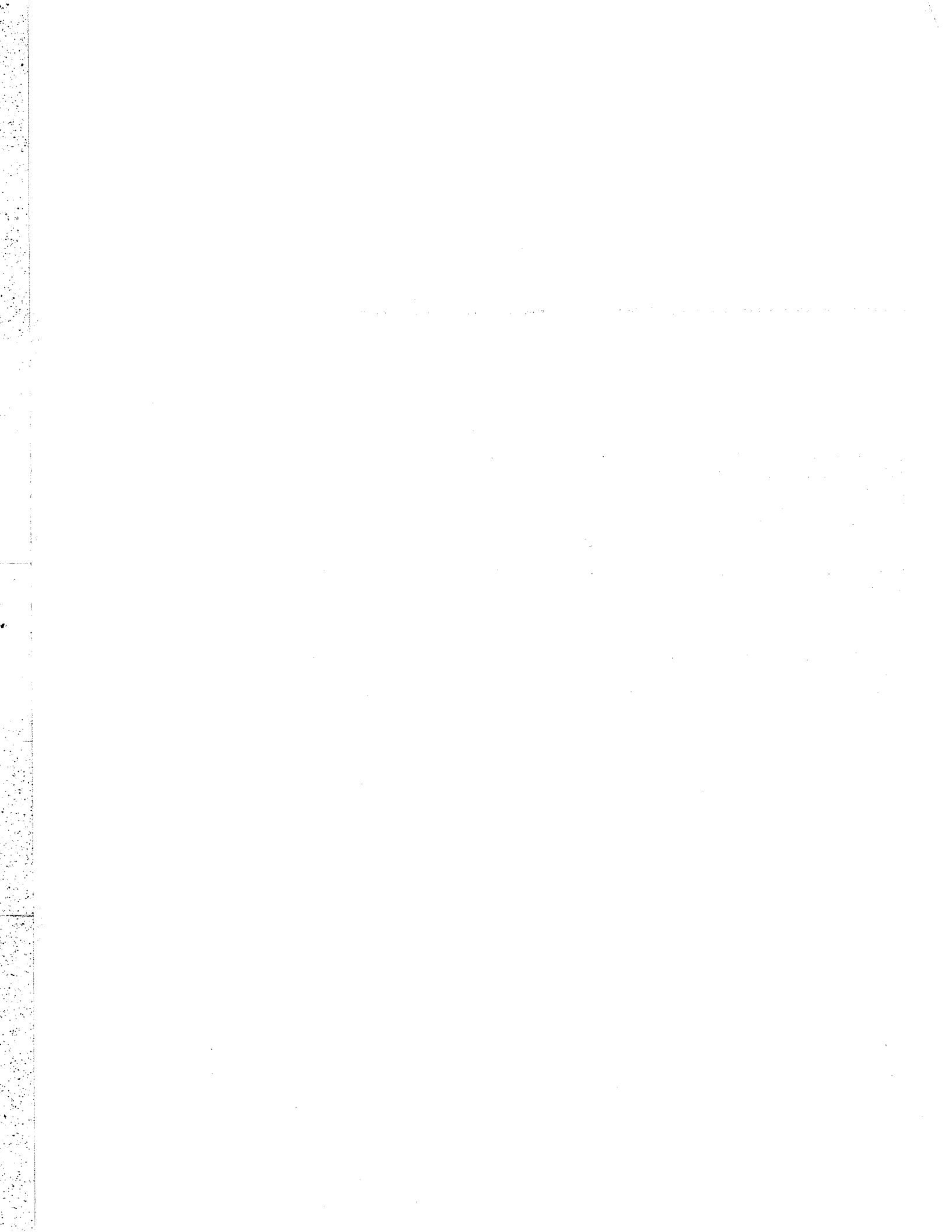
<b>Traffic Distribution Fitting: A Systematic Methodology</b> <i>Russell Thompson and William Young</i>	87
<b>CARSIM: Car-Following Model for Simulation of Traffic in Normal and Stop-and-Go Conditions</b> <i>R. F. Benekohal and Joseph Treiterer</i>	99
<b>Work Zone Analysis Model for the Signalized Arterial</b> <i>C. Thomas Joseph, Essam Radwan, and Nagui M. Roupail</i>	112
<b>Application of the Image Analysis Technique To Detect Left-Turning Vehicles at Intersections</b> <i>Yean-Jye Lu, Yuen-Hung Hsu, and Guan C. Tan</i>	120
<b>Some Properties of Macroscopic Traffic Models</b> <i>Paul Ross</i>	129
<b>Generating Partial Origin-Destination Tables for Streamlined Application of Corridor Models</b> <i>Sam Yagar</i>	135
<b>Modeling the Effect of Traffic Signal Progression on Delay</b> <i>Kenneth G. Courage, Charles E. Wallace, and Rafiq Alqasem</i>	139
<b>Validation of Saturation Flows and Progression Factors for Traffic-Actuated Signals</b> <i>Pano D. Prevedouros and Paul P. Jovanis</i>	147
<b>Modeling of Shared Lane Use in TRANSYT-7F</b> <i>Charles E. Wallace and Frank J. White</i>	160
<b>Critical Movement Analysis for Shared Left Turn Lanes</b> <i>Herbert S. Levinson</i>	167

---

---

<b>Computation of Signalized Intersection Service Volumes Using the 1985 Highway Capacity Manual</b>	<b>179</b>
<i>Kenneth G. Courage and John Zen-Young Luh</i>	
DISCUSSION, <i>Edmond C.-P. Chang</i> , 187	
AUTHORS' CLOSURE, 190	
<hr/>	
<b>Some New Data That Challenge Some Old Ideas About Speed-Flow Relationships</b>	<b>191</b>
<i>B. N. Persaud and V. F. Hurdle</i>	
<hr/>	
<b>Capacity Analysis of Two-Lane Highways</b>	<b>199</b>
<i>David L. Guell and Mark R. Virkler</i>	

---



# Volume-Based Model for Forecasting Truck Lane Use on the Rural Interstate

D. ALBRIGHT AND C. BLEWETT

Truck lane use is important when assessing the impact of trucks on multilane roads. A study was conducted to estimate truck lane use on tangent sections of the New Mexico rural interstate. Following review of related literature, a field study methodology was developed, data were collected, and a model was derived for estimating a truck lane distribution factor. The study determined that total vehicle volume is statistically significant in explaining truck lane use. The percent trucks of total volume also adds to the explanatory ability of the model. Three models were developed from the field data, based on three traffic volume groups. To diminish the possibility of underbuilding a facility, precision ranges were established for the predictive ability of the models. The models were modified to account for variability in the data. An analysis was conducted of the impact of study models on pavement structural design. The study models, when compared with previous lane use estimates, typically reduced pavement thickness by  $\frac{3}{4}$  in. For all study sites the reduction in pavement thickness was between  $\frac{3}{4}$  and 1 in.

The lane distribution factor (LDF) is an integral part of load calculations on multilane facilities. The LDF describes the percent of heavy commercial vehicles in the right lane. This factor is used in design of pavement thickness, affecting pavement design and construction costs. For the New Mexico rural interstate system, the truck lane distribution factor for pavement design assumed that 94% of the heavy commercial traffic travels in the right-hand lane and the remaining 6% occupies the left lane. These values were estimates and were not specifically based on New Mexico traffic data. To determine if these factors accurately described New Mexico rural interstate truck lane use, a research program was begun to review the literature on truck lane distribution.

The review of existing literature yielded two conclusions. First, little research has been published on the subject. Second, the published studies available were completed before enactment of the Surface Transportation Assistance Act of 1982. Because heavy commercial vehicle characteristics have changed significantly since that legislation, the New Mexico State Highway Department felt that the issue was important enough to warrant a study. The objectives of the study were to evaluate the existing procedure and, if appropriate, to develop a new truck lane distribution formula.

## STUDY DESIGN

The truck lane distribution study involved three elements: (1) the methodology to facilitate accurate data collection and

ensure the validity of the data for statistical inquiry; (2) the process of data collection, assimilation and summarization, including the verification of base data; and (3) the statistical analysis of summary data and generation of a valid set of lane distribution postulates.

## Methodology

Based in part on the literature, the research procedure was modified to address the problem of personnel constraints and accurate data collection. The first step was to determine which variables were most important for the study. The percent of heavy commercial traffic in the right lane was defined by the nature of the study as the dependent variable. The remaining issues to be answered by the design were the selection of measurable independent variables, determination of sample size, and site selection.

The selection of independent variables is typically limited by resources. Variables are selected which previous experience or inquiry indicate have promise. These variables are then reduced by some scale of priority, according to the resources available.

Based on previous research, notably that conducted in 1971 by M. M. Alexander and R. A. Graves for the Georgia Highway Department (1), the independent variables for the study were limited to total traffic volume, percent heavy commercial with five and more axles, and percent heavy commercial with less than five axles. The last two independent variables were totaled separately by lane during data collection and analysis.

Vehicle speed was considered as a variable but was not selected. Driver detection of radar might result in atypical data, particularly among heavy commercial vehicles. Additionally, if speed were determined to improve the lane use model, it would have to be collected for each site where lane use was to be calculated, increasing the cost of model implementation. For the remaining variables used in the study, system-wide data were available so that lane use calculations could be readily made. This is a critical element of a useful model for determining truck lane use distribution.

## Sample Size and Site Selection

After selection of independent variables, the next step was to define the universe of the study and identify an adequate sample size. A previous study determined that New Mexico's 900-mile rural interstate system could be subdivided into fourteen sections with unique volume characteristics. Because these fourteen sections represented the entire rural interstate system, and because the determinant criteria for a unique section

were volume and percent heavy commercial of total traffic, the sections were appropriate units for determining sample size. A sample of five sites was randomly selected from these sections. After the five sites were selected, they were reviewed to determine if they were representative of the New Mexico rural interstate. The review resulted in the addition of another site representative of lower traffic volumes and lower percent heavy commercial of total traffic. All subsequent analysis in this study is based on data from these sites.

Once the six sections were identified, the road segment for the count was identified. Each section has a corresponding permanent counter located in a unique road segment. In the interest of accurate data collection, the field data collector would count only heavy commercial traffic. The total volume would be determined by the permanent counter. Therefore, the road segment for the count was located between the same two interchanges as the permanent counter. Because access to the rural interstate system is controlled, the heavy commercial counts could be compared directly with total volume counts of the permanent counter for the same hours to complete the data base.

The remaining concern for site selection was the minimum distance the site had to be located from the nearest interchange. Based on standard acceleration rates, it was determined that the site for the count would preferably be 2 miles from the nearest interchange. This constraint ensures that the site will not be affected by atypical lane distributions that result from merging. With typically 5 miles between interchanges, there was a minimum of a 1-mile road segment within which the specific count site was selected.

### Count Days and Hours

In order to minimize seasonal factors affecting the study, the month of September was chosen as the time to complete the field counts. Historical volume data for the New Mexico rural interstate indicate that the traffic ratio of September mean daily traffic to annual mean daily traffic is close to unity.

At the time of site selection, it was decided that each count would be made on three consecutive days. To avoid weekend variation in traffic volume and to provide adequate time for personnel travel and set-up, the study count days were Tuesday, Wednesday, and Thursday.

An analysis of the daily hourly volumes at the six sites during previous Septembers revealed no discernible morning or afternoon peak in all cases. The typical daily pattern reflected a fairly continuous flow of traffic from 9:00 a.m. until 5:00 p.m. The data did suggest that in most cases the volume decreased from noon to 2:00 p.m. The count hours for the study were selected as 9:00 a.m. to noon, and from 2:00 p.m. to 5:00 p.m. for all six locations. Based on previous years' data, these hours would include any peaking characteristics for all locations and represent characteristic traffic volume at the sites.

### DATA COLLECTION, ASSIMILATION, AND SUMMARIZATION

After all field data were collected, the raw data were assimilated and summarized for analysis. This process involved

surveying the raw data, eliminating unusable records, aggregating the data to different levels, and determining the most useful summary data set for the development of lane distribution factors.

For each of the six sites, data were collected for both directions. This, in effect, doubled the site size for data summarization purposes. Under optimal conditions, 36 hours of data would be available, by direction, for each study site. Of the total of 216 possible hours of observation, it was determined that 44 could not be used in the analysis. Of these, 2 hours were discarded because of lane closures, 6 were lost due to field transportation problems, and the remaining 36 were dropped because volumes from the permanent counter were suspect. Fortunately, the 172 good records were distributed evenly among the six sites and included data for traffic moving in both directions at each site. No sites were dropped from the study because of insufficient data, and no recounting was required.

To analyze the variance of the data, the mean square error of each variable was examined at each of the six sites, by direction. A high within-group variance would suggest that sampled traffic and lane patterns were not uniform. Alternatively, with review of the standard error of the mean to estimate sampling error, high within-group variance might suggest that the three-day sample was inadequate and the study would have to be redesigned. On the other hand, given low within-group variance as indicated by mean square error and standard error, consistent site-specific traffic and lane use patterns could be concluded, and further analysis of the data would be indicated. The within-group mean square error was found to be low. The mean statistics for each site, and each site by direction, adequately represent traffic in the lane distribution factor model.

The relationship among mean traffic data for the sites, by direction, was addressed through regression analysis. The best fit equation was determined for mean values by site and direction. The statistical analysis system (SAS) procedure used was PROC STEPWISE MAX R.

In the stepwise procedure the use of the two independent variables provided a better fit than single variable. The *R*-squared value for the best one-variable model, with total volume the independent variable, was .5845. The *R*-squared value for the best fit two-variable model, total volume and percent trucks of total volume, was .6087.

The two-variable model was selected despite relatively low improvement in the fit of the model. On some segments of the rural interstate, percent trucks can affect the resulting LDF equation. When the percent trucks of total volume exceeds 16%, this variable affects the LDF. At 32% this variable reduces the LDF by 1%. Some New Mexico rural interstate segments currently exceed 38% trucks of total volume. This variable is also intuitively sensible. As the percent trucks of the volume increases, the trucks with higher weight-to-horsepower ratio will on occasion be forced into the left lane.

The two-variable model, or base equation, for the study, is shown below. Figure 1 is a plot of equation 1 with the assumption of 26% truck traffic.

$$\begin{aligned} \text{LDF} = & .98144 - .0004(\text{Volume}) \\ & - .000293(\text{Percent Trucks}) \end{aligned} \quad (1)$$



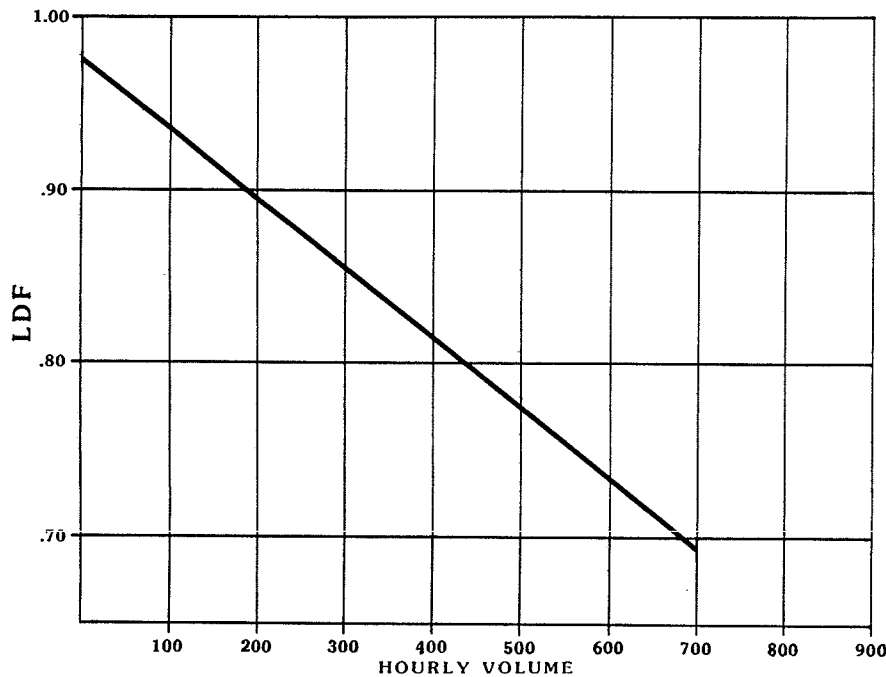


FIGURE 1 Equation 1 (assume 26% trucks).

This equation is only for tangent segments of the rural interstate. Because the data used for model development cover a limited range of volumes, it is important to note the traffic volumes at which the model does not provide useful information. The equation fails at volumes greater than 1,230 per hour, per direction. At this point the equation provides less than half the vehicles in the right lane. No more than a 50/50 split, without other access/egress factors, would occur, even at level of service E. The volume or heavy truck percent at which the rural interstate LDF reduces to equilibrium, a 50/50 split, is not known and cannot be accurately extrapolated from the data collected for this study.

What is the volume at which the LDF equation should not be applied? The equation should not be used with volumes greater than 700 vehicles per hour, per direction. At this level there is an LDF of 70%—the lowest observed right lane percent of heavy commercial trucks on the rural interstate. Extrapolation beyond the observed percent is not recommended because it could result in underbuilding a section of rural interstate.

While the primary purpose of the study is to check the assumption of 94/6 distribution between the right and left lanes and perhaps to prevent overbuilding of a facility, underbuilding is also of concern. This concern not to underbuild a section of interstate directed the study to modifications of equation 1.

The results of the study demonstrate that current distribution assigns too much of the heavy commercial traffic to the right lane. At no location surveyed was the LDF as high as the standard LDF value previously employed by the New Mexico State Highway Department. This suggests that the interstate construction cost could be reduced through a corrected load analysis resulting from more accurate LDF.

Equation 1, based around a line fit to the mean data by site and direction, reduces the possibility of overbuilding through error in the LDF. To prevent underbuilding through LDF

error, the LDF must not be underestimated. This concern is the basis for modification to the equation.

#### RANGE OF ACCEPTABLE ERROR

To use the equation would essentially “miss” the mean as high as to miss it low because of the sum of squares step in fitting the regression line to the data. The equation should be modified so it is fit not only to the mean, but also to the range of desirable values. This range will be characterized by a smaller margin of acceptable error below the mean, and a larger margin of acceptable error above the mean.

Two ranges of error were identified in discussions with personnel from the Materials Laboratory and Technical Services Engineering. These can be used to compare and evaluate the results of equation 1 with field observation data. A standard for the result of the equation is an LDF equal to or greater than  $-1$  percentage point of the observed value, and equal to or less than  $+3$  percentage points of the observed value.

A less precise, but acceptable, standard would be the equation LDF providing dependent variable values equal to or greater than  $-2$  percentage points of the observed value, and equal to or less than  $+5$  percentage points of the observed value. The ranges of values identified reflect the concern to overestimate rather than underestimate the LDF.

At the lower precision level, equation 1 results in eight of twelve sites being defined within the range. However, at the higher precision level, equation 1 drops to four of twelve sites within the range. Equation 1 results in low mean and cumulative errors, of  $-.00907$  and  $-.10885$  respectively. However, both errors are minus, more typically below the mean field observations.

The model developed to this point does not satisfactorily interpret all of the stations. Equation 1 produces acceptable results for study sites with traffic volumes between 400 and 450 vph.

**VOLUME GROUPS**

There are three groups of total volumes. Given the desirable range around the mean, the equation 1 intercept parameter operates in a linear manner with the coefficients of the independent variables only within one range of total traffic volume. Volumes less than 400 and greater than 450 vph are modeled to the desirable range. This was accomplished by adding half a standard deviation to the intercept for the low volume group, and one standard deviation for the high volume group. Adding the standard deviation to the equation increases the LDF. The equation is expressed with the standard deviation added to the intercept parameter. The intercept parameter is, therefore, modified to model the characteristics of the three volume groups. The volume groups, and the modified parameters, are noted below by range.

Equation	Volume Range (vph)	Equation Number
LDF = 1.02144 - .0004(Volume) - .000293(T)	405-700	2
LDF = .98144 - .0004(Volume) - .000293(T)	400-449	1
LDF = 1.00144 - .0004(Volume) - .000293(T)	10-399	3

With these equations, the projected LDF would be within the more precise range five of twenty sites, and twelve of twelve within range at the less precise range. For the less precise range, six sites were overestimated but in range, four sites were underestimated but in range, and there was no error at two sites.

Equation 1 when applied to all traffic volumes performed satisfactorily to 700 vph. The use of three equations for the LDF by volume range applies to 800 vph.

**FURTHER ADJUSTMENT FOR THE HIGHER VOLUME RANGE**

A further adjustment may be made to ensure that there is no underestimation of the LDF. No adjustment is needed for volume ranges 10-399 and 400-449. Underestimation at higher volumes suggests that the intercept adjustment for this volume is not sufficient. An additional .03 added to this intercept would provide, at higher volume sites, LDFs that are overestimated and in range, rather than underestimated and in range. This modification will ensure that the LDF is not understated, and that the design and cost gains through the study are only slightly diminished. The effect of this additional modification is to increase the equation by slightly under two standard deviations. The precision range result is similar to that of a confidence interval modification of the base equation.

**LDF EQUATIONS**

The final equations by volume on the rural interstate are noted below.

Equation	Volume Range (vph)	Equation Number
LDF = 1.05144 - .0004(Volume) - .000293(T)	405-700	4
LDF = .98144 - .0004(Volume) - .000293(T)	400-449	1
LDF = 1.00144 - .0004(Volume) - .000293(T)	10-399	3

The above equations were evaluated by testing for homogeneity of variance. PROC TABLES from the SAS Supple-

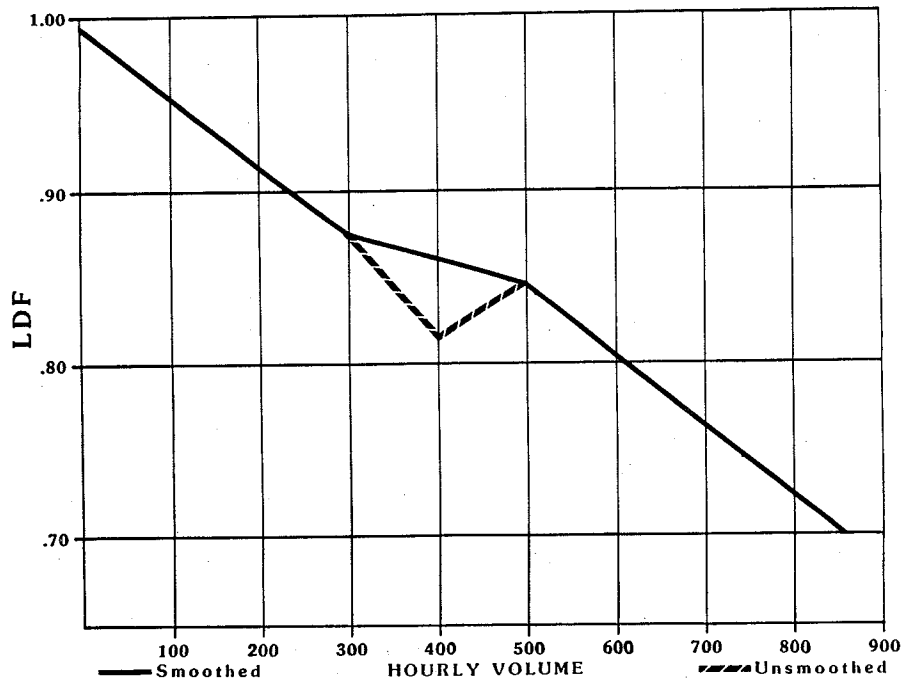


FIGURE 2 Adjusted equation (assume 26% trucks).

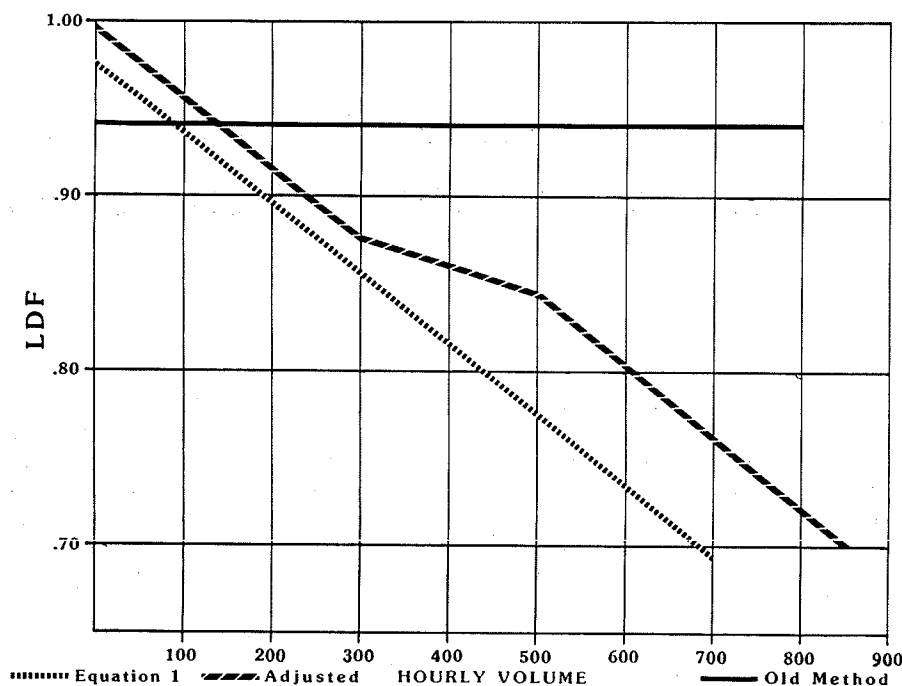


FIGURE 3 Comparison of methods (assume 26% trucks).

mental Library was used to test the homogeneity of variance in volume and percent trucks to volume. Using Bartlett's procedure at  $p \leq 0.001$ , the variance was found to be homogeneous. For purposes of comparison with the above equations, despite acceptance of homogeneity, weighted regression analysis was conducted using SAS PROC REG. The resulting high and low volume group equations derived lower LDFs than calculated by the above equations. If the concern is not to underbuild a facility, the equations would not be acceptable.

To present the resulting equations for daily departmental use, the LDF relationship to volume and trucks is presented in graphic form. The equations line is smoothed at the middle volume range. In the process of smoothing the regression equation line, the resulting LDF for the middle volume range is also higher than that calculated from the weighted regression equation. Figure 2 illustrates the use of the three equations with volume group 2 smoothed.

## CONCLUSION

The lane distribution factors put forth in the study represent a significant improvement in accuracy over earlier methods. Figure 3 compares the earlier method with the results of the present study, for the case of 26% truck traffic. The models can be applied easily because the data required are available on a system-wide basis and because the LDFs can be represented in graphic form. The models reflect the differences in

volume and heavy commercial percent of total volume that are characteristic of New Mexico's rural interstate system.

Using the lane distribution factors will result in better road design and may provide considerable savings. A sensitivity analysis of the impact of this procedure on pavement structural characteristics was conducted by the Materials Testing Laboratory and Planning Bureau of the New Mexico State Highway Department. The analysis determined that for longer-term pavement design the lane distribution factors reduce pavement thickness by an average of  $\frac{3}{4}$  in. For all of the sites in the present study, the pavement design was reduced between  $\frac{3}{4}$  and 1 in. pavement thickness. It may be concluded that the lane distribution factors derived from this study improve the accuracy of determining truck lane usage, and that this improvement significantly affects pavement design.

Additional research will be undertaken, including analyzing the effect of positive grades on truck lane use, sampling for monthly variation, and examining the time stability of LDF equations.

## REFERENCE

1. M. M. Alexander, Jr., and R. A. Graves II. *Determination of the Lateral Distribution of Truck Traffic on Freeway Facilities*. Georgia Highway Department, Atlanta, Nov. 1971.

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

# Comparative Analysis of Two Logics for Adaptive Control of Isolated Intersections

FENG-BOR LIN

Adaptive signal control has the potential to provide improved control at isolated intersections. Adaptive control, however, has limitations due to its need to rely on estimated flow conditions for making signal timing decisions. Such estimated flow conditions always differ from the actual conditions, and the discrepancies can offset the benefit of having an elaborate decision making process in a control logic. Therefore, an issue can be raised as to whether it is necessary to rely on strenuous decision-making processes for adaptive control. This study compares the relative merits of a simple queue based logic and a logic that relies on a much more complicated procedure for making timing decisions. It is found that the queue-based logic is nearly as effective as the more complicated logic. This finding points to a direction for the development of new control logics that can be widely used to replace existing traffic-actuated control logics.

Adaptive control, referred to herein, is a mode of control which relies on very short-term advance vehicle arrival information in an attempt to achieve real-time optimization of signal operations. Several adaptive control logics have been tested in the field, implemented, or recommended for use at isolated intersections. Examples of such logics include modernized optimized vehicle actuation strategy (1), Miller's algorithm (2), optimization policies for adaptive control (3), traffic optimization logic (4), and stepwise adjustment of signal timing logic (5). The split, cycle, and offset optimization technique (6), which is intended mainly for signal coordination, has also been tested for the control of isolated intersections (7).

It should be noted that, regardless of the level of sophistication of a control strategy, optimal signal operations can never be achieved in a real life situation. The term optimization is often used casually to represent a process of searching for a better course of action. Such a process can be based on a straightforward trade-off analysis in order to determine whether the current green duration should be extended for a short time interval (4). It can also be based on an elaborate procedure to evaluate alternative signal switching sequences and, subsequently, to identify the best sequence for a relatively long (e.g., 100 sec) future period of time (3). This liberal interpretation of optimization is also adopted in this paper.

## INTRODUCTION

The extent to which adaptive control can improve signal operations depends in part on the specific adaptive control logic

employed, the quality of the information used for making timing decisions, the level of control efficiencies delivered by existing control devices, and the traffic flow patterns involved. Field tests conducted so far have shown mixed but encouraging results for adaptive control. A test of Miller's algorithm (8) at an intersection revealed that the resulting control efficiencies were poorer than those provided by vehicle-actuated controls when the flows approaching the intersection were less than 1,300 vph. A test of modernized optimized vehicle actuation strategy (1) for ten time-of-day periods resulted in delay reductions of 5% to 12% for seven periods, a delay increase of 7% for one period, and delay reductions of 20% and 30%, respectively, for the remaining two periods. The implementation of the traffic optimization logic (4) at one intersection yielded delay reductions of more than 20% when compared with a traffic-actuated operation.

As demonstrated in these field tests, the incorporation of an optimization capability into signal control does not necessarily guarantee improved signal operations. The accuracy of the information utilized to make timing decisions is also critical to adaptive control. This can be a drawback, because adaptive control usually relies on estimated flow conditions rather than on actual flow conditions. To facilitate the estimation of flow conditions, it is necessary to place detectors several hundred feet upstream of the intersection in order to provide advance vehicle arrival information. Such detectors can provide perhaps no more than 115 sec of advance information at most intersections. In order to overcome this limitation, some researchers (3, 8) have resorted to the use of predicted vehicle arrival data to supplement the detector data. This approach tends to introduce errors into the information that is used to make signal timing decisions. Even if the estimation of the flow conditions is based entirely on detector data, the resulting estimates can be expected to deviate from the actual flow conditions. The discrepancies between the estimated and the actual conditions can be attributed to lane changes and to variations in vehicle speeds, queue discharge headways, driver responses to signal change interval, and so forth.

In light of this drawback, it is worth investigating whether strenuous decision-making processes can be replaced by simple decision rules for adaptive control. To address this issue, two adaptive control logics are compared in this study. One logic is stepwise adjustment of signal timing (SAST) (5), which requires the evaluation of alternative signal switching sequences in order to reach a signal timing decision. The other logic is based on the consideration of queue length and determines whether the current green should be terminated by comparing the expected maximum queue length of the current green phase with a threshold queue length.

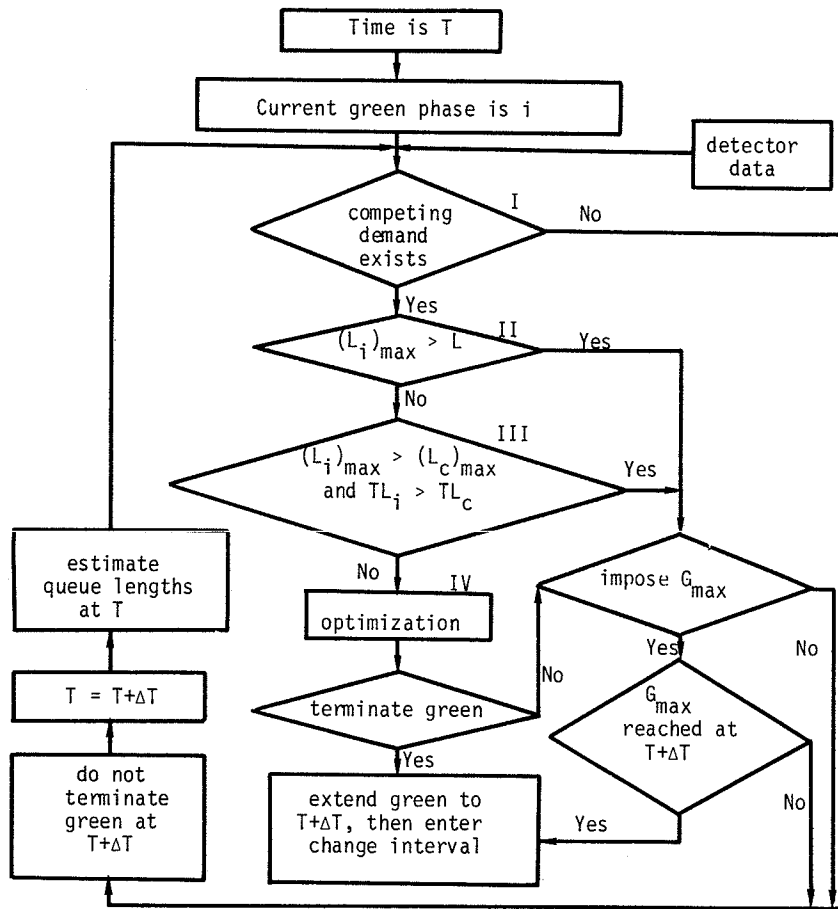


FIGURE 1 Decision-making process for stepwise adjustment of green.

### SAST LOGIC

SAST logic relies on a binary choice decision-making process for stepwise adjustment of signal timing. In this decision-making process, time is divided into small intervals, or steps. In each step, an analysis is made to determine whether the current green should be terminated at the end of that step. The rationale for the development of this logic is discussed elsewhere (5).

The decision-making process adopted in SAST logic for stepwise adjustment of a green duration is shown in Figure 1. This process has four levels of decision-making activities, which are marked in the figure as I, II, III, and IV, respectively. The first three levels employ simple decision rules which either permit the current green to be extended beyond the first step or call for additional analyses. The data processing requirements for these three levels are very limited. Signal optimization, which is the last level of the decision-making process, comes into play only if the first three levels fail to choose a definitive course of action.

The manner in which SAST logic processes and utilizes information to reach signal timing decisions is described in detail below.

### Data Acquisition and Processing

Time is divided into successive steps in SAST logic. Each step is  $\Delta T$  sec in length. A decision must be made in each step

either to extend the green beyond the current step or to terminate the green at the end of that step. Referring to Figure 2, let  $T$  be the beginning of a step. At least two types of data are needed for making a timing decision. One type of data is the vehicle arrival sequence that is expected at the stop line in several steps beyond  $T$ . This type of data is derived from vehicle arrival data obtained by the upstream detectors. The procedure for deriving such data is simple.

Let  $t_i$  be the arrival time of a vehicle at an upstream detector location and  $r$  the average travel time between the detector and the stop line in the absence of interferences by signal operations. Then, the expected arrival time of that vehicle at the stop line is assumed to be  $A_i = t_i r$ . This expected arrival time can be used directly for decision making. It can also be represented as one arrival in a specified step. The latter approach enables more efficient data processing. Therefore, it is adopted in SAST logic. Following this approach,  $A_i$  is transformed into one vehicle arrival in the  $k$ th step beyond  $T$  if  $A_i$  falls in that step. Since the efficiency of adaptive control can be sensitive to the errors in the vehicle arrival sequence that is used for signal optimization, the step size  $\Delta T$  should be sufficiently small. Large step sizes will distort an arrival sequence. Two second steps are a reasonable choice. In each of such steps, the number of vehicle arrivals will rarely exceed one. On the other hand, the step size should be sufficiently large in order to allow time for data processing, signal optimization, and implementation of a timing decision.

Because the number of arrivals in each step is derived from

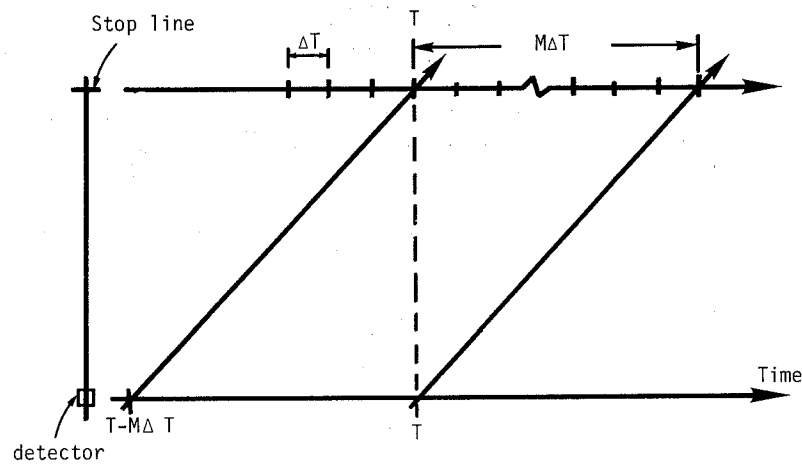


FIGURE 2 Discretization of time for timing adjustment.

detector data, a finite amount of advance information is available at a given point in time. Referring again to Figure 2, if the average travel time between an upstream detector and the stop line is equivalent to  $M$  steps, then only those vehicles detected between  $T - M\Delta T$  and  $T$  will produce advance information for decision making at  $T$ . In other words, the amount of advance information available at  $T$  is  $M\Delta T$  sec.

The second type of data needed for decision making is the expected queue length in each traffic lane at time  $T$ . Such queue lengths are determined from a traffic model, which is an integral part of SAST logic, and are defined as the differences between expected cumulative arrivals and departures as measured at the stop line for a specified point in time.

### Level I Decision Making

If competing demands for the right-of-way do not exist, there is no reason to terminate the current green phase. At a given time  $T$ , competing demands are considered to be nonexistent if all the phases waiting for the right-of-way have the following expected flow conditions:

1. There are no queuing vehicles in any lane at  $T$ , and
2. There are no vehicles expected to arrive at the stop line in  $n$  steps following  $T$ .

A reasonable value of  $n$  is one such that  $n\Delta T$  is about 6 sec. If a competing demand after a period of  $n\Delta T$  results in a decision to terminate the green,  $n\Delta T$  shorter than 6 sec may unnecessarily force an approaching vehicle to a complete stop before it is given the right-of-way.

### Level II Decision Making

This level of decision making is based on the maximum expected queue length  $(L_i)_{\max}$  of the current green phase at  $T + \Delta T$ . If this queue length exceeds a specified threshold value  $L$ , the current green is automatically extended beyond  $T + \Delta T$ , subject to a maximum green constraint. This feature is important: if adaptive control relies exclusively on signal optimization, a phase with a relatively low demand is likely to lose

the right-of-way before most of its queuing vehicles enter the intersection. The result is poor signal operation. This problem can be eliminated when a threshold queue length is used to bypass signal optimization.

For example, if the threshold queue length is set at four vehicles, the queue lengths of the current phase can be reduced to about four vehicles before other phases are allowed to compete for the right-of-way through signal optimization. This ensures that the queue lengths of every phase will not grow at excessive rates due to the existence of short green intervals. A previous study (5) reveals that the best threshold queue length to use appears to be about four vehicles. This implies that it is best to allow the queue lengths of the current green phase to be reduced to approximately four vehicles before other phases are allowed to compete for the green.

To prevent exceedingly long green durations, SAST logic also allows the imposition of a maximum allowable green  $G_{\max}$  on the current phase. This  $G_{\max}$ , however, is imposed only when the maximum queue length of all completing phases exceeds a specified threshold value.

### Level III Decision Making

This level of decision making takes into consideration the queue lengths of the current green phase and those of all competing phases. The current green is extended beyond  $T + \Delta T$  if the following two conditions are satisfied:

1. The maximum queue length  $(L_i)_{\max}$  of the current green phase  $i$  is longer than the maximum queue length  $(L_c)_{\max}$  of all competing phases, and
2. The total number of queuing vehicles  $TL_i$  of the current green phase is larger than the total number of queuing vehicles  $TL_c$  of all competing phases.

### Level IV Decision Making

This level of decision making involves signal optimization. SAST logic allows minimization of the total delay either of all vehicles or of the vehicles in certain critical lanes. The subject vehicles include the queuing vehicles at  $T$  and those

vehicles which are expected to reach the stop line between  $T$  and  $T + M\Delta T$ . The critical lanes to be included for signal optimization can vary from one step to another. They are determined according to the following criteria for each phase:

1. A lane that has a long queue length at time  $T$  is more critical than a lane that has a short queue.
2. If two lanes have equal queue lengths at time  $T$ , the lane that has a larger number of expected arrivals between  $T$  and  $T + 2\Delta T$  is more critical.
3. If two lanes have equal queue lengths at  $T$  and equal numbers of expected arrivals between  $T$  and  $T + 2\Delta T$ , the lane which is ahead in the data processing order is more critical.

### Optimization Process

The signal optimization process is illustrated in Figure 3. The first task in this process is to examine the option of terminating the green at  $T + \Delta T$ . This option leads to several alternative signal switching sequences. These sequences are generated and evaluated in order to estimate the minimum delay  $D_{\min}$  associated with this option. The next task is to determine whether  $D_{\min}$  can be reduced by extending the green beyond  $T + \Delta T$ . This task is carried out by first considering the option of terminating the green at the end of the second step, i.e., at  $T + 2\Delta T$ . The signal switching sequences associated with

this option are generated and evaluated one at a time. If the delay  $D_n$  resulting from such a switching sequence is less than or equal to  $D_{\min}$ , it is more desirable to extend the green beyond  $T + \Delta T$ . In such a case, the current green is allowed to continue unless the maximum green constraint prohibits further extension of the green. If  $D_n$  is greater than  $D_{\min}$  instead, another signal switching sequence is generated and evaluated in the same manner until all alternative sequences associated with terminating the green at  $T + 2\Delta T$  are exhausted. Following that, the option of terminating the green at  $T + n\Delta T$  for  $n = 3, 4, \dots$  may be evaluated.

SAST logic uses a decision variable  $N_{\max}$  to limit the maximum number of options that are to be evaluated. For example, if  $N_{\max} = 3$  is specified, only those signal switching sequences involving the termination of the green at  $T + \Delta T$ ,  $T + 2\Delta T$ , and  $T + 3\Delta T$  are considered for evaluation. With  $M$  steps of advance information, the value of  $N_{\max}$  can vary from 2 to  $M$ .

### Generation of Switching Sequences

Given that the current green is to be terminated at  $T + n\Delta T$  ( $n = 1, 2, \dots, M$ ), SAST logic does not attempt to generate all feasible signal switching sequences for evaluation. Instead, it generates a small number of switching sequences that are likely among the best few of all feasible sequences.

The process of generating signal switching sequence can be

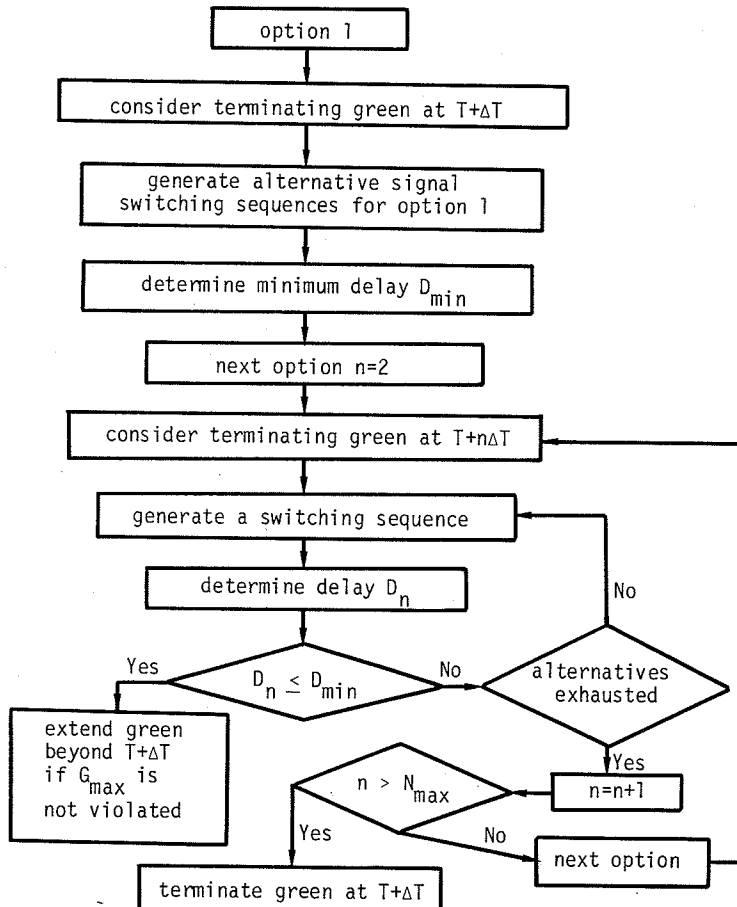


FIGURE 3 Signal optimization process.

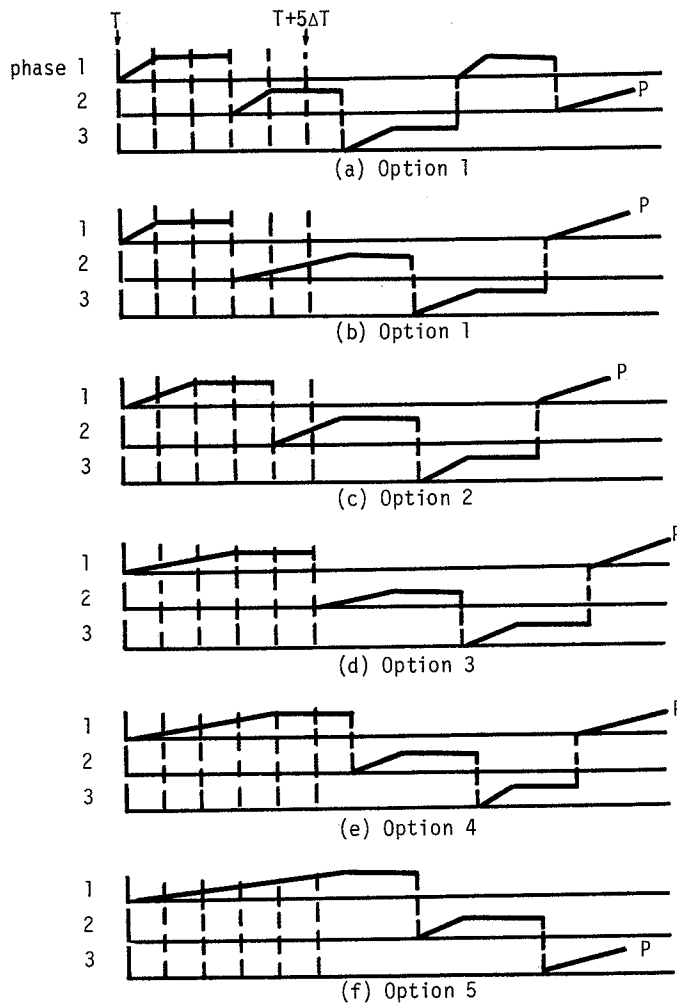


FIGURE 4 Example of generated signal switching sequences.

better described with the example shown in Figure 4. This figure should be interpreted as follows:

1. The total number of signal phases is three.
2. Phase 1 has the green at time  $T$ .
3. The signal change interval following each green equals two steps.
4. A bold ascending line indicates that a phase has the green, and a bold horizontal line signifies a signal change interval.
5. Advance information is available for  $M = 5$  steps beyond  $T$ .
6. Switching sequences marked as (a), (b), (c), (d), (e), and (f), respectively, are arranged according to the order in which they are generated.

In Figure 4a, the current green is terminated at  $T + \Delta T$ . Subsequently, a signal change interval is timed out at  $T + 3\Delta T$ . The green is then given to the next phase, i.e., phase 2, if that phase has a demand for the right-of-way. If phase 2 has no demand, the green is given to phase 3 by skipping phase 2, provided that phase 3 has a demand for the right-of-way. If neither phase 2 nor phase 3 has a demand for

the right-of-way, the green is given to phase 2 which follows the current green phase. The resulting signal switching sequence is likely to be inefficient, and this will be reflected in the timing decision.

A demand is recognized in SAST logic if any lane of the phase in question satisfies either of the following conditions:

1. Expected queue length at the time the last change interval is terminated (e.g.,  $T + 3\Delta T$ ) is greater than zero.
2. Expected number of arrivals within 4 sec after the termination of the change interval is greater than zero.

If phase 2 receives the green at  $T + 3\Delta T$ , this green can be extended by one step (Fig. 4a), or by two steps (Fig. 4b). Once the generated portion of a switching sequence reaches the end of the last step, i.e.,  $T + 5\Delta T$ , one of the following actions is taken in generating the portion of the switching sequences beyond  $T + 5\Delta T$ :

1. If a signal change interval is in effect at the end of the last step, i.e., at  $T + 5\Delta T$ , that interval is allowed to be timed out at or beyond  $T + 5\Delta T$  (Fig. 4a, 4d, and 4e). Afterwards, the green is given to the next phase that still has vehicles waiting to enter the intersection. This green is allowed to continue until all the vehicles in that phase are discharged.
2. If a green interval is in effect at  $T + 5\Delta T$  (Fig. 4b, 4c, and 4f), this green interval is extended beyond  $T + 5\Delta T$  until all the vehicles are discharged.
3. The generation of a signal switching sequence is completed when all the vehicles included in the analysis are presumably discharged. This point in time is denoted as  $P$  in Figure 4.

#### Estimation of Delays

The delay experienced by a vehicle is measured as the expected departure time minus the expected arrival time at the stop line in the absence of interferences by signal operations. SAST logic estimates only the delays of those vehicles which are expected to reach the stop line by  $T + M\Delta T$ . Therefore, it is assumed that there are no additional vehicle arrivals beyond  $T + M\Delta T$ .

Delays are estimated simultaneously with the generation of each signal switching sequence. When the front portions of several signal switching sequences are identical, the delays related to such portions are only estimated once in order to reduce the CPU time. For example, the first two signal switching sequences depicted in Figure 4 have the same switching pattern between  $T$  and  $T + 3\Delta T$ . Therefore, the delays incurred in this period and estimated for the first switching sequence are used directly for the second sequence.

The total delay associated with a signal switching sequence is the sum of the delays incurred in each step. The delays in each step can be estimated from the cumulative arrivals and departures both at the beginning and at the end of that step. To estimate such delays for each lane, the following two quantities are defined first:

$$CA_o = \text{Que}(T) \quad (1)$$

$$CD_o = 0 \quad (2)$$



where

$$\begin{aligned} CA_o &= \text{cumulative number of expected arrivals at } T, \\ \text{Que}(T) &= \text{queue length at } T, \text{ and} \\ CD_o &= \text{cumulative number of expected departures at } T. \end{aligned}$$

Next, the cumulative number of arrivals  $CA_i$  at the end of Step  $i$  ( $i = 1, 2, 3, \dots$ ) can be determined as

$$CA_i = CA_{i-1} + NA_i \quad (3)$$

where  $NA_i$  = number of expected arrivals in step  $i$ .

For  $i = 1, 2, \dots, M$ ,  $NA_i$  is derived from the detector data. For  $i > M$ ,  $NA_i$  is set equal to zero because SAST logic does not consider the delays of those vehicles not yet detected by  $T$ .

The cumulative number of departures  $CD_i$  at the end of step  $i$  is determined as

$$CD_i = CD_{i-1} + ND_i \leq CA_i \quad (4)$$

where  $ND_i$  = number of expected departures in step  $i$ .

In the original SAST logic, vehicle departures from the stop line were treated as discrete events. In this study, the departures of queuing vehicles are treated as a continuous variable in accordance with a nonlinear function of saturation flow. After a queue is completely discharged from the stop line, the number of expected departures in a step is set equal to the number of expected arrivals in the same step, i.e.,  $ND_i = NA_i$ .

Equations 3 and 4 are applied in a stepwise manner, beginning with the first step ( $i = 1$ ) that starts at  $T$ . The combined delay of the vehicles in a lane in step  $i$  is estimated as

$$\text{DELAY}_i = (CA_{i-1} + CA_i - CD_{i-1} - CD_i)\Delta T/2 \quad (5)$$

#### Choice of $N_{\max}$

$N_{\max}$  is used in SAST logic to limit the maximum number of options to be evaluated in the signal optimization process (Fig. 3). If  $N_{\max} = 4$  is chosen, the signal optimization is limited to the evaluation of the options of terminating the current green at  $T + \Delta T$ ,  $T + 2\Delta T$ , and  $T + 4\Delta T$ , respectively. For the example depicted in Figure 4, this means the optimization process will be forced to terminate after the first five switching sequences are generated and evaluated. Of course, the optimization process may end as soon as the third switching sequence is generated and evaluated.

Given that  $M$  steps of advance information are available, SAST logic allows the evaluation of up to  $M$  options of terminating the current green. It is not necessary, however, to consider all the options. Simulation analyses (5) reveal that the efficiencies of SAST-based signal operations are not very sensitive to the choice of  $N_{\max}$ . With detectors installed at a distance of 400 ft upstream of the intersection,  $N_{\max} = 2$  is sufficient to achieve a high level of control efficiencies. With a longer detector setback of 600 ft, a larger  $N_{\max}$  may be desirable. Nevertheless,  $N_{\max}$  of 3 or 4 is sufficient in such a case.

#### A QUEUE-BASED CONTROL LOGIC

The queue-based logic analyzed in this study is depicted in Figure 5. This logic retains the same decision-making structure

as the SAST logic. The only difference is that the optimization process of the SAST logic (level IV) is replaced by a simple decision rule based on queue length. It should be noted that the level II decision making of this queue-based logic can be eliminated without affecting the resulting control efficiencies. This level of decision making, however, was not removed in this study when the relative merits of the SAST logic and the queue-based logic were evaluated.

The queue-based decision rule that replaces the optimization process of SAST logic involves a comparison of the following two values:

1.  $(L_i)_{\max}$  = expected maximum queue length of the current green phase at  $T + \Delta T$ , and
2.  $L_f$  = predetermined threshold queue length.

If the maximum queue length  $(L_i)_{\max}$  is less than or equal to the threshold value  $L_f$ , the current green phase is terminated at  $T + \Delta T$ . Otherwise, the green phase is allowed to extend beyond  $T + \Delta T$ , subject to a maximum green constraint.

In appearance, the queue-based control logic is as simple as the gap seeking logic of traffic-actuated control. They differ, however, in significant ways. Traffic-actuated control attempts to terminate the current green when the arriving vehicles can no longer utilize the green efficiently. Unfortunately, the gap-seeking logic of this mode of control can easily misjudge the actual flow conditions. The queue-based logic also attempts to terminate the current green when the arriving vehicles can no longer utilize the green efficiently. However, it requires the synthesis of detected vehicle arrivals to form a reasonably accurate picture of the conditions for decision making. Consequently, this decision-making process is more intelligent than the gap-seeking logic of traffic-actuated control. For general application, the queue-based logic also needs a more sophisticated vehicle detection system, because accurate estimation of queue lengths is not necessarily a simple problem.

The idea of using queue length as a control criterion is not new (9, 10). A conventional wisdom of queue-based control is to extend the current green until the vehicles in the governing queue are completely discharged into the intersection. In analyzing SAST logic, it was found that allowing queuing vehicles to dissipate completely before calling for optimization can be detrimental to the control efficiency (5). Therefore, it is necessary to choose a proper threshold queue length for implementing the queue-based logic.

#### ASSESSMENT OF CONTROL LOGICS

A microscopic, event-oriented simulation model was used to compare the performances of SAST logic and the queue-based logic. A major component of this model is a flow processor, which generates vehicle arrivals and processes vehicles downstream according to prevailing signal indications. Another major component is a signal processor, which allows various signal logics to be implemented for evaluation. Delays estimated from the model agree very well with the estimates obtained from Webster's formula (11) for pretimed operations (5). Simulated delays under traffic-actuated operations have also been compared with delays measured on six intersection

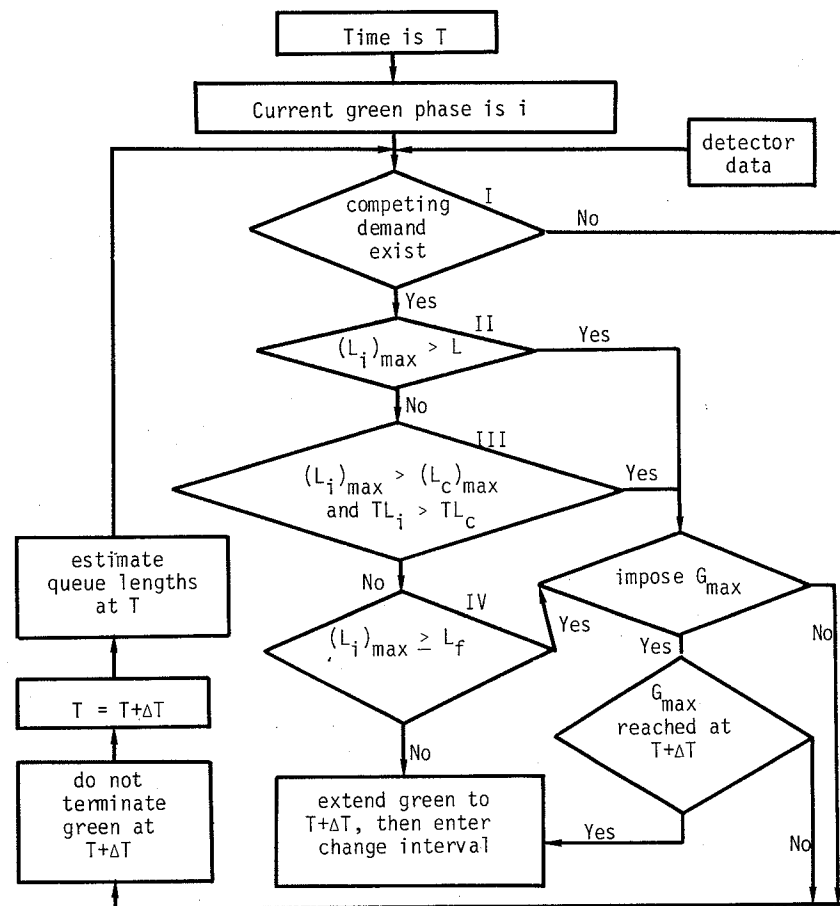


FIGURE 5 Queue-based control logic.

approaches (5). They differ by less than 3% when actual vehicle arrival sequences are used as inputs into the simulation model. The differences between the simulated and the measured delays can be greater if vehicle arrivals are generated from specified flow rates.

To provide an insight into the desirability of replacing traffic-actuated control with adaptive control, SAST logic was compared with conventional loop occupancy control logic. This comparison was based on various hourly flow patterns which were subjected to either two- to four-phase control. The number of lanes associated with a phase was varied from two to four. The lane flows ranged from 100 to 750 vph per lane. The total flows approaching the intersection were in the range of 600 to 5,600 vph. No conflicting movements were present. The four-phase control had two protected left-turn phases to accommodate vehicles in continuous left-turn lanes. The simulated vehicle had an average approach speed of 40 ft/sec.

For the loop occupancy control, the maximum allowable green was set at 60 sec. When two-phase operations were encountered, 50-ft detectors were used if a phase was associated with four lanes, and 70-ft detectors were used if a phase was associated with two lanes. For four-phase operations, 50-ft detectors were used in left-turn lanes, while 70-detectors were used in others. The extension interval was set at zero seconds for all the cases examined. These timing settings and detector configurations yield near optimal operations under heavy flow conditions.

For SAST-based operations, 5-ft detectors were placed 400

ft upstream of the intersection to detect vehicle arrivals. The step size  $\Delta T$  was set at 2 sec and  $N_{max}$  was limited to 2. The maximum green was set at 60 sec. This constraint, however, took effect only when a queuing vehicle was stopped by a red light. In addition, the current green was extended automatically if the maximum queue length of the current green phase exceeds four vehicles.

Referring to Figure 6, it can be seen that the advantages of the SAST based control over the loop occupancy control can vary from one hourly flow pattern to another. When the flow rates are low, and the delays under the loop occupancy control are less than 10 sec/veh for two-phase operations, the SAST-based control can be only as efficient as the loop occupancy control. For four-phase operations, the SAST-based control cannot be expected to deliver significant improvements when the delay under the loop occupancy control is less than 20 sec/veh. Under moderate to heavy flow conditions, however, the SAST-based control can improve the control efficiency in some cases by more than 20%; the most likely level of improvement appears to be in the range of 8% to 15%. Since traffic-actuated signals are not necessarily utilized to their best ability, the actual improvement through adaptive control can be greater than what is implied in Figure 6.

Traffic-actuated control based on loop occupancy or volume density logic can be very efficient under light flow conditions. Under heavier flow conditions, two problems may emerge. One problem is the failure of the control to allow

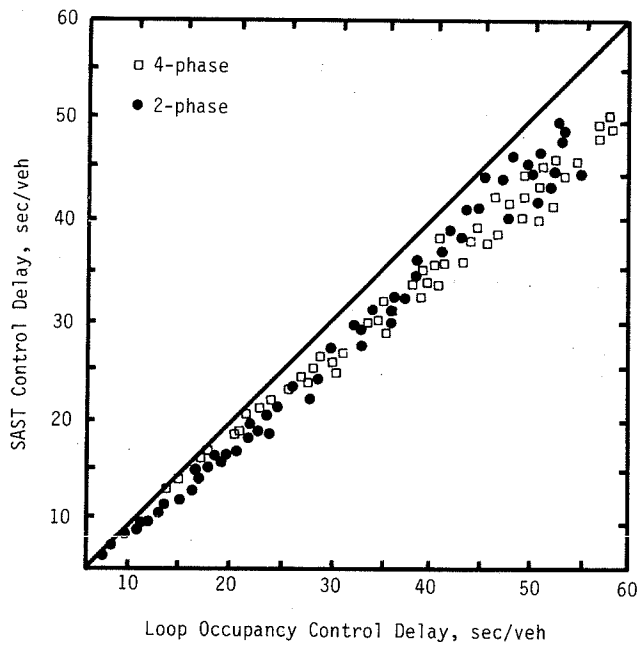


FIGURE 6 Delays of SAST control versus delays of loop occupancy control.

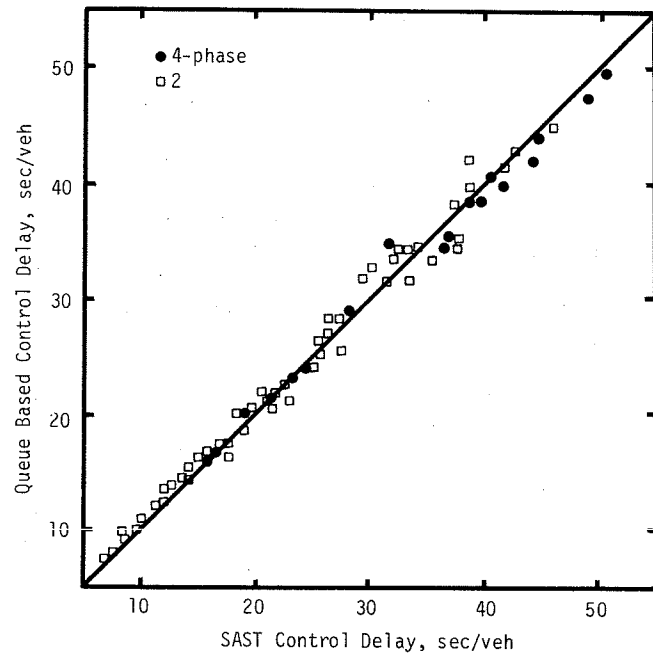


FIGURE 7 Delays of queue-based control versus delays of SAST control (detector setback = 400 ft).

most queuing vehicles to enter the intersection due to premature termination of the green. This problem can arise when short detector lengths, short vehicle intervals, or short maximum greens are employed. On the other hand, excessively long greens may result because of the actuation of detectors by vehicles not in a queue. Such vehicles cannot utilize the intersection capacity as efficiently as queuing vehicles. This problem can become rather acute when many lanes are associated with a signal phase.

Adaptive control can alleviate these weaknesses of traffic-actuated control through real-time optimization. Similarly, the queue-based logic, as shown in Figure 5, can also prohibit vehicles that cannot efficiently utilize the intersection capacity from extending the green. To facilitate a comparative analysis of SAST logic and the queue-based logic, the same simulation model was used to determine the best threshold queue length  $L_f$  that should be used for the level IV decision-making of the queue-based logic. The best threshold queue length was found to be in the range of 0.5 to 1.5 vehicles.

Based on a threshold value of  $L_f = 1.5$  vehicles, the delays produced respectively by SAST logic and the queue-based logic for a number of hourly flow patterns were estimated through simulation. The results are shown in Figure 7, where it can be seen that, for hourly flow patterns that have delays under 20 sec per vehicle, the SAST-based control is slightly better than the queue-based control. Under heavier flow conditions, however, the queue-based control can sometimes perform better than the SAST-based control.

The ability of the queue-based logic to deliver reasonably high control efficiencies is not without a logical basis. Nevertheless, one reason that the queue-based control can sometimes deliver better signal operations than the SAST-based control can be found in the use of  $N_{\max} = 2$  and a threshold queue length of  $L = 4$  vehicles for the SAST-based control. This combination of  $N_{\max}$  and  $L$  is not necessarily the best

for all the hourly patterns tested in this study. At the present time, however, it is unknown which combination of  $N_{\max}$  and  $L$  will result in the best overall operation for an intersection with a wide range of flow conditions. Despite this limitation, it should be noted that, with the exception of a few tested flow patterns that have low flow rates, the queue-based control consistently performs better than the loop occupancy control. This characteristic can be exploited for adaptive control of intersections where short auxiliary lanes, opposed left turns, or frequent right-turn-on-red maneuvers exist. At such intersections, reliable advance information cannot be obtained for all approach lanes. In these cases, real-time information on queuing flows may be obtained and used to complement advance information in order to produce efficient signal operations.

## CONCLUSIONS

Adaptive control has the potential to improve the existing level of signal control efficiencies at isolated intersections. Generally, adaptive control requires a logic to identify flow conditions and to use the identified conditions for making intelligent timing decisions. This process of control usually results in the use of estimated flow conditions for making signal timing decisions. Estimated flow conditions always deviate from actual conditions. The detrimental effects of flawed information on signal timing decisions cannot be compensated for by the use of a strenuous process of searching for better signal operations. Therefore, it is pertinent to examine whether simple decision rules can be effectively used to replace a more strenuous decision-making process for adaptive control.

In comparison with conventional loop occupancy control under simulated conditions, the SAST logic, which uses advance information in a vigorous process for making signal timing decisions, can provide significantly better signal operations.

The level of improvement varies with a number of factors, but it tends to be higher when heavier flows are encountered. Under the same simulated conditions, the simpler queue-based logic can produce comparable results. This implies that real-time information on queuing flow can be used to produce improved signal operations. This understanding is important in the development of a versatile adaptive control logic.

With the possible exception of the modernized optimized vehicle actuated strategy (I), none of the adaptive control strategies mentioned above can be effectively utilized for the control of intersections where short turning bays, opposed left turns, or right-turn-on-red maneuvers exist. At such intersections, reliable advance information cannot be obtained for all traffic movements. Under the circumstances, it would be logical to use real time information on queuing flow, as well as other advance information, for decision making. A major challenge to facilitate such a use of information is to develop a vehicle detector system that can provide reliable real-time and advance information at all or most intersections controlled currently with traffic-actuated signals.

#### ACKNOWLEDGMENT

This study is sponsored in part by the National Science Foundation and in part by the UPS Foundation.

#### REFERENCES

1. R. A. Vincent and C. P. Young. Self-Optimizing Traffic Signal Control Using Microprocessors—the TRRL 'MOVA' Strategy

- for Isolated Intersections. *Traffic Engineering and Control*, Vol. 27, No. 7/8, 1986, pp. 385–387.
2. A. J. Miller. A Computer Control System for Traffic Network. *Proc., 2nd International Symposium on Theory of Road Traffic Flow*, London, pp. 201–220.
3. N. H. Gartner. OPAC: A Demand-Responsive Strategy for Traffic Signal Control. In *Transportation Research Record 906*, TRB, National Research Council, Washington, D.C., 1983, pp. 75–81.
4. K.-L. Bang. Optimal Control of Isolated Traffic Signals. *Traffic Engineering and Control*, July 1976, pp. 288–292.
5. F. B. Lin, N. Wang, and S. Vijayakumar. Development of an Intelligent Adaptive Signal Control Logic. Presented at the Engineering Foundation Conference on Management and Control of Urban Traffic, Henniker, N.H., June 14–19, 1987.
6. P. B. Hunt, D. I. Robertson, R. D. Bretherton, and M. C. Royle. The SCOOT On-Line-Traffic Signal Optimization Technique. *Traffic Engineering and Control*, Vol. 23, No. 4, 1982, pp. 190–192.
7. P. Carden and M. McDonald. The Application of SCOOT Control to an Isolated Intersection. *Traffic Engineering and Control*, Vol. 2, No. 6, 1985, pp. 304–310.
8. L. de la Breteque and R. Jezequel. Adaptive Control at an Isolated Intersection—A Comparative Study of Some Algorithms. *Traffic Engineering and Control*, Vol. 20, No. 7, 1979, pp. 361–363.
9. Caltrans. Diamond Interchange Program, Software Documentation. Implementation Package, FHWA, U.S. Department of Transportation, IP-80-5. 1980.
10. D. O. Gerlough and F. A. Wagner. *NCHRP Report 32: Improved Criteria for Traffic Signals at Individual Intersections*. HRB, National Research Council, Washington, D.C., 1967.
11. F. V. Webster. Traffic Signal Settings. Road Research Technical Paper No. 39. Road Research Laboratory, 1958.

---

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

# New Algorithm for Solving the Maximum Progression Bandwidth

HUEL-SHENG TSAY AND LIANG-TAY LIN

Two popular computer programs, MAXBAND and PASSER II, are widely used in obtaining the maximal bandwidth. However, these bandwidths may not be realized or only be partly realized if the resultant signal timings are actually applied on the arterial. This phenomenon can be observed from field tests or from a time-space diagram. In this paper two examples demonstrate the problem. A new algorithm is proposed for solving the bandwidth problem and provides the user with a more realistic maximum progression bandwidth. The algorithm uses a general mixed-integer programming formulation, and a program BANDTOP based on this formulation has been developed to obtain the real progression bandwidth. It has been tested on street networks in Taiwan, where it has proved very effective. The major variation from traditional methods is that the bandwidth has a saw-toothed pattern in both directions instead of parallel and uniform. Any vehicle in the segment is allowed to travel through the entire section of an arterial with at most one stop.

The coordination of traffic signals on the arterial is an effective way of reducing stops, delay, and excessive fuel consumption. Previously, signal settings were determined from the time-space relation of signal timing and traffic flow by manual methods. As researchers begin to use computers to increase analysis flexibility and reduce computational effort, it becomes possible to develop new approaches that take into account more variables and complex equations involved in reflecting the real situation. The United States and many other nations use Maximal Bandwidth (MAXBAND) and Progression Analysis and Signal System Evaluation Routine (PASSER II), which are based on maximizing two-way traffic through bandwidth. MAXBAND and PASSER II can automatically take traffic demands to determine two-way progression bandwidths and to provide users with other information, such as cycle length, phase sequence, offsets, phase lengths, and a time-space diagram for practical use. Here, the bandwidth is defined as the vehicles within a time interval, present at a given traffic signal or point, that can travel through downstream signals of an arterial without stopping.

The progression bandwidth of MAXBAND was mainly calculated from Little's general mixed-integer linear programming formulation (1-5). It obtains a global optimum of bandwidth, cycle length, offsets, and phase sequence with no starting solution. MAXBAND also has the capability to allow small deviations from the arterialwide progression speed on individual links (6). PASSER II is a macroscopic, deterministic program that obtains the optimal timing set-

tings from maximizing progression bandwidth in both directions. It was developed by Messer through Little's half-integer synchronization and expansion of Brook's algorithm by selecting the offsets that minimize the total interference to the progression band (2, 7-10). The newest version of PASSER II-84 can analyze the phase sequence of any NEMA style from two to eight phases and find minimum delay through fine-tuning of the offsets while allowing the maximization of bandwidth to dominate (11). The heuristic optimization technique used in PASSER II does not produce the widest possible green bands, unlike MAXBAND, which guarantees a global optimum (12).

Since both approaches have impressive bandwidths through time-space diagrams, a substantial number of practicing engineers may use the output as the arterial signal timing settings. Unfortunately, these bandwidths actually will not be realized or only be realized in fraction. The phenomena can be observed from the field or time-space diagrams. One could argue that many fairly restrictive hypotheses related to these bandwidth approaches exist. The assumptions include a uniform platoon, no platoon dispersion, low volumes, and no or few vehicles entering the arterial from side streets; but situations corresponding to these assumptions are rare and unreasonable.

Many traffic engineers prefer a maximization of synchronized green phases using time-space diagrams to satisfy the public's demand. Since the assumptions made in the two programs are unrealistic, the use of MAXBAND and PASSER II output on an arterial may result in unexpected stops, delay, and even more fuel consumption to the entire system. It is necessary therefore to develop a new algorithm for solving the maximum progression bandwidth that allows the driver to travel at the design speed without any stop. In this paper, as the first step, two examples define the existing problem of bandwidths obtained from MAXBAND and PASSER II. Then, a new algorithm is developed to provide users a real maximal bandwidth without interference. A complete mixed-integer programming formulation of the new algorithm is proposed and discussed in detail. Finally, the new algorithm is tested on street networks and proved effective.

## THE EXISTING PROBLEM

When the lights are red, queues build up as a result of turning movements into the arterial at the previous intersection before the appearance of green. The queue includes not only turning vehicles from the previous intersection but also the vehicles that do not pass through the arterial at the end of the last

green time. The phenomena are quite obvious and should not be neglected at any intersection during the entire day.

Although the assumptions of no or very few flow left on the arterial and entering the arterial from side streets were made by MAXBAND and PASSER II, the two programs still allow the user to specify a queue clearance time at any intersection in either direction. A queue clearance time can either be a fraction of the cycle length or actual time units to deal with queues (1, 10). The program then adjusts the through vehicles to arrive at the intersections after the queue has cleared and leave the intersections with the queue as a part of the band. This puts a jog into the through band, advancing it upon leaving the intersection by an amount equal to the queue clearance time (1). In other words, MAXBAND and PASSER programs admit the existence of queue at each intersection and try to use the queue clearance time to handle this unavoidable situation. If a queue clearance time is being considered at each intersection of the arterial, however, its maximal bandwidth will be severely affected and sometimes reduced to a very small value. In addition, since queue clearance time is an arbitrary number specified by the user, it is difficult to provide the user with guidelines for setting a reasonable value at a particular intersection.

Figures 1 and 2 show the time-space diagrams of PASSER II-84 and MAXBAND for Zin-Wha Arterial with four intersections in Tainan City, Taiwan. Both programs were run on

the same information of arterial configuration and traffic flows as input. The major difference between the two figures lies in the dot points in PASSER II that represent signal green time but indicate red time in MAXBAND and vice versa. From Figures 1 and 2, it can be seen that, in the outbound direction, as the light of intersection 2 (MING-CHEN) turns green, the queueing vehicles at this intersection will move toward the adjacent downstream intersection [i.e., intersection 3 (MING-SEN)] and have to wait at the red light at this intersection. The newly arriving vehicles then join the existing queue to form a new composite queue at intersection 3.

Based on the first-come first-served principle, the vehicles involved in the composite queue have to use the front portion of the next green time. It equals the time needed to depart the total queue at saturation flow rates as the signal turns green. Hence, the incoming through-band vehicles cannot cross intersection 3 unless all queues have cleared. Under such a circumstance, most of the vehicles in the through band are hindered and have to stop. This phenomenon can also be observed from the trajectories of several leading vehicles at intersection 3 in Figures 1 and 2. Here, any intersection that has the bandwidth located in the very front of green time but with a different band location of green time at an adjacent upstream intersection is the critical intersection. The bandwidth of MAXBAND and PASSER II will be affected or decreased at each critical intersection.

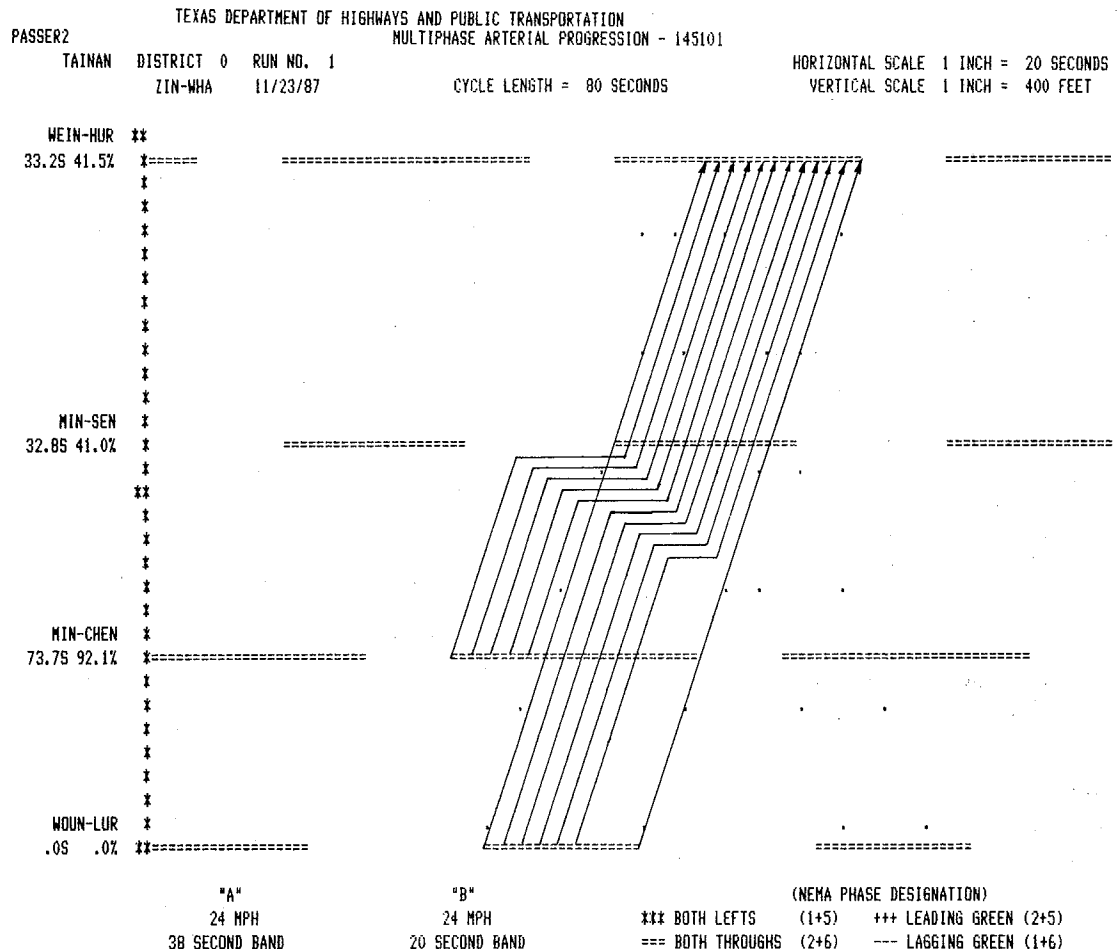


FIGURE 1 PASSER II-84 time-space diagrams of Zin-Wha arterial in Tainan, Taiwan.

\*\*\* MAXBAND TIME-SPACE DIAGRAM \*\*\*

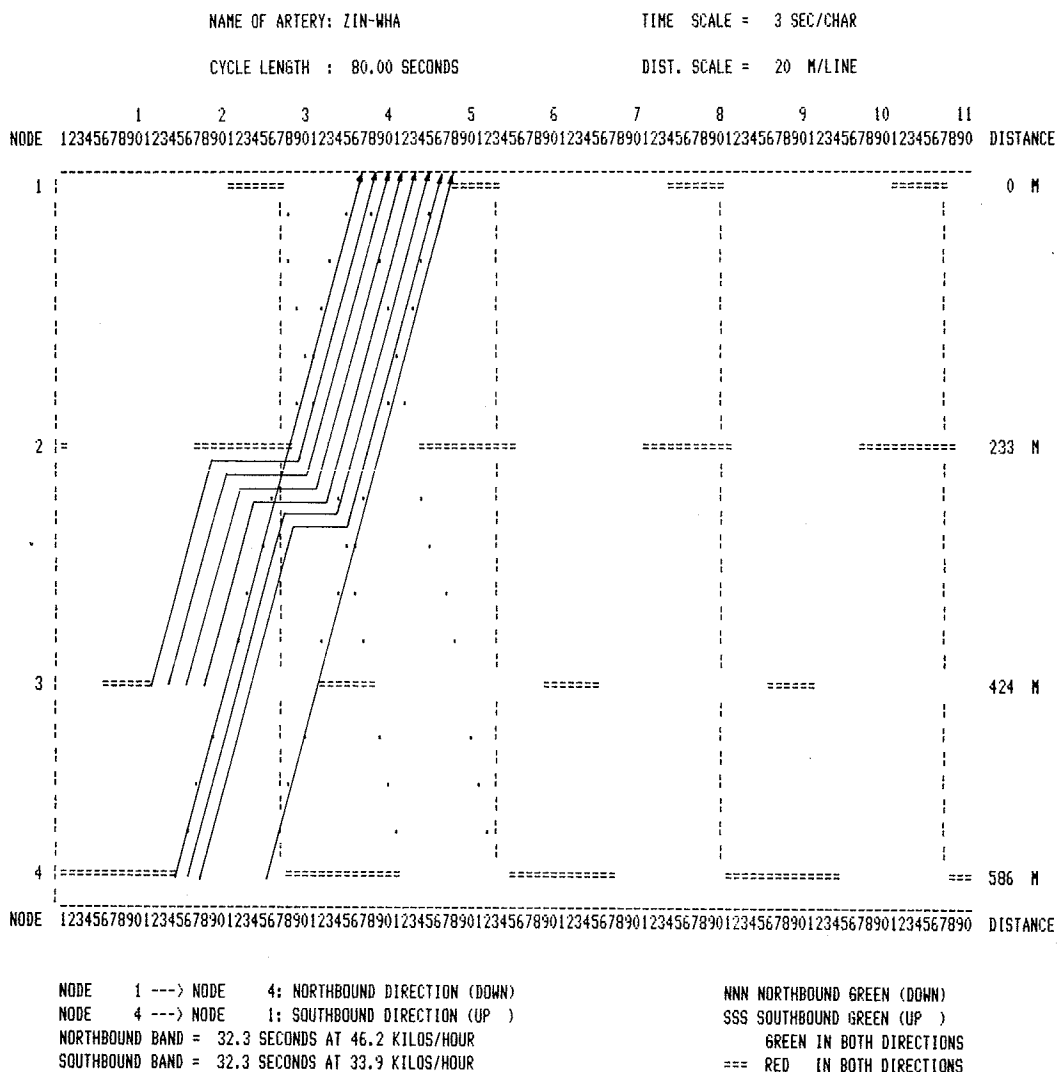


FIGURE 2 MAXBAND time-space diagrams of Zin-Wha arterial in Tainan, Taiwan.

One may consider using queue clearance time to avoid the problem encountered at the critical intersection. However, because queue clearance time set in MAXBAND and PASSER II attempts to clear the queue due to turning movements, it cannot handle the composite queue involving the existing queue and incoming upstream vehicles. In addition, the user has to specify queue clearance time at any intersection. It is impossible to know which intersection needs queue clearance time and what value should be used. Even if the value is assumed, the critical intersection will soon be changed to another intersection based on the output of MAXBAND and PASSER II. The existing problem still remains unsolved.

At this point, several questions arise concerning the critical intersection, for example, intersection 3 in Figures 1 and 2. How much time is needed to clear the composite queue? Can we prespecify the queue clearance time to prevent this phenomenon? How many seconds of the maximum bandwidth from the two programs will not be used due to the composite queue? Will this value just be equal to the time required to relieve the queue? As a result, how many seconds of band-

widths that vehicles can travel from the critical intersection till the last intersection of the designated arterial without stopping will be realized?

Three outputs of MAXBAND and PASSER II have been tested on three arterials in two cities in Taiwan, Tainan and Taipei, under various flow conditions. The results showed that these bandwidths could not be realized and their signal timing plans gave more stops and delay than the original one. Traffic engineers in both cities could not explain the reason. One claimed it was probably due to different driver behavior and cultural background as programs developed in the United States may not be suitable in other countries. In fact, the problems mainly come from inaccurate progression theory because of unrealistic assumptions involved in the two programs.

According to several tests in the field which utilized the bandwidths obtained from PASSER II and MAXBAND, at least double the time needed to clear the composite queue will not be available at the critical intersection. That is to say, if the width of green band minus double the composite queue

clearance time is less than zero, the progression band will not be available under given conditions. For example, in Figure 1, the outbound bandwidth (38 sec) will not be realized if the composite queue requires more than 19 sec to clear at intersection 3. This is the serious drawback of current MAXBAND and PASSER in practical applications; therefore, it becomes necessary to develop an algorithm that can obtain the real maximum bandwidth. One way to perform this study is to discuss the progression theory used in MAXBAND as the first step. The following section serves this purpose.

**MAXBAND FORMULATION**

The time-space diagram of MAXBAND showing green bands was presented by Little et al. (1) and is shown in Figure 3. Inbound and outbound green bands pass through signals  $S_h$  and  $S_i$ . Quantities with bars refer to inbound reds, are drawn solid, and above inbound reds need not coincide (1). The definitions of variables shown in Figure 3 are as follows:

- $b$  = outbound bandwidth,
- $S_i$  =  $i$ th signal ( $i = 1, \dots, n$ ),
- $r$  = outbound red time at  $S_i$ ,
- $W_i$  = time from right side of red at  $S_i$  to left edge of outbound green band,
- $t(h, j)$  = outbound travel time from  $S_h$  to  $S_i$ ,
- $\phi(h, i)$  = time from center of an inbound red at  $S_h$  to the center of a particular outbound red at  $S_i$ ,
- $\Delta_i$  = time from center of  $r_i$  to nearest center of  $r_i$ , and
- $\tau_i$  = queue clearance time.

A general mathematical programming formulation of MAXBAND given by Little et al. (1) is presented in the following. All variables and symbols are based on Figure 3 except that signal  $h$  is replaced by symbol  $i$  and signal  $i$  is

substituted by  $i + 1$ . It is defined as  $x = x(i, i + 1)$ , for  $x = t, \bar{t}, m, \phi, \bar{\phi}$ .

$$\text{Max } b + k\bar{b}$$

$$ST(1-k)\bar{b} \geq (1-k)kb$$

$$1/T_2 \leq Z \leq 1/T_1$$

$$W_i + b \leq 1 - r_i \quad i = 1, \dots, n$$

$$\bar{W}_i + \bar{b} \leq 1 - \bar{r}_i \quad i = 1, \dots, n$$

$$(W_i + \bar{W}_i) - (W_{i+1} + \bar{W}_{i+1}) + (t_i + \bar{t}_i) + \delta_i l_i - \bar{\delta}_i \bar{l}_i - \delta_{i+1} \bar{l}_{i+1} + \bar{\delta}_{i+1} l_{i+1} - m_i = (r_{i+1} - r_i) + (\bar{r}_i + \bar{r}_{i+1}), \quad i = 1, \dots, n-1$$

$$(d_i/f_i)Z \leq t_i \leq (d_i/e_i)Z, \quad i = 1, \dots, n-1$$

$$(\bar{d}_i/\bar{f}_i)\bar{Z} \leq \bar{t}_i \leq (\bar{d}_i/\bar{e}_i)\bar{Z}, \quad i = 1, \dots, n-1$$

$$(d_i/h_i)Z \leq (d_i/d_{i+1})t_{i+1}$$

$$-t_i \leq (d_i/g_i)Z, \quad i = 1, \dots, n-2$$

$$(\bar{d}_i/\bar{h}_i)\bar{Z} \leq (\bar{d}_i/\bar{d}_{i+1})\bar{t}_{i+1}$$

$$-\bar{t}_i \leq (\bar{d}_i/\bar{g}_i)\bar{Z}, \quad i = 1, \dots, n-2$$

$$b, \bar{b}, Z, W_i, \bar{W}_i, t_i, \bar{t}_i \geq 0$$

$$m_i = \text{integer}$$

$$\delta_i, \bar{\delta}_i = 0, 1$$

where

- $K$  = target ratio of inbound to outbound bandwidth;
- $T$  = cycle length;
- $Z = 1/T =$  signal frequency;

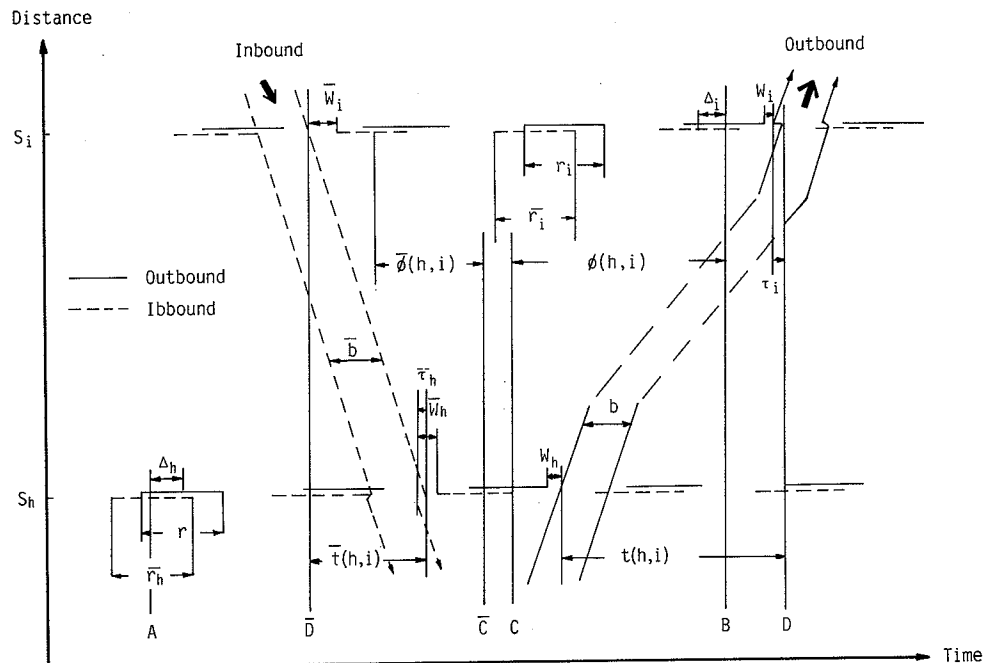


FIGURE 3 Time-space diagrams of MAXBAND showing green bands (1).



- $T_1, T_2$  = lower and upper limits on cycle length (i.e.,  $T_1 \leq T \leq T_2$ );  
 $d(h, i)$  = distance between  $S_h$  and  $S_i$  outbound;  
 $d_i = d(i, i+1)$  = distance between outbound intersections  $i$  and  $i+1$ ;  
 $e_i, f_i$  = lower and upper limits on outbound speed;  
 $1/h_i, 1/g_i$  = lower and upper limits on change in outbound reciprocal speed; and  
 $t_i$  = travel time from outbound intersection  $i$  to intersection  $i+1 = (d_i/V_i)Z$ ,  $V_i$  = travel speed.

This formulation has the following deficiencies:

1. The progression band cannot be fully realized or can only be partly realized.
2. Queue clearance time is prespecified by the user. In fact, it cannot be a fixed value and will be varied with the queue length of each intersection. This value should be determined internally through the computation of traffic flow movements.
3. The traffic flow model is oversimplified. No account is taken of secondary flows turning from side streets and platoon dispersion (6).
4. The time lag between the remaining portion of green time after the band and start of red time and the time difference between the beginning of green time and of bandwidth in either direction are not clearly distinguished. Only a variable  $W_i$  is used to represent this time difference and may cause confusion.
5. A symbol error exists in the inbound speed change of the above MAXBAND formulation. For consistency, the equation should be changed to the following form:

$$(\bar{d}_{i+1}/\bar{h}_{i+1})Z \leq (\bar{d}_{i+1}/\bar{d}_i)\bar{t}_i - \bar{t}_{i+1} \leq (\bar{d}_{i+1}/\bar{g}_{i+1})Z$$

## MATHEMATICAL FORMULATION OF NEW ALGORITHM

Based on the above discussion, a new algorithm for obtaining the real maximal bandwidth will be developed in this section. To explain the new algorithm more easily, similar notations and definitions considered in the MAXBAND formulation are used. The following described variables refer to the above section unless otherwise specified. Major features of this new algorithm compared to the MAXBAND formulation are:

1. Divide the time between the start of green time and of the bandwidth into two parts: queue clearance time ( $Q_i$ ) and incoming flow clearance time ( $H_i$ ) at each intersection. Here, the queue clearance time is used to clear queues due to turning vehicles during red time and through vehicles that do not go through the arterial at the end of the last green time. Incoming flow clearance time represents the time needed by the incoming vehicles that come from the adjacent upstream intersection but excludes vehicles in the through band, to depart the upstream intersection.
2. Specify the time lag from the right side of the bandwidth to the left edge of outbound red as  $W_i$  at intersection  $i$ . It is noted that this new definition of  $W_i$  is different from the one used in the MAXBAND.
3. Add a composite queue clearance time constraint.
4. Add a constraint that guarantees the progression bandwidth to be fully used by vehicles without stopping.

5. Add the minimum green time of side streets.
6. Provide the selection of eight left-turn phase patterns.
7. Find the minimum cycle time by making minor changes of the objective function and constraints under given bandwidths in both directions.

A time-space diagram of the new algorithm concerning green bands is given in Figure 4. All variables involved and equations derived later are based on the relationships shown in this figure. Any variable with a bar represents the inbound flow. Otherwise, it indicates the outbound flow. Since Little et al. have provided detailed information to derive some equations (1), similar derivations of these equations are omitted and only the final equations with the newly added variables will be shown.

### Objective Function

$$\text{Max } b + \bar{b} \quad (1)$$

### Constraints

#### Geometric Relationship

$$t_i + \bar{t}_i + (1/2)(r_i + \bar{r}_i) - (1/2)(r_{i+1} + \bar{r}_{i+1}) + (Q_i - Q_{i+1}) + (H_i - H_{i+1}) + (\bar{W}_i - \bar{W}_{i+1}) + \Delta_i - \Delta_{i+1} = I_i \quad (2)$$

where  $I_i$  is an integer.

#### Offsets

$$\text{OFF}_i = t_i + (Q_i - Q_{i+1}) + (H_i - H_{i+1}) \quad (3)$$

$$\bar{\text{OFF}}_i = \bar{t}_i + (\bar{Q}_{i+1} - \bar{Q}_i) + (\bar{H}_{i+1} - \bar{H}_i) \quad (4)$$

#### Common Cycle

In the coordinated signal intersections of an arterial, every intersection within the segment has the same cycle length. Therefore,

$$Q_i + H_i + b + W_i + r_i = 1 \quad (5)$$

$$\bar{Q}_i + \bar{H}_i + \bar{b} + \bar{W}_i + \bar{r}_i = 1 \quad (6)$$

#### Bandwidth

To guarantee a real bandwidth, the following equations have to be added as bandwidth constraints:

$$H_{i+1} \geq H_i + Q_i \quad (7)$$

$$\bar{H}_i \geq \bar{H}_{i+1} + \bar{Q}_{i+1} \quad (8)$$

From equations (7) and (8), the final shape of the progression bandwidth will be a saw-toothed pattern.

#### Queue Clearance Time

Before discussing the queue length, it is necessary to explain the arrival types of incoming vehicles from the adjacent

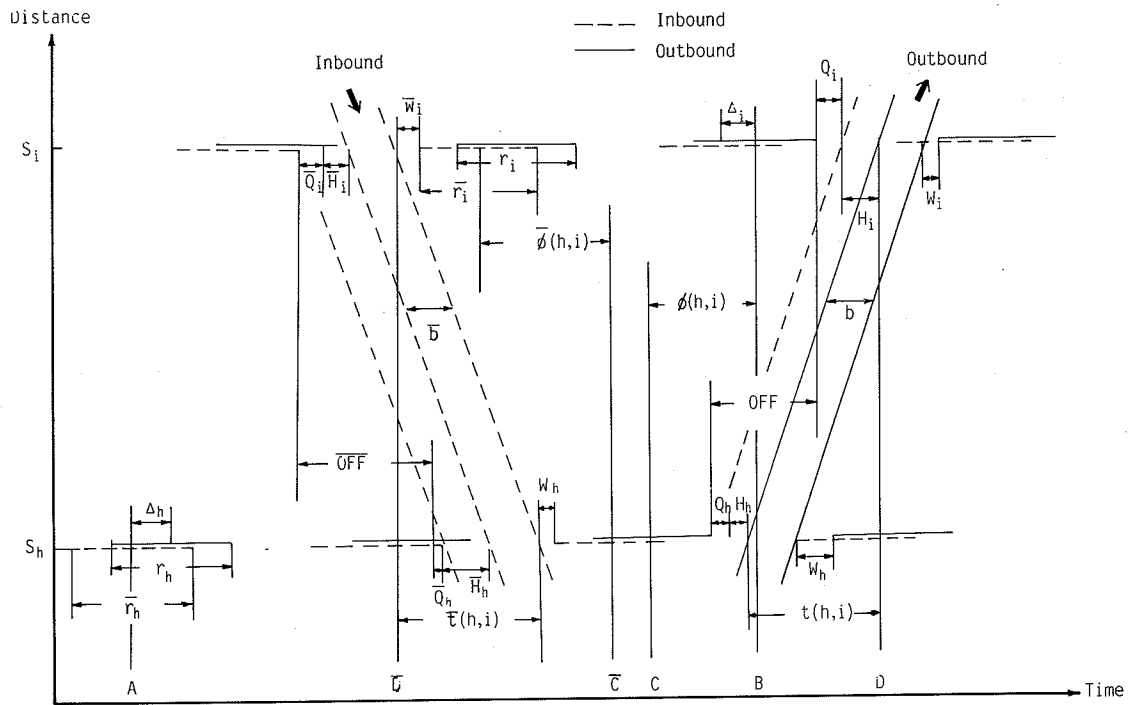


FIGURE 4 Time-space diagrams of new algorithm.

upstream intersection. The vehicle arrival type depends on through traffic volume, turning vehicles, and timing plan at the upstream intersection. Two types of arriving vehicles exist. Vehicles departing from the upstream intersection during green belong to type 1. Similarly, vehicles leaving the upstream intersection during red are referred to as type 2.

For outbound flow, the arrival rate of type 1 from upstream intersection  $i$  to intersection  $i+1$  is

$$\lambda_{i+1,1} = V_{i,T} / [(g_i/C) \times 3600 \times N_{i+1}] \quad (9)$$

where

- $\lambda_{i+1,1}$  = arrival rate of type 1 at intersection,
- $V_{i,T}$  = through traffic volume at intersection  $i$ ,
- $g_i$  = green time of intersection  $i$ ,
- $N_{i+1}$  = number of lanes at intersection  $i+1$ , and
- $C$  = cycle time.

The arrival rate of type 2 at the intersection  $i+1$  is

$$\lambda_{i+1,2} = (V_{i,R} + V_{i,L}) / [(r_i/C) \times 3600 \times N_{i+1}] \quad (10)$$

where

- $\lambda_{i+1,2}$  = arrival rate of type 2 at intersection  $i+1$ ,
- $V_{i,R}$  = right turn volume at intersection  $i$ ,
- $V_{i,L}$  = left turn volume in the opposite approach at intersection  $i$ , and
- $r_i$  = red time interval at intersection  $i$ .

Because the model has to satisfy the bandwidth constraint given in equations (7) and (8), two cases can be drawn to show the relationships of any two neighboring intersections.

Case 1:  $W_{i+1} \geq W_i$ . The time lag from the right side of the bandwidth to the left edge of outbound red at an intersection is greater than or equal to that of the adjacent upstream intersection. This can be displayed in Figure 5(a). The queuing

vehicles  $QV_{i+1,1}$  of intersection  $i+1$  in the outbound direction can be obtained from the following equation:

$$QV_{i+1,1} = \lambda_{i+1,2} \times [r_i - (W_{i+1} - W_i)] \times C \quad \text{if } W_{i+1} \geq W_i \quad (11)$$

Case 2:  $W_{i+1} \leq W_i$ . This case is shown in Figure 5(b). Queuing vehicles at intersection  $i+1$  are

$$QV_{i+1,2} = [\lambda_{i+1,1} \times (W_i - W_{i+1}) + \lambda_{i+1,2} \times r_i] \times C \quad \text{if } W_{i+1} < W_i \quad (12)$$

where  $QV_{i+1,2}$  is queuing vehicles at intersection  $i+1$  under case 2.

After obtaining queuing vehicles and assuming saturation flow rates, the queue clearance time of outbound flow becomes

$$Q_{i+1} \geq SH_{i+1} \times \lambda_{i+1,2} \times [r_i - (W_{i+1} - W_i)] \quad \text{if } W_{i+1} \geq W_i \quad (13)$$

or

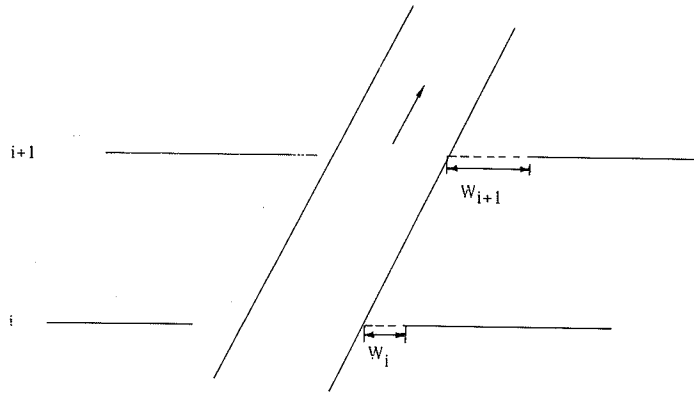
$$Q_{i+1} \geq SH_{i+1} \times [\lambda_{i+1,1} \times (W_i - W_{i+1}) + \lambda_{i+1,2} \times r_i] \quad \text{if } W_{i+1} < W_i \quad (14)$$

where  $SH_{i+1}$  is saturation flow headway for outbound time flow at intersection  $i+1$ . Similarly, the queue clearance time of inbound flow is

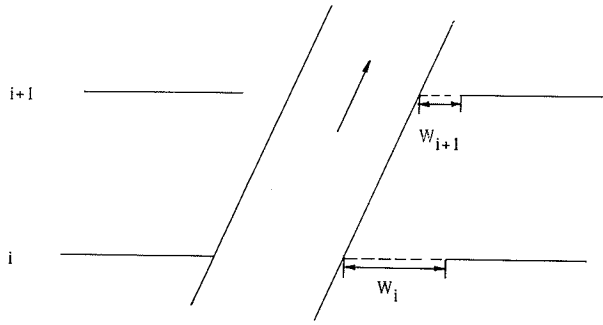
$$\bar{Q}_i \geq \bar{S}H_i \times \bar{\lambda}_{i,2} \times [\bar{r}_{i+1} - (\bar{W}_i - \bar{W}_{i+1})] \quad \text{if } \bar{W}_{i+1} \leq \bar{W}_i \quad (15)$$

or

$$\bar{Q}_i \geq \bar{S}H_i \times [\bar{\lambda}_{i,1} \times (\bar{W}_{i+1} - \bar{W}_i) + \bar{\lambda}_{i,2} \times \bar{r}_{i+1}] \quad \text{if } \bar{W}_{i+1} > \bar{W}_i \quad (16)$$



(a) Case 1:  $W_{i+1} \geq W_i$



(b) Case 2:  $W_{i+1} < W_i$

**FIGURE 5** Two cases of queuing vehicles.

Because most of the green time of downstream intersection  $i+1$  will be used by the queue clearance time and incoming flow clearance time, it is recommended equations (14) and (16) be used as the queue clearance time of the outbound flow and inbound flow, respectively.

#### Bandwidth of Each Intersection

The bandwidth constraint shown by equations (7) and (8) gives a real bandwidth for vehicles to travel through all downstream intersections of an arterial without interference. This new algorithm also provides additional progression opportunities at intersections in both directions except the through bandwidth. For example, vehicles move before the left edge of through band during green time and can arrive at the last intersection without stopping for the outbound direction. This is defined as the bandwidth of each intersection; the bandwidth is saw-toothed. The bandwidths of intersection  $i$  in either direction are represented in the following two equations:

$$b_i = b + Q_i + H_i \quad (17)$$

$$\bar{b}_i = \bar{b} + \bar{Q}_i + \bar{H}_i \quad (18)$$

From equations (17) and (18), it can be determined that

vehicles outside the through band need to stop at most once to pass through the entire arterial.

#### Directional Bandwidth Weight

We can set up weights for different directions:

$$\bar{b} = K b \quad (19)$$

where  $K$  is a relative weight ratio between inbound and outbound bandwidths.

If  $K$  is greater than 1, the inbound bandwidth is wider than the outbound one.

#### Minimum Green Time

This constraint guarantees that each side street has a minimum green time to prevent overdelay of vehicles from the side street and give pedestrians enough time to cross the arterial safely.

$$\frac{1}{2}(r_i + \bar{r}_i) - \Delta_i \geq MIG_i$$

$$r_i \geq MIG_i$$

$$\bar{r}_i \geq MIG_i \quad (20)$$

where  $MIG_i$  is minimum green on side street at intersection  $i$ .

*Cycle Limit*

$$1/C_2 \leq Z \leq 1/C_1 \tag{21}$$

*Speed Limit*

For outbound flow:

$$(d_i / f_i)Z \leq t_i \leq (d_i / m_i)Z \tag{22}$$

For inbound flow:

$$(\bar{d}_i / \bar{f}_i)Z \leq \bar{t}_i \leq (\bar{d}_i / \bar{m}_i)Z \tag{23}$$

where  $m_i, f_i$  are lower and upper limits of outbound speed.

*Speed Change Limit*

For outbound flow:

$$(d_i / h_i)Z \leq (d_i / d_{i+1})t_{i+1} - t_i \leq (d_i / n_i)Z \tag{24}$$

For inbound flow:

$$(\bar{d}_{i+1} / \bar{h}_{i+1})Z \leq (\bar{d}_{i+1} / \bar{d}_i)\bar{t}_i - \bar{t}_{i+1} \leq (\bar{d}_{i+1} / \bar{n}_{i+1})Z \tag{25}$$

where  $1/h_i, 1/n_i$  are lower and upper limits on change in outbound reciprocal speed.

*Left-Turn Phase*

Eight possible patterns of left-turn green phases exist in this new algorithm. The eight left-turn phase patterns are shown in Figure 6. Let

$g_i$  = outbound green time for through traffic,

$l_i$  = time allocated for outbound left-turn green at intersection  $i$ , and

$R$  = common red time in both directions to provide for side street movements.

This gives

$$r_i = R + \bar{l}_i$$

$$\bar{r}_i = R + l_i$$

$$r_i + g_i = 1$$

$$\bar{r}_i + \bar{g}_i = 1 \tag{26}$$

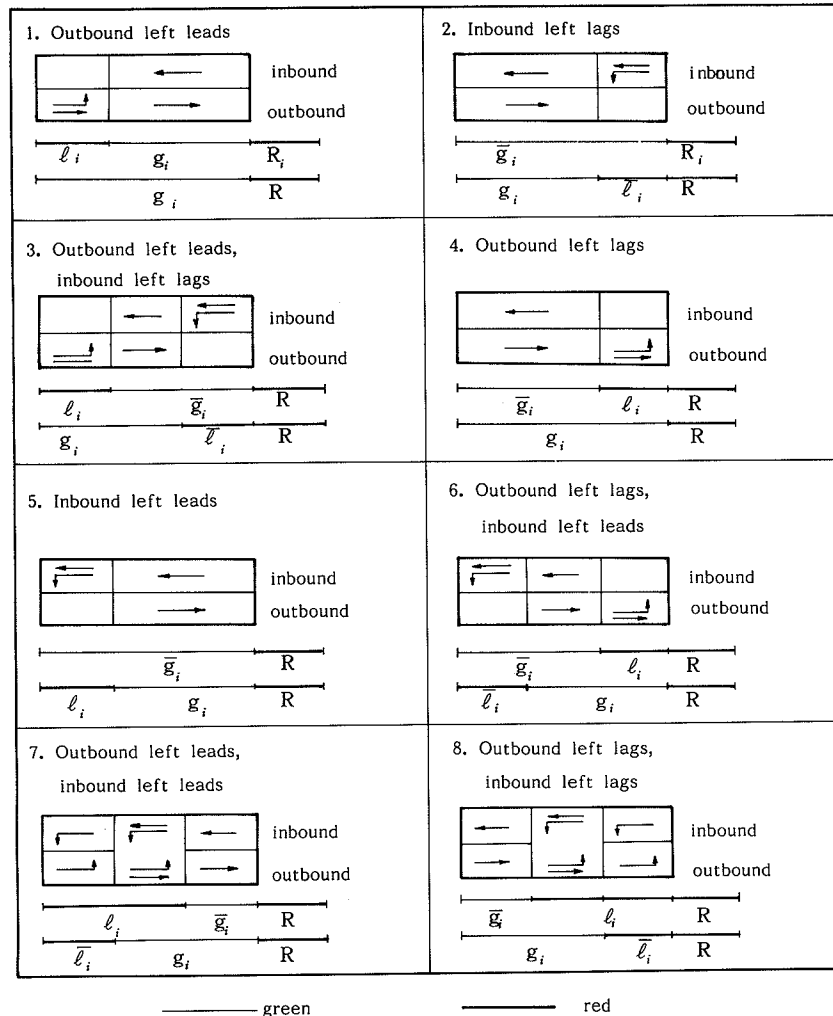


FIGURE 6 Eight possible patterns of left-turn phases.

TABLE 1 EIGHT POSSIBLE PATTERNS OF LEFT-TURN PHASE IN TERMS OF  $l_i, \bar{l}_i$ , AND 0-1 VARIABLES

Left-turn phase pattern	$\Delta_i$	$\beta_i$	$\bar{\beta}_i$	$\alpha_i$	$\bar{\alpha}_i$
1	$-\frac{1}{2} l_i$	0	1	1	0
2	$-\frac{1}{2} \bar{l}_i$	0	1	0	1
3	$-\frac{1}{2} (l_i + \bar{l}_i)$	0	1	1	1
4	$\frac{1}{2} l_i$	1	0	1	0
5	$\frac{1}{2} \bar{l}_i$	1	0	0	1
6	$\frac{1}{2} (l_i + \bar{l}_i)$	1	0	1	1
7	$-\frac{1}{2} (l_i - \bar{l}_i)$	0	0	1	1
8	$\frac{1}{2} (l_i - \bar{l}_i)$	1	1	1	1

These eight phases  $\Delta_i$  (time from center of inbound red to nearest center of outbound red at intersection  $i$ ) can be expressed by  $l_i$  and  $\bar{l}_i$ . The relationships of eight left-turn phases related to  $l_i$  and  $\bar{l}_i$  are given in Table 1. Furthermore, each left-turn phase can be represented by the following general equation:

$$\Delta_i = (1/2)[(2\beta_i - 1)\alpha_i l_i - (2\bar{\beta}_i - 1)\bar{\alpha}_i \bar{l}_i] \quad (27)$$

where  $\alpha_i, \beta_i, \bar{\alpha}_i, \bar{\beta}_i$  are 0-1 variables.

The values of  $\alpha_i, \beta_i, \bar{\alpha}_i, \bar{\beta}_i$  corresponding to each left-turn phase are also given in Table 1.

Based on the previous discussion, a complete mathematical programming formulation of this new algorithm is given as follows:

$$\text{MAX } b + \bar{b}$$

$$ST \bar{b} = K b$$

$$1/C_2 \leq Z \leq 1/C_1$$

$$t_i + \bar{t}_i + (1/2)(r_i + \bar{r}_i) - (1/2)(r_{i+1} + \bar{r}_{i+1}) + (Q_i - Q_{i+1}) + (H_i - H_{i+1}) + (\bar{W}_i - \bar{W}_{i+1}) + \Delta_i - \Delta_{i+1} = I_i \quad i = 1, \dots, n-1$$

$$OFF_i = t_i + (Q_i - Q_{i+1}) + (H_i - H_{i+1}) \quad i = 1, \dots, n-1$$

$$\overline{OFF}_i = \bar{t}_i + (\bar{Q}_{i+1} - \bar{Q}_i) + (\bar{H}_{i+1} - \bar{H}_i) \quad i = 1, \dots, n-1$$

$$Q_i + H_i + b + W_i + r_i = 1 \quad i = 1, \dots, n$$

$$\bar{Q}_i + \bar{H}_i + \bar{b} + \bar{W}_i + \bar{r}_i = 1 \quad i = 1, \dots, n$$

$$H_{i+1} \geq H_i + Q_i \quad i = 1, \dots, n-1$$

$$\bar{H}_i \geq \bar{H}_{i+1} + \bar{Q}_{i+1} \quad i = 1, \dots, n-1$$

$$Q_{i+1} \geq SH_{i+1} \times [\lambda_{i+1,1} \times (W_i - W_{i+1}) + \lambda_{i+1,2} \times r_i] \quad i = 1, \dots, n-1$$

$$\bar{Q}_i \geq \bar{S}\bar{H}_i \times (\bar{\lambda}_{i,1} \times (\bar{W}_{i+1} - \bar{W}_i) + \bar{\lambda}_{i,2} \times \bar{r}_{i+1}) \quad i = 1, \dots, n-1$$

$$b_i = b + Q_i + H_i \quad i = 1, \dots, n$$

$$\bar{b}_i = \bar{b} + \bar{Q}_i + \bar{H}_i \quad i = 1, \dots, n$$

$$(d/f_i)Z \leq t_i \leq (d/m_i)Z \quad i = 1, \dots, n-1$$

$$(\bar{d}/\bar{f}_i)Z \leq \bar{t}_i \leq (\bar{d}/\bar{m}_i)Z \quad i = 1, \dots, n-1$$

$$(d/h_i)Z \leq (d/d_{i+1})t_{i+1} - t_i \leq (d/n_i)Z \quad i = 1, \dots, n-2$$

$$(\bar{d}_{i+1}/\bar{h}_{i+1})Z \leq (\bar{d}_{i+1}/\bar{d}_i)\bar{t}_i - \bar{t}_{i+1} \leq (\bar{d}_{i+1}/\bar{n}_{i+1})Z \quad i = 1, \dots, n-2$$

$$(1/2)(r_i + \bar{r}_i) - \Delta_i \geq MIG_i \quad i = 1, \dots, n$$

$$r_i \geq MIG_i \quad i = 1, \dots, n$$

$$\bar{r}_i \geq MIG_i \quad i = 1, \dots, n$$

$$\Delta_i = (1/2)[(2\beta_i - 1)\alpha_i l_i - (2\bar{\beta}_i - 1)\bar{\alpha}_i \bar{l}_i] \quad i = 1, \dots, n$$

$$b, \bar{b}, Z, W_b, \bar{W}_b, H_b, t_b, \bar{t}_b, OFF_b, \overline{OFF}_b, Q_b, \bar{Q}_b \geq 0$$

$$I_i = \text{integer}$$

$$\alpha_b, \beta_b, \bar{\alpha}_b, \bar{\beta}_b = 0 \text{ or } 1$$

It should be noted that one may consider only part of this complete mathematical formulation to obtain the band, depending upon the user's requirements. If fewer constraints are included for analysis, the user obviously will have a wider progression bandwidth. To solve this mixed-integer programming problem, a variety of solution methods can be considered. If solving the new algorithm optimally through the above formulation is needed, the major consideration lies in the effectiveness of the mixed-integer programming packages. The Linear, INteractive and Discrete Optimizer (LINDO) (13) is considered here to solve this formulation. Although the free input form to run LINDO is easy to prepare, this new algorithm still requires substantial effort in learning how to formulate and create an input file to run LINDO. As far as sensitivity analysis and future applications are concerned, it is better to write a computer program that obtains the real progression bandwidth automatically based on the proposed formulation. A FORTRAN-based program named the BANDwidth of Timing Optimization Program (BANDTOP) has been developed to find the maximum progression bandwidth in both directions. It is a user-friendly program that improves the computational efficiency and ease of use by traffic engineers. BANDTOP provides much flexibility and convenience in responding to the changes of formulation or

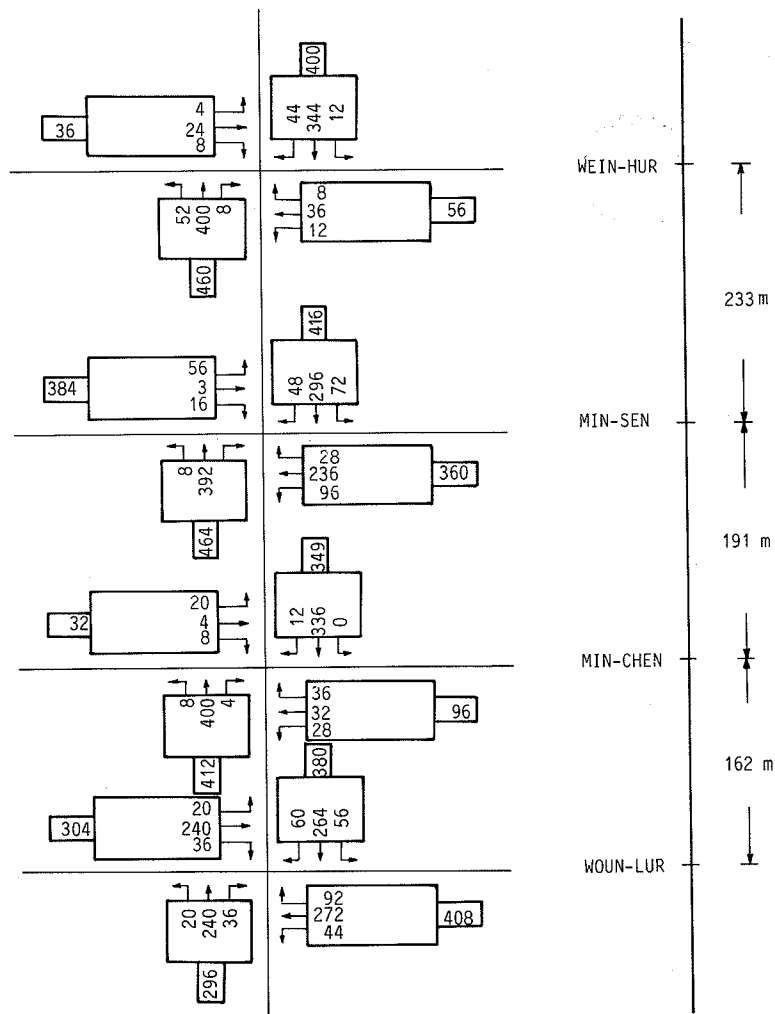
arterial configuration. It is noted that this new bandwidth program gives the optimal solution. BANDTOP can be run on PCs, VAX, IBM, or CDC and has been considered to generate progression signal timing plans for real-time traffic control systems in the cities of Keelung and Taichung, Taiwan.

**AN EXAMPLE**

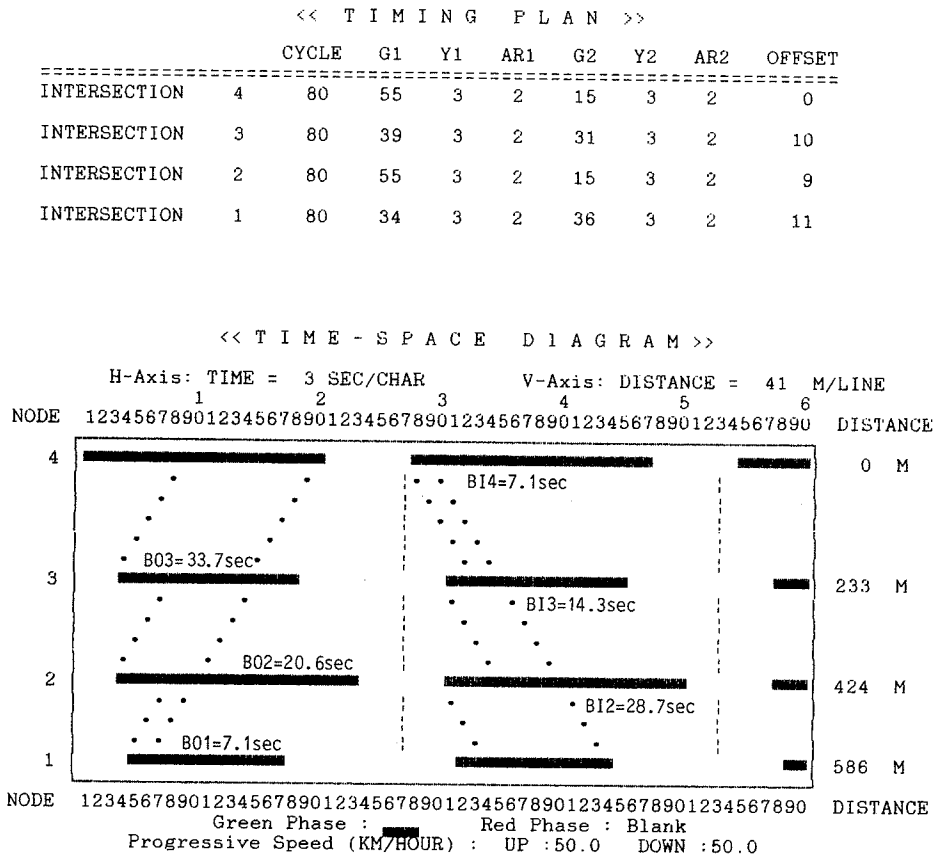
The new bandwidth algorithm has been tested on the Zin-Wha Arterial in Tainan City, south of Taiwan. Four intersections exist in this arterial. Figure 7 gives the network geometry and traffic flows on these four intersections. The inputs for this new algorithm consist of the order and distances of signals between intersections, traffic flows and capacities, range of speed, left-turn phase pattern, acceptable range of cycles, and the target ratio of bandwidths on different directions. The user can either specify the green splits at each intersection as a fraction of the overall cycle or calculate them through Webster's formula (1). The upper and lower limits of speed in this example are assumed to be 50 km/hr and 30 km/hr, respectively. The saturation flow headway equals 2.07 sec based on a recent study (14).

Through given information, BANDTOP finds the optimal signal timing plan for four intersections and its time-space diagram is shown in Figure 8. From Figure 8, this new bandwidth approach clearly produces offsets and other signal timing parameters. The maximum bandwidths in both directions are 14.2 sec. Any vehicle within this band, unlike MAX-BAND and PASSER II, can travel through all downstream intersections without stopping. As far as intersection 2 (Ming-Chen) is concerned, 20.6 sec outbound and 28.7 sec inbound bandwidths exist. Similarly, at intersection 3 (Min-Sen), outbound and inbound bandwidths have 33.7 sec and 14.3 sec, respectively. The new algorithm also recognizes partial progression opportunities over the shorter sections of the arterial. The partial progression bandwidth becomes wider as a vehicle moves toward downstream intersections. This is important because through this partial progression bandwidth one can conclude that vehicles outside the through band will need to stop at most once to pass the entire section of an arterial. In other words, vehicles will not stop or stop only once to travel through the arterial if the timing plan generated from BANDTOP is to be implemented.

To make a consistent comparison, the same network information and traffic flows were used to prepare the inputs for



**FIGURE 7** Geometrics and traffic volumes of Zin-Wha arterial in Tainan, Taiwan.



**FIGURE 8** Output of BANDTOP for Zin-Wha arterial in Tainan, Taiwan.

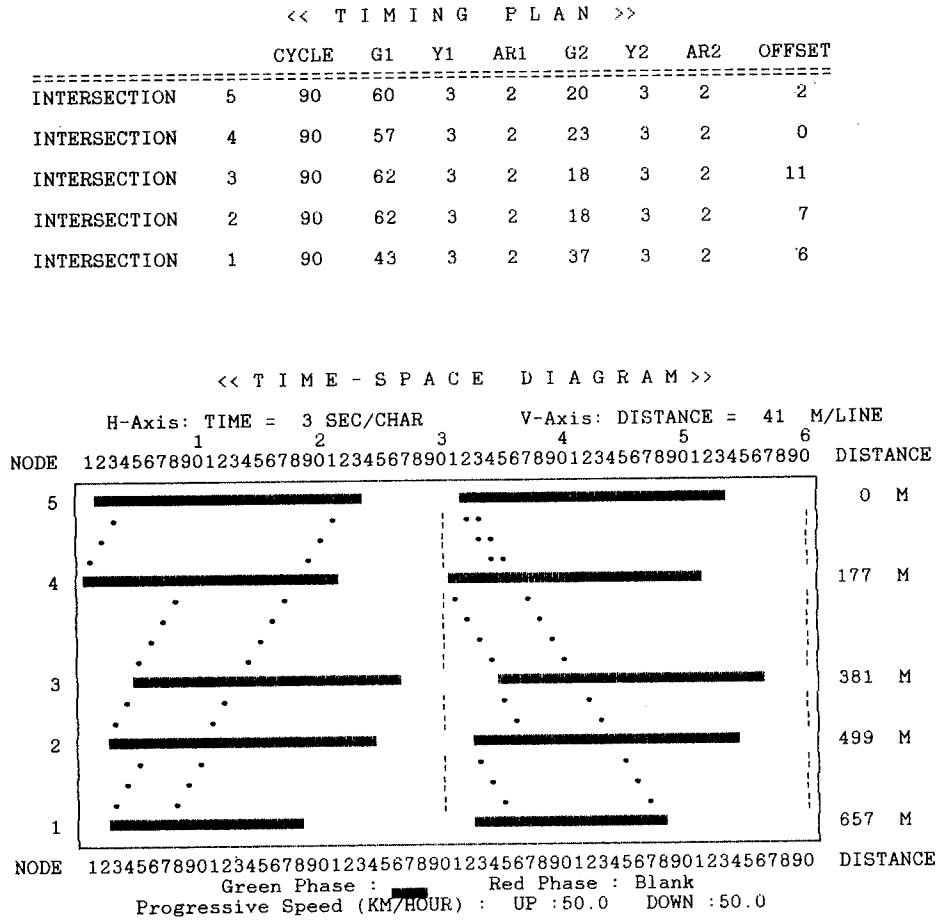
**TABLE 2** COMPARISON OF BANDTOP, MAXBAND, AND PASSER II SYSTEM PERFORMANCE THROUGH TRANSYT AND NETSIM

	System Performance	5 Intersections			4 Intersections		
		BANDTOP	MAXBAND	PASSERII	BANDTOP	MAXBAND	PASSERII
TRANSYT	PI	52.9	58.55	56.15	32.8	34.83	34.04
	Average Delay (sec/veh)	13.81	15.19	14.17	12.42	12.78	12.38
	Stop (%)	49	54	54	48	53	52
NETSIM	Average Delay (sec/veh)	28.27	30.9	29.03	25.0	23.48	24.46
	Stop (stops/veh)	0.94	1.12	1.04	0.83	0.82	0.9

MAXBAND and PASSER II. Figures 1 and 2 give time-space diagrams of two programs. By comparing Figure 8 with Figures 1 and 2, the new algorithm obviously produces a more reliable and acceptable progression bandwidth than MAXBAND and PASSER II in practical applications. Signal timing plans obtained from BANDTOP, MAXBAND, and PASSER II for four and five intersections are also analyzed by running TRANSYT-7F and NETSIM to evaluate their system performance. Results are given in Table 2. From this table, it can be seen that stops and average vehicle delay of BANDTOP are almost all less than those of MAXBAND and PASSER II. Figure 9 shows the computer output of BANDTOP for five intersections. The computer time of running BANDTOP on a PC/AT for three cases is available in Table 3. It takes only 36 sec and 72 sec to obtain the optimal solutions

for four and six intersections with a math coprocessor 80287-10 on a PC/AT. BANDTOP uses far less computing time than MAXBAND.

Because the input file of BANDTOP is similar to that of MAXBAND, users only need to make a slight modification of MAXBAND input to run BANDTOP. Details of conversion described in the BANDTOP user's manual will be released in the near future. It should be emphasized that this general mixed-integer mathematical programming formulation does not always guarantee the achievement of a feasible solution under given arterial configuration and traffic flows. If no feasible solution is available, it means that no real progression bandwidth can be realized through given conditions. Under such circumstances, the user may need to change arterial information, target ratio of bandwidths, or the number of



**FIGURE 9** Output of BANDTOP for five intersections.

**TABLE 3** COMPARISON OF BANDTOP RUNNING TIME ON PC/AT FOR THREE CASES

Number of Intersections	PC/AT Without Math Coprocessor	PC/AT With Math Coprocessor 80287-10
4	134 sec	36 sec
5	175 sec	46 sec
6	283 sec	72 sec

intersections considered in that segment to obtain the progression bandwidth in two directions.

**CONCLUSIONS**

From the research conducted in this work, it can be calculated that this new algorithm to find the maximal bandwidth in developing an arterial signal timing plan has many advantages over MAXBAND and PASSER II:

1. Unlike the current bandwidth approach, this approach guarantees that any vehicle in the progression band can travel through all downstream signals without stopping. Vehicles

outside the bandwidth will need to stop at most once to pass the entire section of the arterial.

2. It provides several features in practical applications. The program calculates queue clearance time and incoming flow clearance time automatically, provides eight left-turn phase patterns for selection, sets the minimum green time on side streets, and gives the target ratio of direction flow.

3. BANDTOP shows a better system performance than MAXBAND and PASSER II according to the stops and average vehicle delay of two real examples tested on NETSIM and TRANSYT-7F. The real progression bandwidth should have a saw-toothed shape. In addition, the partial progression bandwidth becomes wider as a vehicle moves toward downstream intersections.



4. If no feasible solution can be obtained from the new algorithm, it means that under given conditions no real progression bandwidth is available on the arterial. The user may change the number of intersections considered in the segment or run other kinds of signal timing packages for the designated arterial.

5. The new algorithm can consider leading and lagging phase patterns at each intersection. The use of a leading or lagging phase will result in a wider bandwidth for the arterial but increase the delay of vehicles from side streets.

6. Based on the proposed formulation, BANDTOP provides the optimal solution of bandwidth and requires less computer time to obtain the arterial signal timing plan and its time-space diagram. BANDTOP has been used successfully as a part of generating timing plan software at two real-time traffic control systems in the cities of Keelung and Taichung in Taiwan.

Through field tests, the new approach can give better and more reliable progression bands than MAXBAND and PASSER II. Therefore, it is recommended that this new algorithm be used to obtain the maximum bandwidth if the resultant signal timing plan is to be implemented.

#### ACKNOWLEDGMENT

The authors would like to acknowledge the assistance of Huei-Lung Huang during the phases of developing the BANDTOP program and preparing this final manuscript.

#### REFERENCES

1. J. D. C. Little, M. D. Kelson, and N. H. Gartner. MAXBAND: A Program for Setting Signal on Arteries and Triangular Networks. In *Transportation Research Record 795*, TRB, National Research Council, Washington, D.C., 1981, pp. 40-46.
2. J. D. C. Little. The Synchronization of Traffic Signals by Mixed-Integer Linear Programming. *Transportation Research*, Vol. 14, 1965, pp. 568-594.
3. M. D. Kelson. *Optimal Signal Timing for Arterial Signal System: Volume 1—Summary Report, Volume 2—User's Manual*. FHWA, U.S. Department of Transportation, April 1980.
4. M. M. Cuoms and J. L. Larocca. *MAXBAND User's Manual*. Traffic and Safety Division, New York State Department of Transportation, 1983.
5. M. D. Kelson, J. R. Mekemson, C. C. Liu, and S. L. Cohen. *MAXBAND—Source Program*. FHWA, U.S. Department of Transportation, Aug. 1985.
6. S. L. Cohen. Concurrent Use of MAXBAND and TRANSYT Signal Timing Programs for Arterial Signal Optimization. In *Transportation Research Record 906*, TRB, National Research Council, Washington, D.C., 1983, pp. 81-84.
7. R. L. Bleyl. A Practical Computer Program for Designing Traffic-Signal-System Timing Plans. In *Highway Research Record 211*, HRB, National Research Council, Washington, D.C., 1967, pp. 19-33.
8. E. C. P. Chang, C. J. Messer, and S. L. Cohen. Directional Weighting for Maximal Bandwidth Arterial Signal Optimization Programs. In *Transportation Research Record 1057*, TRB, National Research Council, Washington, D.C., 1986, pp. 10-19.
9. E. C. P. Chang and C. J. Messer. Minimum Delay Optimization of a Maximum Bandwidth Solution to Arterial Signal Timing. In *Transportation Research Record 1005*, TRB, National Research Council, Washington, D.C., 1985, pp. 89-95.
10. E. C. P. Chang. *Arterial Signal Timing Optimization—PASSER II-84 Workshop Manual*. Texas A & M University, College Station, Dec. 1985.
11. B. G. Marsden, E. C. P. Chang, and B. R. Dern. The PASSER II-84 System: A Practical Signal Timing Tool. *ITE Journal*, March 1987, pp. 31-36.
12. S. L. Cohen and C. C. Liu. The Bandwidth-Constrained TRANSYT Signal-Optimization Program. In *Transportation Research Record 1057*, TRB, National Research Council, Washington, D.C., 1986, pp. 1-9.
13. L. E. Schrang. *User's Manual for LINDO*. The Scientific Press, 1987.
14. H. S. Tsay and B. R. Lo. An Analysis of Platoon Dispersion and Saturation Flow Headway at Intersections. *Transportation Planning Quarterly* (Printed in Taiwan), Vol. 15, No. 4, 1986.

## DISCUSSION

EDMOND C.-P. CHANG

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

PASSER II and MAXBAND are two computerized signal timing programs currently available and popularly used for optimizing signal timing plans based on the maximum progression bandwidth concept. The maximum bandwidth approach can simultaneously optimize signal timing settings to provide the maximum weighted sum of arterial progression bands in both directions of an arterial street. This paper describes a modification of the constraints on the determination of the locations of the progression band with respect to the start of the arterial green times. This study examines the resultant reformulation being implemented in the original MAXBAND program and investigates the run-time efficiency after replacing the existing mixed-integer programming technique through microcomputer applications. The major difference between this enhanced algorithm and the original MAXBAND progression solution is that the enhanced algorithm provides wider progression bandwidths travelling farther toward the downstream intersections. It claims that the saw-toothed progression bandwidth generated by this formulation can allow some vehicles to travel through the arterial with at most one stop.

This study has raised several interesting points. First, the perception of the progression concept and definition of the maximum progression bandwidth may sometimes be misinterpreted. Second, it should be pointed out that this algorithm should only be considered an enhancement to the original MAXBAND algorithm, as the formulation and the computer program remain almost the same. The only new term being introduced by the authors is the modified " $W_i$ " variable, which is used to provide the preset maximum queue clearance settings. Third, since no simulation or field control validation of the new algorithm has been performed, other than the computer run-time evaluation, serious reservations exist concerning the effectiveness and validity of the enhanced algorithm. Fourth, it should be clearly stated that the significant improvements on the run-time efficiency from the test case examination were due primarily to the commercially available LINDO code in the program. This replaces the inefficient execution of the 1973 version of the Mixed-Integer Linear Programming Code (MILP) for solving this complex optimization problem. Fifth, due to the feasible number of intersections in the solutions and the mathematical characteristic of the algorithm for not being able to find a feasible solution, sincere reservations

exist about the intended use of this algorithm, as reported, for the real-time signal control applications.

The maximum progression, or the maximum bandwidth concept, is designed to provide the specific time intervals in which vehicles have the opportunity to travel through the downstream signalized intersections without having to stop. The existence of this arterial "Progression Bandwidth," or "Progression Opportunity," is an optimum time period whose existence is conditioned upon the interactions of cycle length, phase sequence, coordinated offsets, phase length, and, most important, intended target progression speeds. It should be noted that the progression bandwidth may be totally independent of the physical vehicle trajectories. "Vehicular Trajectories" represent the locations of vehicles arriving at a certain time. They can be used to examine whether vehicles following certain trajectories can travel through arterial streets. The existence of the progression band, however, does not guarantee that there will be certain vehicles lined up in the progression band. This simply means that the opportunities do exist in that time period for those vehicles that wish to follow the average target progression speed, and they may take advantage of the progression band to travel through without having to stop. However, the realization of the progression opportunities still depends on whether and how the designed progression bands can be utilized by the platooned and random vehicular arrivals in the field.

Progression may not work very well in those cases where target progression speeds were not set according to realistic operating speeds. Also, it may not function properly under those instances in which the progression phenomenon simply cannot exist because of heavy vehicular queue spillback or intersection blockage during arterial green times due to the overcongested operating conditions. However, most progression-based signal timing programs, such as PASSER II and MAXBAND, do allow users to adjust progression bands to some extent through the queue clearance features to tailor the progression time-space diagram to the potential queues observed in the field. On the other hand, multiple solutions may also exist to the same progression problems for given combinations of progression design speeds and coordinated background cycle lengths. This phenomenon is particularly noticeable in the coordinated two-phase operations. Therefore, the realization of the arterial progression bandwidth design approach depends heavily on whether the predicted progression can be achieved or fine-tuned according to actual vehicular performance during coordinated arterial traffic signal system operations.

It should be clearly stated that this algorithm can only be considered an extension to the original MAXBAND program because most of the MAXBAND formulation and all the program features remain exactly the same. Phase sequence, cycle length, green time, and offset optimization already existed in the original MAXBAND and PASSER II model. In addition, the benefits of using combinations of different traffic signal phase sequences to achieve a wider arterial progression bandwidth calculation were demonstrated in several earlier studies. Realistically, the new definition introduced by the authors only serves to modify the existing " $W_i$ " variable to provide a crude estimation of the maximum queue clearance settings without having to add a detailed traffic flow prediction model. The basic question that still remains is whether the saw-toothed type of progression bandwidth can provide a bet-

ter scheme than either the constrained MAXBAND-TRAN-SYT-7F progression approaches or the system delay offset fine-tuning optimization used in the PASSER II-84 approach.

Nevertheless, this paper did illustrate the significant run-time reduction that can be achieved by replacing the existing optimization code in the MAXBAND program to solve complicated optimization problems. This study examined the effectiveness obtained by replacing the relatively inefficient 1973 version of the MIL with the commercially available LINDO code. The inefficiency of the optimization algorithm and the commercial availability of Mixed-Integer Linear Programming Optimization codes have been commonly recognized. In parallel to this investigation, the maximum bandwidth program, MAXBAND 86, was enhanced by the Texas Transportation Institute in 1986 to simultaneously maximize the weighted sum of progression bandwidths on all the arteries of a signalized network. During its development, the same recommendations were made on the program run-time efficiency. It was decided to emphasize the network formulation and traffic engineering interpretation of the optimization results for developing the generalized arterial network optimization program. The results of an in-house study made by the Federal Highway Administration indicate that approximately 90 percent to 95 percent of the computer CPU time was spent on several subroutines of the MILP code of the MAXBAND 86 program during several test case runs. Therefore, the differences in run-time efficiency are contributed primarily to the replacement of the optimization code in MAXBAND, as all the other algorithms remain practically the same.

As indicated in this paper, this algorithm does suffer, as expected, from the inherent limitations of the number of intersections that can be feasibly analyzed to reach practical solutions efficiently. For the algorithm to provide feasible solutions, three important elements must exist. First, the arterial street directions must be given much larger amounts of green time than the cross street direction to provide the chance of generating a wider progression bandwidth farther toward downstream intersections. Second, to fully take advantage of the early start strategy for advancing the green times provided by the program, the saw-toothed type of progression approach will tend to favor those signal systems having short-spaced intersections, a large operating speed differential, and heavy turning traffic from side streets. This also implies that the algorithm tends to encourage the arterial vehicles travelling much faster than traffic turning from cross streets or slower vehicles. Third, full realization of this saw-toothed type of progression bandwidth relies heavily on the existence of equal amounts of green times to achieve maximum arterial progression.

As summarized from the above observations, the most successful operations of this enhanced algorithm are best suited for arterial signal systems having short spacing, small numbers of signals, large amounts of arterial green times, and almost optimum zero-offset coordination traffic operating conditions. In these cases, an alternative computerized signal operation can also be implemented through a series of two-phase signals with real-time green split adjustments without having to use the sophisticated Mixed-Integer Linear Optimization Problem for optimizing the one-line operation. At the same time, because of inherent limitations on the mathematical formulation due to the introduction of more constraints to the original optimization problem, the system does sometimes

suffer from not being able to reach a feasible solution. Therefore, serious reservations exist concerning the intended use of this enhanced MAXBAND algorithm for real-time traffic signal system control. Consequently, it is highly recommended that implementation of this enhanced algorithm be reserved until realistic validation studies, through either simulation studies or field controlled experiments, can be made available for further evaluation. Simulation studies, through either the TRANSYT-7F or NETSIM program evaluations, for examining the potential effectiveness of the enhanced algorithm versus the conventional constant progression bandwidth approach will be beneficial.

## AUTHORS' CLOSURE

Chang's discussion mainly concerns the concept and application of the new algorithm. From his discussion, several points are raised due to the misunderstanding of this algorithm. Each of these points will be discussed here.

Chang, in his second paragraph, states that the new algorithm introduces only a modified " $W_i$ " variable that is used to provide the preset maximum queue clearance settings. Actually, in our paper, seven major characteristics of the new algorithm compared to the MAXBAND are clearly discussed in the beginning of the section "Mathematical Formulation of New Algorithm." The new algorithm uses three new variables: queue clearance time ( $Q_i$ ), incoming flow clearance time ( $H_i$ ), and time lag ( $W_i$ ) for each intersection to take into consideration the clearing of queued vehicles before the arrival of platoons in the progression band. Thus, any vehicle within the band can travel through downstream intersection without stopping. All these variables need not be preset but are internally calculated based on the requirements of different arriving flow volumes. On the other hand, to assure vehicles in the through band cross the critical intersection without stopping, PASSER II and MAXBAND try to use the concept of preset queue clearance time. Nevertheless, it is impossible to know which intersection needs the queue clearance time and what value it should take. Even with an assumed or preset queue clearance time, the critical intersection will soon be shifted to another intersection according to the progression theory used in PASSER II and MAXBAND. The existing problem still remains unsolved. The new algorithm, however, can overcome this problem by introducing those three new variables at each intersection.

The discussant, in the second and fifth paragraphs, has emphasized that the new algorithm should only be considered as an enhancement to the original MAXBAND algorithm because most of the MAXBAND formulation and all the program features remain exactly the same. In our paper, we list complete formulations of the MAXBAND and the new algorithm separately in the context for comparison. Even part of the output of the MAXBAND and BANDTOP are also shown in Figures 2 and 8. The proposed algorithm uses a new progression concept and theory to handle the existing problem encountered by MAXBAND and PASSER II. A new mathematical programming formulation has been developed and enhanced LINDO is applied to obtain a saw-toothed pattern of the bandwidth instead of a parallel and uniform one. Obviously, it is different from the original MAXBAND and

PASSER II. The discussant, however, strongly objects to the term of a "new" algorithm.

The discussant mentions the importance of performing simulation study through either TRANSYT-7F or NETSIM and field tests to the new algorithm and its program BANDTOP. We certainly agree; some of simulation results and comparisons among BANDTOP, PASSER II, and MAXBAND on two arterials with four and five intersections have been shown in Table 2. At present, BANDTOP is used as a part of computing software for timing plan generation on a new real-time traffic control system, named the Traffic Responsive and Uniform Surveillance Timing System (TRUSTS), located in the cities of Keelung and Taichung in Taiwan. Many results can be obtained from the field and considerations have been given to perform further evaluation and to improve the TRUSTS performance and current BANDTOP version.

The discussant tries to reinterpret the concept of bandwidth in the third and fourth paragraphs. In the paper, we have pointed out the existing progression problem if the output of PASSER II and MAXBAND is implemented directly. The vehicles in the front portion of the through band will be hindered and have to stop at the critical intersection. This problem mainly comes from inaccurate progression theory with unrealistic assumptions made in the PASSER II and MAXBAND. The phenomenon can be easily observed from the field and time-space diagrams. Although the discussant, in his fourth paragraph, tries to use a preset queue clearance time to tailor the progression time-space diagram to the potential queues observed in the field, the existing problem still cannot be solved. This is simply because queues are varying from time to time and the estimated or observed queues are only suited for a particular time and day. It is uneconomical for the users to check the potential queue of each intersection in the field every time. In other words, if the current output of the MAXBAND and PASSER II with an impressive bandwidth operates through time-space diagrams without additional manual adjustments, it will provide practicing traffic engineers with false information in determining whether to choose to implement the signal timing plan.

The discussant mentions in the seventh paragraph that for the new algorithm to provide a feasible solution, three important elements must exist. In fact, none of these three points are accurate. First, he states that "... the arterial street direction *must* be given *much* larger amounts of green time than the cross street direction. . . ." From Figure 8 of the paper, it can be seen that the arterial has 34 sec green versus 36 sec green of the cross street at intersection 1 and 39 sec green of the arterial versus 31 sec of the minor street at intersection 3. Second, he states that "... the approach will tend to favor those signal systems that have short-spaced intersections, a large operating speed differential, and heavy turning traffic from side streets." The new algorithm, in fact, can handle long-space intersections even over one cycle travel time of one block distance based on the integer value of  $I_i$  in Equation 2. That is the reason why we claim the new algorithm has the general mixed-integer formulation. Since the new algorithm considers the general case, it will be able to deal with various kinds of street types and flow patterns. Furthermore, the optimal travel speed in Figures 8 and 9 remains 50 km/hr (31 mph) for both directions. We do not understand why the discussant concludes that the new algorithm tends to favor a large operating speed differential, heavy turning traffic,

and the arterial vehicles travelling much faster than traffic turning from cross streets or slow vehicles. Third, he mentions that “. . . bandwidth relies heavily on the existence of equal amounts of green time to achieve maximum arterial progression.” Here we are not sure whether the green time refers to the bandwidth or the green time interval of that intersection. In either case, it is not true, as Figures 8 and 9 demonstrate.

In the last paragraph, the discussant summarizes his observations and states, “. . . the most successful operations of this enhanced algorithm are best suited for arterial signal systems having short spacings, small numbers of signals, large amounts of arterial green times, and almost optimum zero-offset coordination traffic operating conditions.” Some of the points have been explained above. Similarly, the offsets shown in Figures 8 and 9 ranging from 2 to 11 sec reveal that the new algorithm is not only suited for almost zero-offset coordination. Besides, the new algorithm can handle various types of

intersections and traffic flow. In our paper, we mention that if no feasible solution can be obtained from the new algorithm, it means that no real progression bandwidth is available on the arterial under given conditions. This is due to the proposed optimization model that assures any vehicle in the band can travel through all downstream intersections without stopping and outside the band with at most one stop. The bandwidth should be the saw-toothed shape. Therefore, we agree to the point that the new algorithm is suited for small numbers of signals, but not exclusively, because the optimal solution relies mainly upon the traffic flow movement and block distances of intersections under consideration in the segment.

---

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

# Examination of Shared Lane Operations

J. A. BONNESON, C. J. MESSER, AND D. B. FAMBRO

The shared use of a single traffic lane by through and left-turn movements is one of the more complex operations that can occur at signalized intersections. A closed-form solution for evaluating the effect of shared lane use is described in the *Highway Capacity Manual* (HCM). This paper investigates the methodology of the HCM shared lane analysis. It also extends that methodology to recognize the operational interdependence of saturation flow rate and lane use on opposing approaches. This extension is in the form of an iterative modification wherein the saturation flow rate and lane use on opposing approaches are incrementally updated. Using the modified methodology, several investigations were undertaken to determine the behavior of shared lane operations. These investigations included comparing the modified methodology with the original HCM methodology; studying convergence trends; evaluating the effects of various timing and volume conditions; isolating a maximum volume threshold; and identifying the shared versus de facto left-turn lane regime. As a result of this examination, it was found that the HCM methodology consistently estimated slightly lower saturation flow rates than the final flow rate converged upon. A major outcome of the sensitivity analysis and evaluation study was a graphical technique for estimating the operational nature of a shared traffic lane.

The operation of traffic in a lane shared by left-turning and through vehicles is difficult to describe both in general and mathematical terminology. Other authors (1, 2, 3, 4) have described the complex combination of events that occur in shared lanes. However, the sensitivity of shared lane operations to various timing and volume conditions has yet to be adequately described or understood. This paper offers another look at opposed, shared lane operations at traffic signals.

The recent publication of the *Highway Capacity Manual* (HCM) (5) has heightened the need for a better understanding of shared lane operations. Chapter 9 of this manual presents a methodology for estimating the saturation flow rate of an intersection approach having a shared lane. Although this methodology has been well documented in the HCM, a basic understanding of the effects of each of the input variables (e.g., volume and signal timing) is still greatly needed. In essence, without understanding the trends and tendencies of the methodology, the analyst cannot be confident of the results or their implication.

One goal of this paper is to examine the theoretical sensitivity of shared lane operations (as modeled by the HCM methodology) to changes in several control variables. Another goal is to examine the interdependent relationships among operations on opposing approaches. Finally, the realm of de facto left-turn lane operation (i.e., a shared lane operating as exclusive left-turn lane) will be quantified and described in terms of the conditions that induce its occurrence.

## APPROACH

An intersection having two opposing, two-lane, shared lane approaches was considered to be the typical shared lane situation. This intersection, shown in figure 1, has two lanes on each approach; the inside lane could be shared by through and left-turn vehicles, while the outside, or curb, lane would be used exclusively by through vehicles. The intersection is served by a two phase signal, that is, one phase for each street. For the sake of simplicity, neither cross street nor right-turn traffic volumes were considered.

The variables considered for their effect on shared lane operation are shown in table 1. They include the approach volume ( $V_a$ ), the opposing approach volume ( $V_o$ ), the proportion of approach traffic that turns left ( $P_{LT}$ ), the proportion of opposing traffic that turns left ( $P_{LTO}$ ), the cycle length ( $C$ ), and the total green plus yellow time ( $G$ ). These variables were selected because they were felt to have the greatest influence on shared lane operation.

Each variable was independently varied over a range of typical values to determine its effect on the overall operation of a shared lane. The measures used to monitor these variational effects were the lane group saturation flow rate ( $S_a$ ) and the proportion of left-turns on the inside lane ( $P_L$ ). For each of the variables studied, a pair of figures was generated to illustrate the trends and sensitivities of  $S_a$  and  $P_L$ . These figures are included below in the discussion of sensitivity analysis.

It should be noted that one variable affecting shared lane operation, the saturation flow rate for through traffic ( $S_T$ ), was not varied in this study. Although it is recognized that this variable could have a significant effect on shared lane operations, it was reasoned that investigation of the other variables would be more pertinent and relevant to the goals of this paper. The saturation flow rate used for this study was less than the ideal rate of 1,800 vehicles per hour of green time per lane (vphgpl). This was deliberately done to illustrate the effects of a less-than-ideal saturation flow rate on shared lane operations.

## SHARED LANE MODELS

As noted above, several models have been offered for the analysis of shared lane operations. All of these models are based on a probabilistic approach wherein the opportunities for left-turn or through movement departures are quantified in terms of expected or average rates. This type of model has the advantage of providing both a logical and tractable solution, but the disadvantages of being both complex and iterative, by nature of its dependency on opposing lane operations. Of these models, the shared lane methodology identified

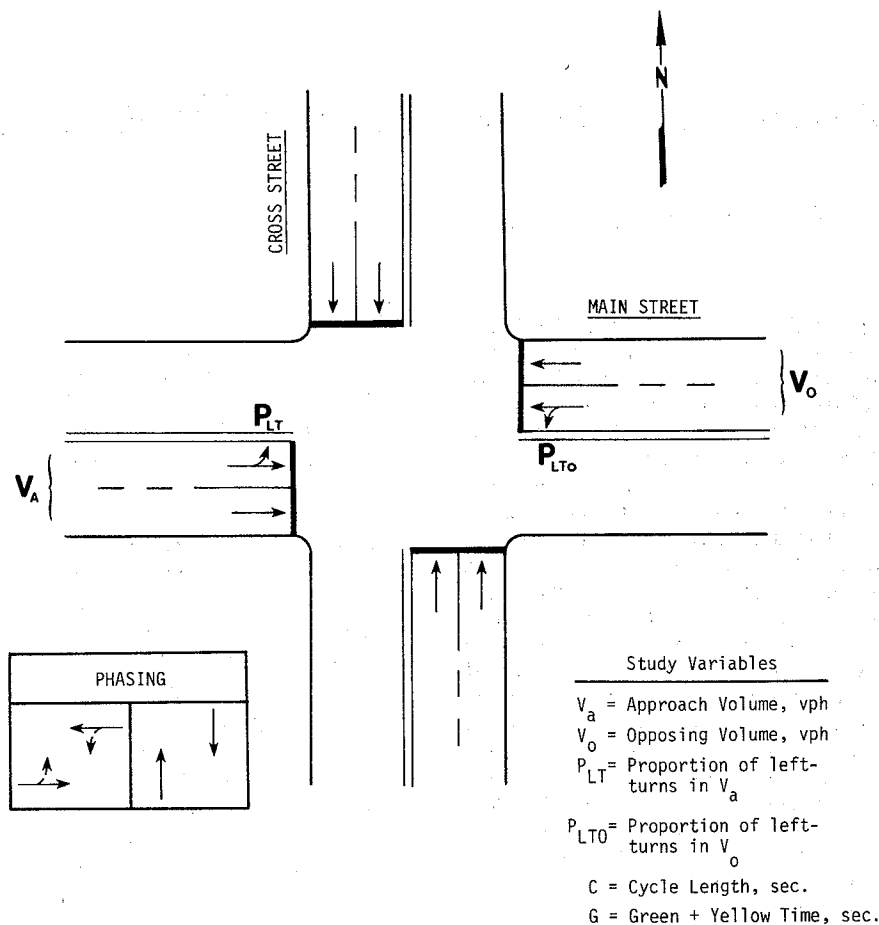


FIGURE 1 Typical shared lane configuration.

in chapter 9 of the HCM was selected for this evaluation for two reasons: its publication in the HCM makes it the more popular methodology, and its simplicity with respect to other probabilistic models provides a reasonable balance between theory and practicality.

#### HCM Model

The HCM methodology should theoretically provide a good estimate of shared lane operations. As shown in figure 2, the

TABLE 1 STUDY VARIABLES

$P_{LT}$	$G/C$	$C$ (sec)	$V_a, V_o^a$ (vph)	$P_{LT}, P_{LTO}^b$	$S_r^c$ (vphgpl)
0.01	0.20	50	200	0.01	<u>1650</u>
0.10	0.30	60	600	0.10	
<u>0.15</u>	0.40	<u>70</u>	800	<u>0.15</u>	
0.20	0.50	80		0.20	
0.30	0.60	90		0.30	
	0.80	100			
		110			
		120			

NOTE: Underlined values represent the base condition. These values were fixed during the sensitivity analysis of a particular variable.

<sup>a</sup>All volume elements were set equal unless otherwise noted.

<sup>b</sup>Proportion of left-turns on the subject and opposing approaches were set equal. For each volume level,  $P_{LT}$  and  $P_{LTO}$  were varied through their entire range.

<sup>c</sup>Saturation flow rate was not varied.

process has been reduced to a series of intermediate calculations. The final result is a factor that can be used to calculate the approach saturation flow rate.

The HCM model evaluates shared lane operation in terms of three components. These components, as they occur from the commencement of green, are

- Period 1—an initial portion, wherein some through vehicles can proceed before the first, blocking left-turn vehicle arrives at the head of the queue,
- Period 2—an interval subsequent to the clearance of the opposing queue wherein both through and left-turn vehicles can depart, and
- Period 3—a final period wherein left-turn vehicles clear the intersection before the initiation of the cross street phase.

Several authors have proposed the existence of other capacity components (1, 2). The contribution of these components, however, is generally small. As a result, the increased complexity of calculation of these components does not appear to be justified, given the overall accuracy of the process.

The HCM method should provide a good estimate of shared lane operations. However, the calculated saturation flow rate can only be taken as an estimate, due to several assumptions and simplifications embedded in the methodology. In particular, assumptions are embedded in the equations for calculating the opposing saturation flow rate ( $S_{op}$ ) and the pro-

SUPPLEMENTAL WORKSHEET FOR LEFT-TURN ADJUSTMENT FACTOR, $f_{LT}$				
INPUT VARIABLES	EB	WB	NB	SB
Cycle Length, C (sec)	70	70	70	70
Effective Green, g (sec)	27	27	37	37
Number of Lanes, N	2	2	1	1
Total Approach Flow Rate, $v_a$ (vph)	800	833	466	667
Mainline Flow Rate, $v_m$ (vph)	800	833	433	623
Left-Turn Flow Rate, $v_{LT}$ (vph)	72	33	33	44
Proportion of LT, $P_{LT}$	0.09	0.04	0.07	0.07
Opposing Lanes, $N_o$	2	2	1	1
Opposing Flow Rate, $v_o$ (vph)	833	800	623	433
Prop. of LT in Opp. Vol., $P_{LTO}$	0.04	0.09	0.07	0.07
COMPUTATIONS	EB	WB	NB	SB
$S_{op} = \frac{1800 N_o}{1 + P_{LTO} \left[ \frac{400 + v_m}{1400 - v_m} \right]}$	3333	3012	1698	1648
$Y_o = v_o / S_{op}$	0.250	0.266	0.367	0.263
$g_u = (g - CY_o) / (1 - Y_o)$	12.67	11.42	17.87	25.24
$f_s = (875 - 0.625 v_o) / 1000$	0.354	0.375	—	—
$P_L = P_{LT} \left[ 1 + \frac{(N-1)g}{f_s g_u + 4.5} \right]$	0.360	0.163	0.070	0.070
$g_q = g - g_u$	14.33	15.58	19.13	11.76
$P_T = 1 - P_L$	0.640	0.837	0.930	0.930
$g_f = 2 \frac{P_T}{P_L} \left[ 1 - P_T^{0.5} g_q \right]$	3.41	7.70	13.29	9.22
$E_L = 1800 / (1400 - v_o)$	3.17	3.00	2.32	1.86
$f_m = \frac{g_f + g_u}{g} \left[ \frac{1}{1 + P_L (E_L - 1)} \right] + \frac{2}{g} (1 + P_L)$	0.490	0.690	0.859	0.950
$f_{LT} = (f_m + N - 1) / N$	0.75	0.85	0.86	0.95

FIGURE 2 Shared lane analysis worksheet (5).

portion of left-turn vehicles in the left lane ( $P_L$ ). In both of these equations, estimates must be made regarding shared lane operations on the opposing and subject lane groups before the evaluation can be completed.

Other assumptions inherent in the formulation are the use of ideal saturation flow rates (*i.e.*, 1,800 vphgpl) in the equations for the initial portion of green time ( $g_f$ ), left-turn equivalency ( $E_L$ ), opposed saturation flow rate ( $S_{op}$ ), and the left-turn factor ( $f_m$ ). It will be shown later that the use of ideal saturation flow rates tends to yield conservative estimates of shared lane capacity. This was an intentional adjustment in

recognition of the closed-form computational approach recommended by the HCM.

#### Iterative Model

Recognizing the limitations of the aforementioned assumptions in the HCM formulation, a modification was proposed and investigated for its effect on the calculated saturation flow rate. This modification was directed towards an iterative approach wherein the values of  $S_{op}$  and  $P_L$  were recalculated

based on previous results, or iterations. This procedure has the advantage of being able to incorporate better estimates of  $S_{op}$  and  $P_L$  during each iteration, which is particularly important in those instances where the opposing approach also has shared lane operations. The proposed model is described below as a sequence of calculation steps.

- Step 1.  $Y_o = V_o / S_{op}(i-1)$
- Step 2.  $g_u = (g - C * Y_o) / (1 - Y_o)$
- Step 3.  $P_L = P_{LT} * [1 + (N - 1) / f_m(i-1)]$
- Step 4.  $g_q = (g - g_u)$
- Step 5.  $g_f = 2.0 * [(1 - P_L) / P_L] * [1 - (1 - P_L) * (g_q * S_T / 3600)]$
- Step 6.  $E_L = S_T / (1400 - V_o)$
- Step 7.  $f_{m(i)} = [g_f + g_u / (1 + P_L * (E_L - 1)) + 3600 * (1 + P_L) / S_T] / g$
- Step 8.  $f_{LT} = (f_{m(i)} + N - 1) / N$
- Step 9.  $S_{a(i)} = S_T * f_{LT} * N$

where

- $C$  = cycle length, sec.
- $E_L$  = through-vehicle equivalent for opposed left-turns.
- $f_{m(i)}$  = left-turn factor for the shared lane (Note:  $f_{m(i)} < 1.0$ ).
- $g$  = effective green time, sec.
- $g_f$  = duration of initial portion of green phase, sec.
- $g_q$  = portion of green phase blocked to left-turning vehicles by the clearing of an opposing queue of vehicles (=  $g - g_u$ ), sec.
- $g_u$  = portion of green not blocked by the clearing of an opposing queue, sec.
- $i$  = current calculation sequence.
- $N$  = number of lanes on approach.
- $P_L$  = proportion of left-turn vehicles in shared lane.
- $P_{LT}$  = proportion of left-turn vehicles on the approach.
- $S_{a(i)}$  = saturation flow rate for subject lane group, vphg.
- $S_{op(i-1)}$  = saturation flow rate on opposing approach taken from the previous calculation, vphg.
- $S_T$  = through-vehicle saturation flow rate on subject approach (=  $1800 * f_w * f_{HV} * f_g * f_a$ ), vphgpl.
- $V_o$  = opposing flow rate, vph.
- $Y_o$  = flow ratio on the opposing approach.

As shown in the preceding steps, the proposed model adopts the same format and sequence of calculation as that of the HCM methodology. The only deviations are the use of information from preceding calculations and the use of a less-than-ideal saturation flow rate where ideal values had previously been assumed.

In recognition of the iterative nature of the proposed procedure, the methodology was programmed in BASIC to automate the analysis process. The function of this program was to make an initial calculation using the original HCM methodology and to use the modified methodology for second and subsequent iterations.

For this study, a fairly strict convergence criterion was selected such that the nature of the solution process could be fully explored. In fact, a solution was said to have "converged" when the difference between approach saturation flows for two successive iterations was less than 0.5 vphg. While the allowable deviation for the final calculations was quite small,

experience with the procedure indicates that the likelihood of arriving at a convergence solution is almost certain.

Using this criterion, the number of iterations was found to range from one to thirty iterations. Convergence was commonly achieved, however, in less than ten iterations for the conditions of this study. In general, two iterations beyond the initial HCM solution would typically yield a saturation flow rate that was very near the convergence flow rate.

Two conclusions were drawn from this study of convergence. First, it appeared that the HCM methodology would generally yield a conservative estimate of the saturation flow rate for a shared lane. This tendency is consistent with the intent of its formulation. However, it should also be noted that for some near-capacity conditions the HCM solution was found to overestimate the shared lane saturation flow rate. Second, the saturation flow rate from the first (HCM) iteration was almost always found to be within 5% of the final solution.

## SENSITIVITY ANALYSIS

This section examines the sensitivity of shared lane operations to variations in volume and signal timing. For each of the variable combinations identified in table 1, an iterative analysis was performed which included an initial calculation of saturation flow using the original HCM methodology. All subsequent calculations used the modified methodology wherein information from the last iteration was used as a better estimate of actual conditions. The iterations were stopped when the approach saturation flow rate ( $S_a$ ) between two successive calculations did not change by more than 0.5 vphg. It should be noted that the results reported in this paper reflect under-capacity conditions for all approaches.

### Left-Turn Percentage

The sensitivity of the shared-lane methodology to variation in left-turn percentage was investigated for this analysis. As shown in figure 3, the effect of an increase in left-turn percentage caused a reduction in saturation flow rate for the subject approach. In particular, the flow rate varied between 3,300 vphg and 2,400 vphg for a left-turn percentage in the range of 1% and 30%, respectively. It should also be noted that when the percentage of left-turns was increased, lane use on the interior, shared lane approached that of an exclusive left-turn lane (*i.e.*,  $P_L = 1.0$ ). This result is reasonable and consistent with general expectation.

Concerning the first (HCM) solution versus the final (convergence) solution, the trends in convergence—*noted above in the discussion of the modified methodology*—were well illustrated. Not only did the HCM and final solution "track" one another, but it appeared that the HCM solution would yield a conservative, or lower, estimate of saturation flow rate.

### Cycle Length

For this investigation, a range of typical of cycle lengths was considered to determine the effect of cycle length on shared-



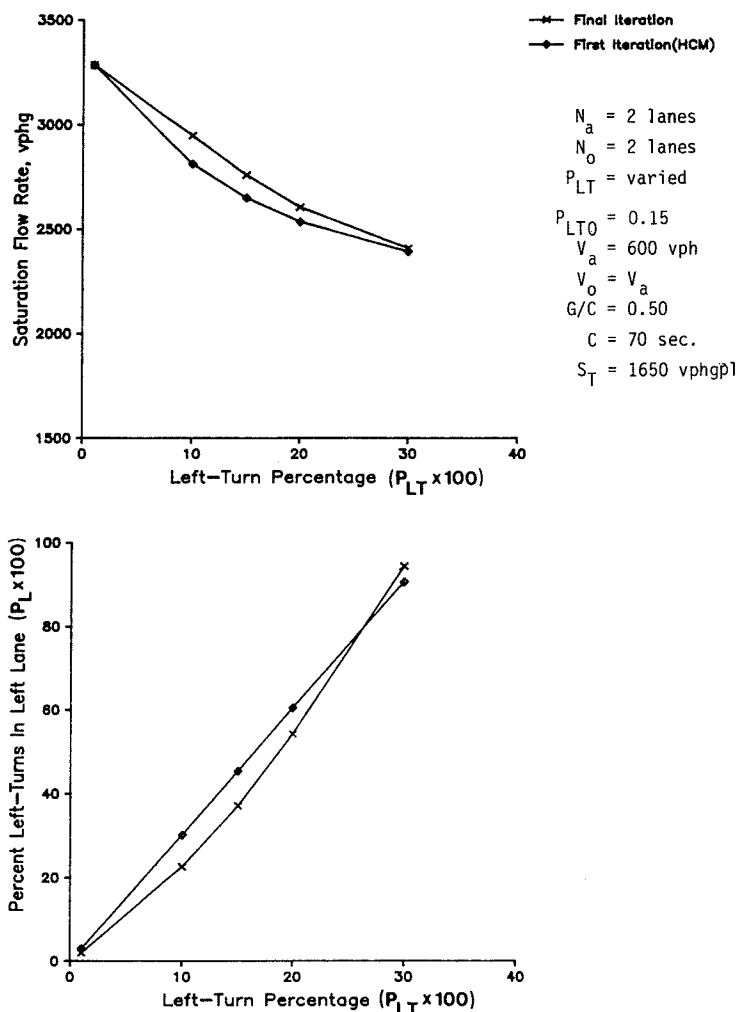


FIGURE 3 Effects of left-turn percentage.

lane operations. As shown in figure 4, the convergence trends appear to be consistent with those found in the investigation of left-turn percentage. In particular, the HCM solution again yielded a slightly lower value of saturation flow rate than the convergence solution. In addition, a positive relationship was again found to exist between the two solutions regarding their general agreement with trends in the changes in cycle length.

One of the more interesting results of this sensitivity analysis was the lack of any significant change in flow rate with cycle length. This result implies that the duration of the cycle length has a minimal effect on the operation of shared lanes. This can be explained as the result of two secondary effects. One is the increase in capacity per cycle during period 2 with an increase in cycle length. Vehicular capacity per cycle during this period is a direct function of the amount of unsaturated green time ( $g_u$ ) available. This green time increases almost linearly with cycle length when the green-to-cycle-length ( $G/C$ ) ratio is held constant.

The other secondary effect is the lower number of cycles and hence, clearance opportunities that can occur each hour when cycle length is increased. These two effects tend to cancel one another and thereby minimize the influence of cycle length on saturation flow. It should be noted that the

capacity of period 1 was not found to vary significantly with cycle length.

### Green-to-Cycle-Length Ratio

For this analysis, the effects of a change in the green time for a given cycle length were investigated. As shown in figure 5, changes in the  $G/C$  ratio were found to have an effect on the operation of the shared lane approach. This was evidenced by the wide variation in saturation flow rates for  $G/C$  ratios of less than about 0.40. In addition to this wide variation, there also appeared to be a deviation from the expected conservative nature of the HCM methodology.

Further investigation of the conditions that created the observed anomalies revealed that as  $G/C$  varies from larger to smaller ratios, the volume-to-capacity ratio ( $v/c$ ) of the approach neared 1.0. In fact, for  $G/C$  ratios less than 0.35, both approaches were found to be over capacity. Hence, it appeared that, for some near-capacity conditions, the saturation flow rate found using the HCM methodology could be greater than the iterated solution (*i.e.*, not conservative). It was also noted that, for  $G/C$  ratios greater than 0.40, the

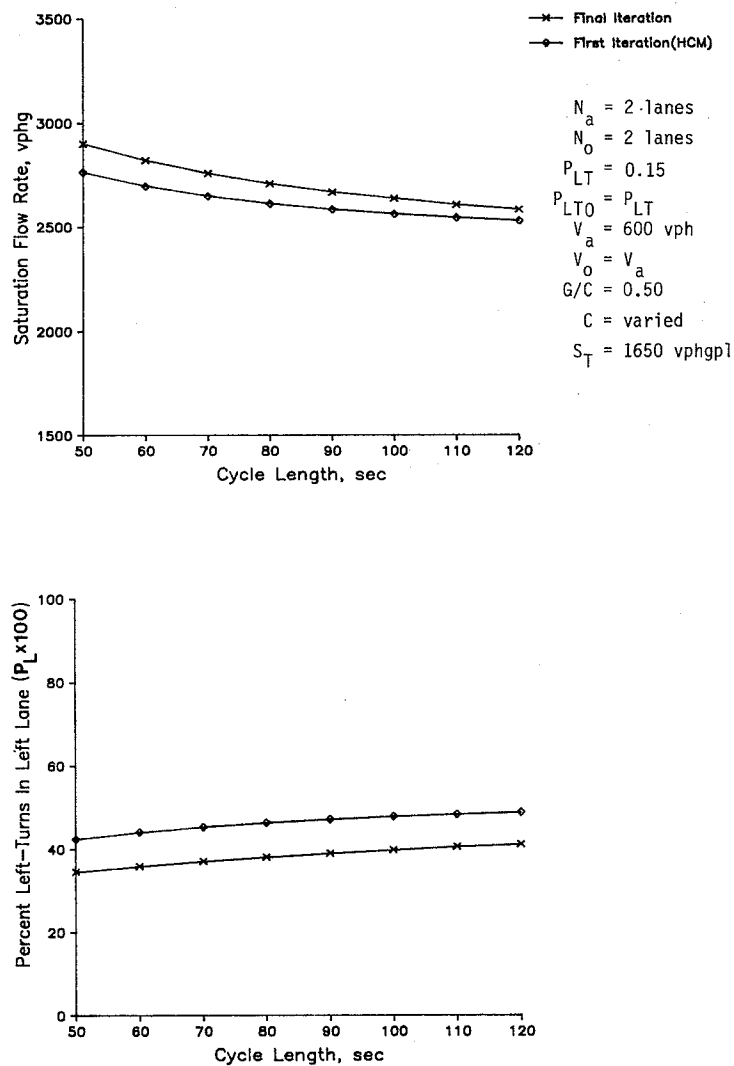


FIGURE 4 Effects of cycle length.

effect on saturation flow rate did not appear to be as significant.

#### Entering Volume and Left-Turn Percentage

The investigation of entering volume presented many possible combinations of opposing and subject approach volumes. Moreover, it was resolved that a thorough examination should consider effects of changes in both through and left-turn volumes. In recognition of the need for variation in these volumes, it was felt that a balanced volume condition would be the most reasonable compromise. Hence, for all volume conditions reported in this section, the volumes and left-turn proportions on each approach were held equal throughout all analyses.

As a result of equalizing the volumes and left-turn proportions, the investigation was reduced to an analysis of symmetric supply and demand conditions. In other words, the operation of the shared lane on either side of the intersection was identical given the same geometry, volume, and timing conditions. For each value of entering volume, five values of

left-turn proportion were examined for their effect on approach saturation flow rate. This approach was taken to ensure a consistency with the preceding analyses.

Figure 6 illustrates the results from this analysis. Based on these results, it was concluded that combinations of high left-turn and entering volumes would have a definite adverse effect on the saturation flow rates of two opposing shared lane approaches. This conclusion is intuitively reasonable, since the demand on a shared lane approach indirectly affects the capacity of the approach opposing it.

One trend observed in this analysis was that the HCM methodology did give conservative results under a wide range of volume conditions. In the few situations where the HCM solution was not conservative, it was found that the  $v/c$  ratio of the left-turn or through movement was near 1.0. This trend was consistent with that observed during the sensitivity analysis of the  $G/C$  ratio.

An explanation of the HCM solution's conservative results lies primarily in its formulation. Several major effects interact within the HCM methodology and result in a lower estimation of saturation flow rate for the shared lane. The first effect is embedded in the equations for  $S_{op}$  and  $P_L$  (see Figure 2). The

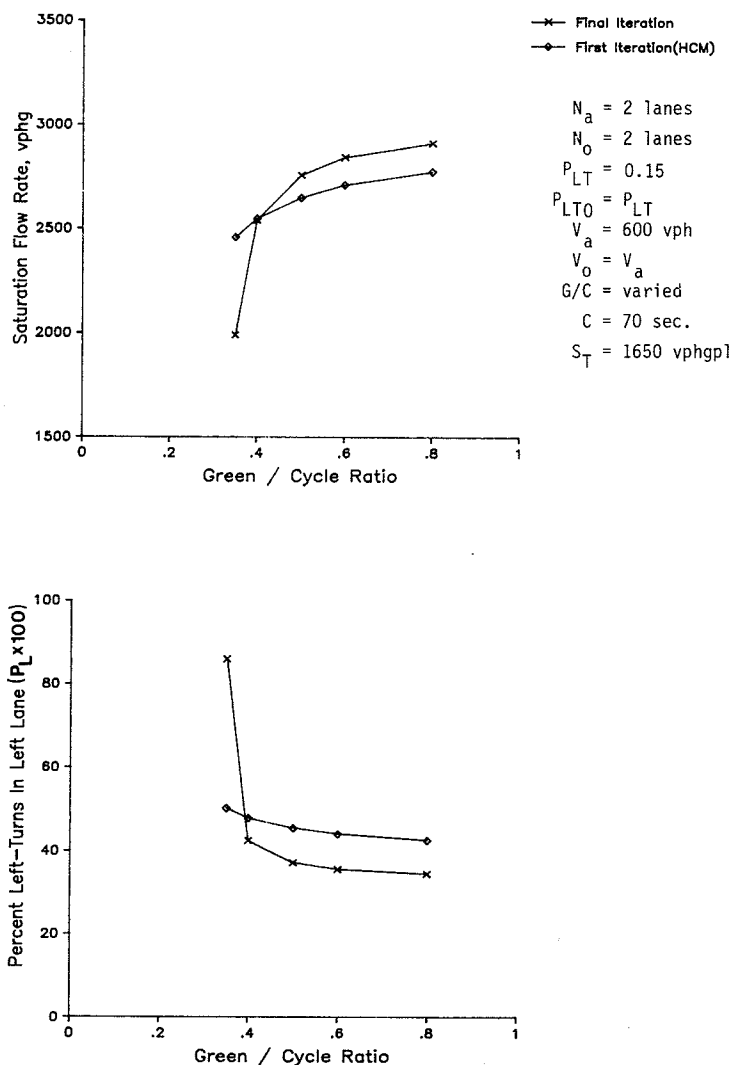


FIGURE 5 Effects of green/cycle ratio.

equation used to estimate  $S_{op}$  will almost always yield a higher value than that found in the final iteration. This implies an expectation of higher estimates of unsaturated green time ( $g_u$ ) and hence, a greater shared lane capacity during period 2. However, this is not generally found because of the effect of another factor,  $P_L$ . The equation used to estimate  $P_L$  typically overestimates the proportion of left-turn vehicles in the inside, or shared, lane. The implication of this approximate procedure can be found in a lower estimate of through vehicle capacity during period 1.

The effect of overestimating  $P_L$  explains some of the conservative nature of the HCM's saturation flow rate estimate. In contrast, the effect of overestimating  $S_{op}$  (and thereby  $g_u$ ) suggests a more liberal flow rate estimate. However, the combination of  $P_L$  and  $g_u$  in the equation used to calculate the left-turn adjustment factor ( $f_m$ ) tends to offset the effects of  $S_{op}$ . As a result, period 2 capacity is also conservatively estimated in most instances.

Another effect worthy of consideration is the calculation of capacity during period 3. As noted above, the HCM methodology generally overestimates the magnitude of  $P_L$  and results in an overestimate of capacity during the clearance interval

(i.e., period 3). However, the ultimate indirect factoring of this capacity component by the other saturation flow rate adjustment factors tends to overcompensate for the effect of an overestimated  $P_L$  and typically results in a net conservative estimate of period 3 capacity. It should be noted that the capacity of this period has traditionally been calculated as the number of vehicles clearing at the end of the cycle which is totally independent of the saturation flow rate experienced during the green phase.

### SHARED LANE RELATIONSHIPS

Based on the preceding sensitivity analysis, it appeared that the four dominant variables (with respect to shared lane operations) were proportion of left-turns ( $P_{LT}$ ), opposing proportion of left-turns ( $P_{LTO}$ ), approach volume ( $V_a$ ), and opposing volume ( $V_o$ ). Green time and cycle length were found to have a limited effect on shared lane operations.

In an attempt to understand better the interaction of the dominant variables, figure 7 was generated using the modified methodology. In this figure, each axis represents an inde-

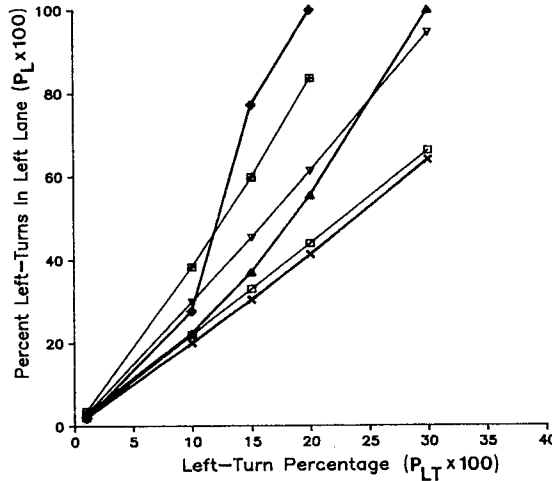
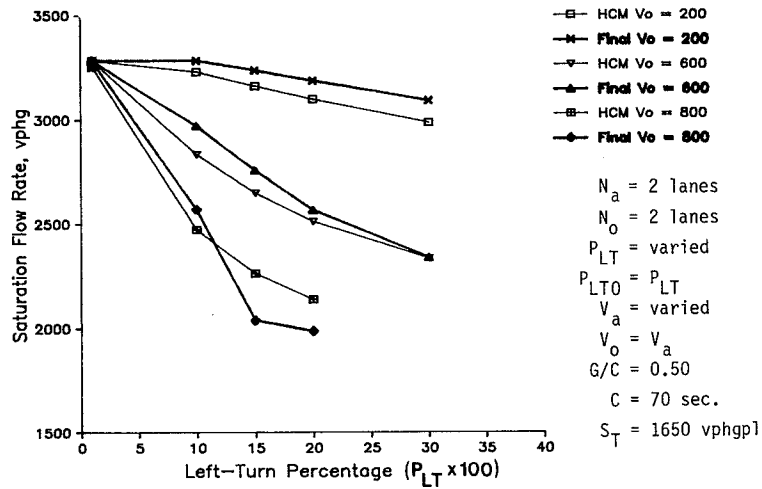


FIGURE 6 Effects of left-turn percentage and entering volume.

pendent variable, while the dependent variable is represented as a shaded region. In general, these variables were chosen to illustrate their interrelated effect on capacity and lane use. As might be expected, opposing and approach volumes were found to have a significant impact on intersection capacity. Similarly, opposing volume and approach left-turn percentage were found to have the greatest influence on approach lane use.

**Results of Specific Analyses**

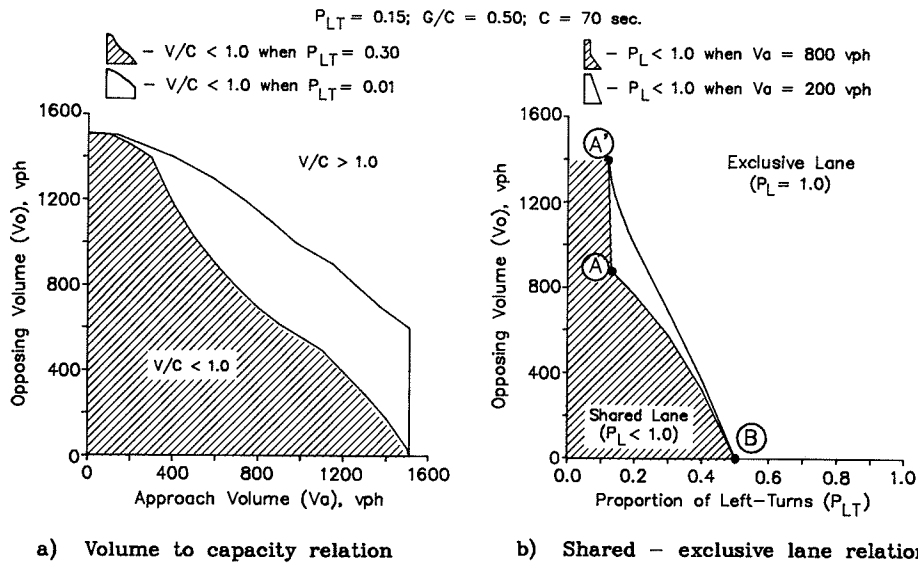
Using the modified HCM methodology, the volume-to-capacity ( $v/c$ ) and shared-exclusive lane relationships shown in figure 7 were developed. The sloping lines shown on each figure represent the unique combination of independent variables such that either the  $v/c$  ratio or the proportion of left-turns in the left-lane ( $P_L$ ) would equal 1.0. These threshold values were selected because they describe a boundary between operational states (*i.e.*, over-under capacity or shared-exclusive lane use).

As suggested by figure 7a, a range of maximum approach

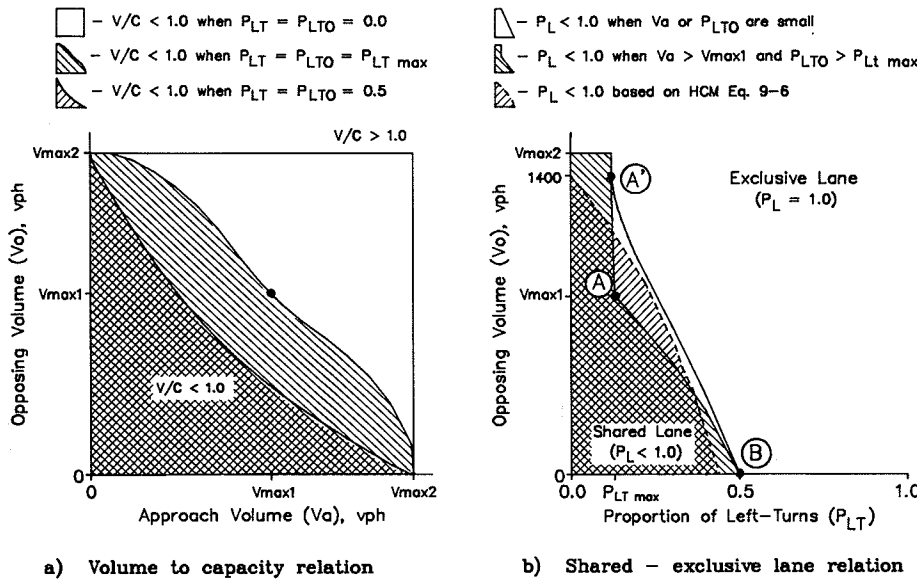
and opposing volume combinations was found such that one of the opposed approaches operated at its capacity. Obviously, any combination of approach volumes that had the same proportion of left-turns and intersected below this threshold should experience a  $v/c$  ratio less than 1.0.

Two capacity thresholds ( $P_{LT} = 0.01$  and  $0.30$ ) are shown in figure 7a. For the case where  $P_{LT}$  equaled  $0.30$ , it was found that the combinations of  $V_o$  and  $V_a$  that produced a  $v/c$  equal to  $1.0$  were less than those found when  $P_{LT}$  equaled  $0.01$ . This trend is intuitively reasonable considering the degrading effect that left-turn vehicles have on the capacity of an approach.

Figure 7b illustrates the effect of left-turn proportion ( $P_{LT}$ ) and opposing volume ( $V_o$ ) on lane usage. The line A'-A-B identifies the threshold combinations of  $P_{LT}$  and  $V_o$  that would cause the inner lane to operate an exclusive, or de facto, left-turn lane for an approach volume of  $800$  vph. Likewise, the line A'-B identifies the exclusive lane threshold for an approach volume of  $200$  vph. This threshold can be described as an equilibrium condition wherein the volume-to-capacity ratios of the through and left-turn movements are equal and the proportion of left-turn vehicles on the interior lane is  $1.0$ .



**FIGURE 7 Shared lane relationships: specific case.**



**FIGURE 8 Shared lane relationships: general case.**

It is reasonable to assume that the equilibrium condition describes the threshold where lane selection becomes movement specific on a shared-lane approach. The essence of this assumption is predicated on the inherent nature of motorists to base their lane selection on minimum travel time. This capacity equilibrium concept is embedded within the formulation of the shared lane analysis methodology.

As shown in Figure 7b, there was a left-turn proportion below which the approach always operated as a shared lane, regardless of opposing volume. For left-turn proportions above this minimum value, operation of the approach was found to be primarily a function of opposing volume. However, approach volume ( $V_a$ ) and opposing left-turn proportion ( $P_{LTO}$ ) were also found to affect the location of the equilibrium threshold (i.e., line A'-A-B vs. A'-B). In fact, the secondary effects of

$V_a$  and  $P_{LTO}$  explain the triangular region A'-A-B wherein the threshold condition was found to vary.

**Formulation of General Case**

Based on concepts introduced in the preceding section and supplemented with additional analyses, figure 8 was generated to illustrate the basic shared lane relationships. This figure is presented in a general form with only unique points and boundaries described. The intent of the general format was primarily to identify the sensitivity of shared lane use to changes in certain variables. However, it was assumed that the identification of all unique points or boundaries on each figure would aid others in applying these results to different timing

and geometric conditions. Each of these points and boundaries is defined below.

$V_{\max 2}$ —This is the capacity of two through lanes of traffic having no left-turn vehicles.

$$V_{\max 2} = (N * S_T * [G - l]) / C \quad (1)$$

where

- $N$  = number of lanes on the approach.
- $S_T$  = through saturation flow rate, vphgpl.
- $G$  = total green plus yellow time, sec.
- $l$  = lost time, assumed = 3.0 sec.
- $C$  = cycle length, sec.

Using the values from figure 7a (*i.e.*,  $G = 35$  sec.,  $S_T = 1650$  vphgpl,  $C = 70$  sec.,  $N = 2$ ),

$$V_{\max 2} = (2 * 1650 * [35 - 3]) / 70 = 1509 \text{ veh/hour.}$$

$V_{\max 1}$ —This represents the capacity of one lane of through traffic plus the number of left-turn vehicles clearing the intersection during the clearance interval (*i.e.*, sneakers).

$$V_{\max 1} = [V_{\max 2} * (N - l) / N] + [S_n * 3600 / C] \quad (2)$$

where

- $V_{\max 2}$  = maximum through lane capacity for  $N$  lanes, vph.
- $N$  = number of lanes on the approach.
- $S_n$  = maximum number of left-turns during the change interval, assumed equal to two under maximum volume conditions.
- $C$  = cycle length, sec.

For this example,  $V_{\max 1}$  can be calculated as

$$\begin{aligned} V_{\max 1} &= [1509 * (2 - 1) / 2] + [2 * 3600 / 70] \\ &= 857 \text{ veh/hour.} \end{aligned}$$

$P_{LT\max}$ —This value represents the threshold proportion of left-turn vehicles. Any proportion of left-turn vehicles less than this value would theoretically guarantee shared lane operation.

$$P_{LT\max} = (S_n * 3600 / C) / V_{\max 1} \quad (3)$$

where

- $V_{\max 1}$  = maximum capacity of one lane plus sneakers, vph.
- $S_n$  = number of left-turns during the change interval.
- $C$  = cycle length, sec.

For this example,

$$P_{LT\max} = (2.0 * 3600 / 70) / 857 = 0.12.$$

$V_o = 1400$  vph—This constant is identified in figure 8 because it represents the opposing volume that would theoretically have insufficient gap size to permit the filtering of left-turn vehicles. Thus, when the opposing volume exceeds 1,400 vph, the only left-turn capacity available would be as sneakers during the clearance interval.

$P_{LT} = 0.50$ —This proportion represents the upper boundary value for shared-lane operation. Theoretically, any left-turn proportion exceeding 0.50 would have exclusive use of the inside lane.

In figure 8a, three separate capacity threshold lines are

shown. These lines represent three different combinations of left-turn proportion: 0.0,  $P_{LT\max}$ , and 0.50. Of these, 0.0 and 0.50 represent lower and upper boundaries, respectively, on the region of shared lane use. The  $P_{LT\max}$  line is also shown because it is related to the shared-exclusive lane threshold and because it can be located using  $V_{\max 1}$ .

The left-turn proportions described in figure 8a are equal for both the subject and opposing approaches. This "symmetric" situation was chosen for two reasons. The first reason was that it more nearly represented the typical intersection where both approaches have roughly the same left-turn proportions. The other reason was based on the results of several analyses, from which it appeared that the symmetric case represented a worst-case combination. Thus, with respect to figure 8a, any volume combination that intersects below the capacity threshold (*i.e.*,  $v/c < 1.0$ ) for a given symmetric left-turn proportion should also be under capacity when one approach has a lesser left-turn proportion. For example, if the proportion of left turns on one approach was 0.05 while the proportion of left-turns on the opposing approach was 0.15 and their volume combination intersected below the  $P_{LT} = P_{LTO} = 0.15$  threshold, then it could be concluded that both approaches will operate at a  $v/c$  of less than 1.0. In summary, the symmetric left-turn proportion threshold should conservatively predict the under-capacity situation.

When considering figure 8a, it must be remembered that only one left-turn proportion line can describe the capacity threshold. The line chosen should equal or exceed the larger left-turn proportion of either the subject approach or its opposing approach. One of the three threshold lines shown in figure 8a could be used as the capacity threshold or another threshold could be located by interpolation. In recognition of the limitation of the shared lane methodology, the  $P_{LT} = P_{LTO} = 0.50$  threshold probably represents a practical boundary for conservatively estimating the volume-to-capacity condition of two, two-lane, opposing approaches.

Figure 8b represents the general relation between the approach left-turn percentage ( $P_{LT}$ ) and the volume opposing it ( $V_o$ ). In particular, the line A'-A-B represents the boundary between shared and exclusive lane operation on the subject approach. If the approach left-turn proportion and opposing volume combination intersect below the line A'-A-B, then the subject approach should operate as a shared lane. If the intersection is beyond line A'-B, then the approach should operate as an exclusive lane.

The triangular region A'-A-B bounds the combinations of  $P_{LT}$  and  $V_o$  that may or may not cause an exclusive lane operation. The location of the exact threshold boundary between point B and a point on line A'-A is a function of approach volume ( $V_a$ ) and opposing left-turn proportion ( $P_{LTO}$ ). The reason for this secondary influence is somewhat complex, but is based primarily on the capacity of period 2. In those situations where either the approach volume ( $V_a$ ) or the opposing left-turn proportion ( $P_{LTO}$ ) is light, there is a maximum likelihood of unused green time ( $g_u$ ). This usually results in a maximum left turn capacity. Conversely, if both  $V_a$  and  $P_{LTO}$  are large, then it is probable that  $g_u$  is small. Obviously if  $g_u$  equals zero, then period 2 capacity is also zero.

Based on the results of the preceding examination, it was possible to make the following generalization: If the left-turn proportion ( $P_{LT}$ ) and volume ( $V_a$ ) of an approach are less than  $P_{LT\max}$  and  $V_{\max 1}$ , respectively, then that approach will

operate below its capacity and its inner lane will be shared by left and through traffic.

In those cases where  $P_{LT}$  and  $V_a$  are greater than  $P_{LTmax}$  and  $V_{max1}$ , but less than 0.50 and  $V_{max2}$ , respectively, the operation of the approach must be determined using figures 8a and 8b or by analysis using the HCM methodology (*i.e.*, supplemental worksheet).

### Comparison With HCM De Facto Lane Procedure

HCM equation 9-6 (5, p. 9-9) describes a simple procedure for determining the operation of a shared lane approach. When the appropriate volume conditions are satisfied, the inside lane can be considered a de facto left-turn lane, and the HCM recommends it be analyzed as such.

The equation 9-6 procedure is compared, in figure 8b, with the threshold curves previously described. As indicated by the dashed curve, there is a general agreement in both shape and orientation. This agreement is particularly good for those situations having low to moderate approach volumes or low opposing left-turn proportions. However, it is also apparent from figure 8 that there are some combinations of opposing volume and left-turn proportion where the simplified procedure could incorrectly predict exclusive or shared lane operation.

### CONCLUSIONS

Based on a comparison of the original (HCM) approach and the modified, iterative approach, it was concluded that the saturation flow rate calculated using the HCM methodology would almost always be less than the final, convergence flow rate. In general, the HCM solution was found to be within 5% of the final saturation flow rate.

As a result of the examination of the shared lane sensitivity, several tendencies were noted. One of the more obvious trends observed was the dramatic reduction in saturation flow rate with small increases in the proportion of left-turns. Conversely, the sensitivity of shared lane operations to the magnitude of cycle length or  $G/C$  ratio was slight.

The combined effects of the study findings were incorporated into figure 8. Using this figure, it is possible to describe,

or predict, the nature of any shared lane operation. By inspection of figure 8, the following generalities were formulated:

1. Only when left-turn percentage is very small does approach capacity near that of two through lanes.
2. When there is a moderate number of left-turn vehicles, the effective number of lanes are reduced from four to roughly two. In other words, the combined maximum volume of two opposing, two-lane approaches, both having a moderate amount of left-turn vehicles, can be conservatively estimated as equal to about 90% of the capacity of two through lanes.
3. If the left-turn percentage and volume of an approach are less than  $P_{LTmax}$  and  $V_{max1}$ , respectively, then that approach will operate below its capacity, and its inner lane will be shared by left and through vehicles.

The results of this research were based entirely on the assumption of reasonableness of the HCM methodology. Furthermore, these results are limited by the assumptions imbedded in the formulation of the HCM methodology. In recognition of these limitations, it is recommended that field studies be conducted to verify the reasonableness of the HCM methodology and, thereby, the results of this paper.

### REFERENCES

1. J.H. Shortreed. Left-Turn Equivalencies For Opposed, Shared, Left-Turn Lanes. In *Transportation Research Record 971*, TRB, National Research Council, Washington, D.C., 1984, pp. 14-20.
2. W.R. McShane and E.B. Lieberman. Service Rates of Mixed Traffic on the Far Left Lane of an Approach. In *Transportation Research Record 772*, TRB, National Research Council, Washington, D.C., 1980, pp. 18-23.
3. D.B. Fambro, C.J. Messer, and D.A. Andersen. Estimation of Unprotected Left-Turn Capacity at Signalized Intersections. In *Transportation Research Record 644*, TRB, National Research Council, Washington, D.C., 1977, pp. 113-119.
4. K. Bang. Swedish Capacity Manual—Part 3. Capacity of Signalized Intersections. In *Transportation Research Record 667*, TRB, National Research Council, Washington, D.C., 1978, pp. 11-21.
5. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.

---

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

# Estimation of Independence of Vehicle Arrivals at Signalized Intersections: A Modelling Methodology

GANG-LEN CHANG AND JAMES C. WILLIAMS

The assumption of independent vehicle arrivals at traffic signals, such as that in the Poisson distribution, has been widely used for modelling delay at urban intersections. The degree of correlation among vehicles determines whether this convenient assumption of independence is realistic. Collection of vehicle headways seems to be the only well-known method by which an estimate of their autocorrelation can be found. This, however, is a tedious and time-consuming task. To cope with this problem, an effective methodology using a discrete model for estimating the degree of interaction among vehicles under given traffic and geometric conditions is proposed in the present study. As an example of this technique's usefulness, a model is proposed and estimated, using information collected from 40 locations. Preliminary results appear to confirm the strength and applicability of the proposed method even though the developed model is constrained by the limited available data.

One of the prerequisites for the effective design and evaluation of the operation of traffic signals is to accurately estimate the delay incurred by traffic passing through the signalized intersection. The use of delay in determining intersection level of service in the 1985 *Highway Capacity Manual* (1) highlights the need to accurately estimate delay. Clayton was one of the first to make an attempt in this respect (2); he proposed a model to calculate the delay at fixed-time signals, where vehicle arrivals at and departures from a signalized intersection were presumed to be at strictly regular intervals. Winsten (3) revised this model, using a more realistic distribution (binomial) to simulate the pattern of arrivals. Webster (4) and Newell (5) have contributed delay formulas using a similar methodology but a different distribution (Poisson) for the arrival of vehicles. Subsequent models by Miller (6, 7) and Newell (8) have incorporated the variance-to-mean ratio of the vehicle arrivals to accommodate arrival distributions other than the Poisson, and in Hutchinson's numerical comparisons of various delay equations this term was added to Webster's model (9). Because little significant progress has occurred in this fundamental aspect over the past two decades, the delay formulas developed by Webster, Miller, and Newell still predominate in practice (particularly Webster's delay equation), and are incorporated in many popular traffic computer packages, such as SOAP (10) and PASSER-II (11).

A common feature of the above-mentioned delay formulas is that the arrival pattern of vehicles at intersections is pre-

sumed to be an independent Markov process. The average delay and queue length are then estimated based on the given flow rate and distribution, such as the Poisson or binomial. The explicit assumption of independence among arrivals of vehicles, as embedded in the basic properties of the Poisson distribution, is convenient for deriving the desired performance indicators such as the average delay, maximum delay, and average queue, and indeed provides a reasonable approximation of reality as long as the traffic is light. It is, however, obviously inconsistent with what can be observed at highly congested intersections where vehicles significantly interact with others in the arriving flows.

As is well recognized, to characterize the complex traffic patterns at the desired level of accuracy is a difficult yet essential task that enables the model to possess realistic features. The uniform distribution, as used in these well-accepted delay models, generally provides a good representation of departure distributions, since queued vehicles at the beginning of green time are usually discharged at a more or less constant rate. The Poisson or binomial distributions, however, cannot capture the possible interaction between arriving vehicles that may vary with the degree of congestion, driver behavior, physical features between adjacent intersections, speed limit, and so on. In very light traffic, it seems reasonable to expect that vehicles will arrive independently and follow a Poisson process at intersections. The degree of independence decreases as the degree of interference among consecutively arriving vehicles, resulting from congestion and other factors, increases. A well-known phenomenon is the car-following relationship (12) that describes the action-response effect between the leading and following vehicles. As long as such interrelationships are developed in the traffic flows, the assumption of independence among arrivals is obviously no longer realistic, and will inevitably lead to a biased estimate of the degree of delay and the other performance measures, such as queue length.

The determination of the distribution of vehicle arrivals, however, is a tedious and time-consuming process, necessitating the collection of vehicle headways or, at the very least, vehicle arrivals in consecutive time periods (of 30 sec or less). Hutchinson's numerical work (9) shows that the commonly used delay equations show little difference when vehicle arrivals are Poisson distributed; otherwise, significant differences exist among models. In addition, preliminary results of a data collection effort by the authors indicate significant differences between estimated delay from independent and nonindependent arrivals (13).



While independent observations are not necessarily Poisson distributed, this distribution is typically assumed in most traffic-related studies if vehicle arrivals are independent. Therefore, it would be useful to have a simple test by which the independence of vehicle arrivals for specific traffic and geometric conditions could be determined without collecting vehicle headways. Of course, if the test indicates nonindependent arrivals, assumption of the Poisson distribution would not be valid, and further field studies and statistical tests would be necessary to select an appropriate distribution for the proper application of existing delay models.

A general model is presented in the next section that can be used to estimate the probability of independent vehicle arrivals to determine whether the commonly used Poisson assumption is applicable. Geometric features and traffic factors likely to affect the degree of interaction are included in this conceptual model. Techniques used to determine an appropriate specific model are also briefly discussed. An example of an application of this model is given in the third section; it is based on data collected from 40 intersections. In the final section, some conclusions and directions for further research are presented.

#### MODELLING CONCEPT AND ESTIMATION METHODOLOGY

A conceptual modelling system is presented in this section. The system relates the key features of a traffic system to its service load from which the degree of interaction among vehicles in the system can be estimated. This problem considers a traffic system consisting of an urban road section with  $N$  traffic lanes connecting two signalized intersections that are not interconnected, shown in Figure 1. Interconnected and actuated signals are not taken into account in this model, as each would require additional factors than those considered here. The goal, given the geometric factors and traffic conditions of such a road section, is to estimate the probability that vehicle arrivals at the downstream location are independent.

A single direction of a traffic system, such as that in Figure 1, can be analogized with a one-way channel with a unique entrance and exit at each end. Every vehicle entering the system, from the microscopic perspective, can then be viewed as a particle following a predetermined path (the available lanes) to pass through the channel. As such, whether the interference among particles in the channel is significant or not apparently depends on the channel's key physical features (e.g., length, number of paths), the number of particles (or the flow of particles), and their characteristics. Similarly, interaction among vehicles in the sort of traffic system shown in Figure 1 may vary with factors associated with the road section's physical features and the traffic flow characteristics.

More specifically, the system's key features primarily determine the available space for the flows. This space can be represented by

$$SA_i = f_1 (L_i, N_i, G_i, O_i) + \zeta_i \quad (1)$$

where

$SA_i$  = the amount of space available for traffic flows;  
 $L_i$  = the road section length, a major factor in the degree

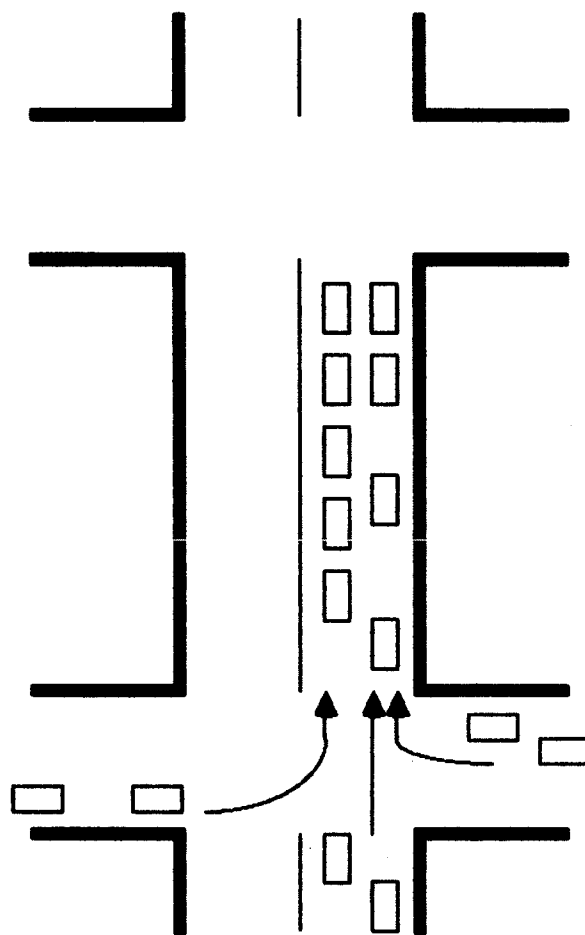


FIGURE 1 Graphical representation of two non-interconnected traffic signals.

of platoon dispersion from the upstream intersection;

$N_i$  = the number of available lanes;  
 $G_i$  = the road section grade;  
 $O_i$  = other associated factors; and  
 $\zeta_i$  = a random variable used to capture the effect of unobserved factors.

The subscript  $i$  identifies a particular direction of a road section.

On the other hand, volume, average speed, concentration, and driver behavior can characterize the roadway space needed to provide the independent-arrival environment. This can be stated as

$$CS_i = f_2 (S_i, Q_i, K_i, DB_i, OF_i) + \varepsilon_i \quad (2)$$

where

$CS_i$  = the critical amount of space needed for independent arrivals at the downstream intersection;  
 $S_i$  = average vehicle speed;  
 $Q_i$  = the flow (or volume);  
 $K_i$  = the average concentration;  
 $DB_i$  = driver behavior, often characterized by various indicators;  
 $OF_i$  = other associated factors; and

$\varepsilon_i$  = random variable used to capture unobserved associated factors.

As before, the subscript  $i$  identifies a particular direction of a road section.

As such, it can be expected that the interaction among arriving vehicles in road section  $i$  may occur if  $CS_i$  is greater than  $SA_i$ . Considering the uncertainties arising from unobservable factors, this statement can further be elaborated as a probabilistic formulation. If  $X_i$  and  $Y_i$  denote the vectors of the observable explanatory variables for  $SA_i$  and  $CS_i$ , respectively, as shown in Equations 1 and 2, and if  $A_i$  and  $B_i$  represent vectors of parameters associated with the variables in  $X_i$  and  $Y_i$ , respectively, this probabilistic formulation is as follows:

$$\begin{aligned} & \text{Prob [having independent vehicle arrivals in road section } i | X_i, Y_i] \\ &= \text{Prob [} SA_i > CS_i] = \text{Prob [} f_1(A_i, X_i) \\ & \quad + \zeta_i > f_2(B_i, Y_i) + \varepsilon_i] \\ &= \text{Prob [} f_1(A_i, X_i) - f_2(B_i, Y_i) > \varepsilon_i - \zeta_i] \end{aligned} \quad (3)$$

Accordingly, given adequate observations (varying with the number of unknown parameters) and presumed properties of the error terms, estimation of parameters can be carried out by using the maximum likelihood method. Note that the specifications for  $SA_i$  and  $CS_i$  (equations 1 and 2) vary with the available information, measurable key factors, and their interrelationships. Many formal statistical procedures, such as the likelihood ratio test, the Lagrangian multiplier test, and tests of non-nested hypotheses are available for specification testing (14-16).

#### Prediction of the Probability of Vehicle Interaction Leading to Independent Arrivals

As is well recognized, the methodology for estimating parameters of a specified model varies with the presumed properties of the error terms. A description follows of two commonly used econometric approaches (binary logit and probit models) that can be applied for the estimation of equation 3. A detailed discussion of their statistical features is, however, not within the scope of this paper and is available elsewhere (17).

#### Binary Probit Model

Assume  $\varepsilon_i$  and  $\zeta_i$  follow normal distributions with zero means, a covariance of  $\sigma_{12}$  and variances of  $\sigma_1^2$  and  $\sigma_2^2$ , respectively. Accordingly,  $\varepsilon_i - \zeta_i$  is also normally distributed with zero mean, but with a variance  $\sigma^2 (= \sigma_1^2 + \sigma_2^2 - 2\sigma_{12})$ . The probability of independent vehicle arrivals can then be solved as follows:

$$\begin{aligned} & \text{Prob [independent vehicle arrivals]} \\ &= \text{Prob [} f_1(A_i, X_i) - f_2(B_i, Y_i) < \zeta_i - \varepsilon_i] \\ &= \Phi [f_1(A_i, X_i) - f_2(B_i, Y_i) / \sigma] \quad \sigma > 0 \end{aligned} \quad (4)$$

where  $\Phi$  denotes the the standardized cumulative normal distribution.

#### Binary Logit Model

Another commonly used technique is to assume that  $\varepsilon_i^*$  ( $= \zeta_i - \varepsilon_i$ ) is logistically distributed (is Gumbel distributed), with a cumulative distribution:

$$F(\varepsilon_i^*) = 1 / (1 + \exp(-u \cdot \varepsilon_i^*)) \quad (5)$$

$$u > 0, -\infty < \varepsilon_i^* < \infty$$

where  $u$  is a positive scale parameter. This distribution approximates the normal distribution (as used in the probit model) quite well, but is much more convenient in terms of analytical computation. Under this assumption, the probability of independent vehicle arrivals is given by the following expression:

$$\begin{aligned} & \text{Prob [independent vehicle arrivals]} \\ &= \exp(u \cdot f_1(A_i, X_i)) u \\ & \quad \div [\exp(u \cdot f_1(A_i, X_i)) + \exp(u \cdot f_2(B_i, Y_i))] \end{aligned} \quad (6)$$

Note that for convenience, but without loss of accuracy, the scale parameter,  $u$ , is generally assumed to equal 1.

#### Estimation of Model Parameters

Let each road section  $i$ , with associated key attributes as described previously, be viewed as one observation, then, given  $N$  observations, the likelihood function of the parameters in vectors  $A$  and  $B$  (Equation 3) can be constructed as follows (18):

$$L^*(A, B) = \prod_{i=1}^N [P_i(a)^{\delta_i} \cdot (1 - P_i(a))^{1-\delta_i}] \quad (7)$$

where

$$\begin{aligned} P_i(a) &= \text{Prob [independent vehicle arrivals in road section } i] \\ \delta_i &= 1 \quad \text{if the independent vehicle arrivals were observed in road section } i \\ \delta_i &= 0 \quad \text{otherwise} \end{aligned}$$

To facilitate computation, Equation 7 is often rewritten in the following logarithmic form, denoted as  $L$ :

$$\begin{aligned} L(\beta_1, \dots, \beta_k) \\ &= \sum_{i=1}^N [\delta_i \ln P_i(a) + (1 - \delta_i) \ln (1 - P_i(a))] \end{aligned} \quad (8)$$

By differentiating  $L$  with respect to each of the  $\beta$ s (parameters in vectors  $A_i$  and  $B_i$ ) and setting the partial derivatives equal to zero, parameters satisfying  $\max L(\beta_1, \beta_2, \dots, \beta_k)$  can be obtained. In many cases of practical interest it has been proven that the likelihood function (Equation 7) is globally concave and is, therefore, unique if a solution to the first-order conditions exists.

#### ILLUSTRATIVE EXAMPLE

This section presents an example that illustrates the modelling procedures and methods used for specification testing. The

data set for parameter estimation consists of 40 noninterconnected signalized intersections, all located in the Salt Lake City metropolitan area in Utah. Due to limited resources, our data collection focused on those variables relatively inexpensive to collect, yet are critical to the model development. The available data and proposed formulation for a model that will evaluate the degree of correlation, or interaction, among arriving vehicles are stated below. Each variable corresponds to a particular road section  $i$  and the traffic approaching the particular intersection in question. The available key variables are as follows:

1. Length of road section  $i$  ( $L_i$ ), measured from the stop line of the selected intersection back to the nearest signalized intersection upstream;
2. Number of lanes, omitting exclusive turn lanes ( $N_i$ );
3. Average vehicle speed ( $S_i$ );
4. Flow rate to the selected intersection ( $Q_i$ ); and
5. Vehicle headways, taken separately for each lane (if  $N_i > 1$ ), between each successive vehicle.

The vehicle arrivals at the downstream intersection are assumed to be independent if the vehicle headways are not significantly correlated. Estimation of the correlation was carried out by analyzing the vehicle headway time series collected over a 30-min interval in the selected road section with a general ARIMA model (19).

### Model Specification

Given the available information above, several plausible specifications for Equation 1 and 2 were proposed and examined based on the estimated results. Primary criteria used to carry out the comparisons are the  $t$ -statistic, likelihood ratio index, and physical implications of the estimated parameters. Of the specifications tested, the one providing the best fit (highest likelihood ratio index) and having a reasonable physical meaning reflected by the proper parameter signs is

$$\text{Amount of space available} = SA_i = N_i \cdot L_i \quad (9)$$

Critical amount of space needed =

$$CS_i = a_1(Q_i)^{a_2} + a_3(S_i)^{a_4} + \varepsilon_i \quad (10)$$

where  $a_1$ ,  $a_2$ ,  $a_3$ , and  $a_4$  are model parameters. The notion embedded in this formulation is that under the given environment (as characterized by  $SA_i$ ) the flow rate and speed are critical factors that contribute to the formation and the degree of interaction among vehicles. More specifically, the vehicle arrivals along road section  $i$  may be significantly correlated if  $SA_i$  is less than  $CS_i$ . The probability of existing dependent arrivals of vehicles is as follows:

$$\begin{aligned} \text{Prob}[SA_i < CS_i] &= \text{Prob}[N_i \cdot L_i < a_1(Q_i)^{a_2} + a_3(S_i)^{a_4} + \varepsilon_i] \\ &= \text{Prob}[(N_i \cdot L_i) - a_1(Q_i)^{a_2} - a_3(S_i)^{a_4} < \varepsilon_i] \end{aligned} \quad (11)$$

This proposed formulation cannot be considered the standard model for predicting the independence among vehicle arrivals because of limitations of the collected data; it simply serves as an illustrative example. For instance, the value of  $SA_i$ , in

practice, depends on the number of lanes and the length of the road section and varies with the system's other geometric characteristics, such as grade. In addition, information regarding drivers' behavior or risk attitude that could be one of the critical factors was not collected due to the prohibitive cost of data acquisition and classification. Also, the arrival pattern depends, to some extent, on the departure pattern from the upstream signalized intersection; and for the same flow rate, many possible departure patterns exist. Although this is not directly modelled, the degree of variation of the arrival flow patterns is constrained by the distance between the signalized intersections, which has been incorporated in the model.

### Parameter Estimation and Implications

Because the variable  $CS_i$  is not directly observable, the parameters in Equations 9 and 10 should be estimated using discrete methods, such as either the logit or probit model, as presented in the second section. Justification of a proper specification for a discrete model, as for regression, is often carried out by examination of (1) the exhibited sign of estimated parameters, (2) the asymptotic standard error (or  $t$ -statistic), and (3) a goodness-of-fit index such as  $\rho^2$ , the likelihood ratio index. This index measures the fraction of an initial log likelihood value explained by the presented model and is defined as  $[1 - L(\beta)/L(0)]$ , where  $L(0)$  and  $L(\beta)$  denote the initial value (when all the parameters are zero) and the maximum value (at convergence) of the log likelihood function, respectively. The  $\rho^2$  is analogous to  $R^2$  used in the regression models, and must lie between zero and one for a binary discrete model. It is particularly useful in comparing alternative specifications developed on the same data set. A more in-depth discussion regarding the statistical properties of the likelihood ratio index is available elsewhere (18).

Table 1 summarizes the results of the parameter estimation for the illustrative model, as presented in Equations 9 and 10 using the discrete logit model. As expected, all parameters exhibit positive signs, except  $a_3$ . The implications are that, within a given road section, as partly characterized by  $SA_i$ , an increase in the flow will significantly increase dependent vehicle arrivals. On the other hand, a facility with better geometrics generally allows a higher speed under a given traffic volume, and thus results in less interaction among vehicles as reflected by the negative sign of parameter  $a_3$ .

Turning to the  $t$ -statistics as shown in Table 1, it can be noted, using a 0.95 statistical confidence interval, that both parameters  $a_2$  and  $a_4$  are not significantly different from one. As such, the model represented by Equations 9 and 10 was re-estimated with a simple linear specification presuming that parameters  $a_2$  and  $a_4$  are known to equal one. Estimated results of the simplified specification present similar information and are reported in Table 2.

In addition to the commonly used  $t$ -statistics, Tables 1 and 2 present the result of the log likelihood ratio test, defined as  $-2[L(0) - L(\beta)]$ , a statistic used to test the null hypothesis that all parameters are zero. This statistic is asymptotically distributed with an  $\chi^2$  distribution with  $K$  degrees of freedom, where  $K$  is the number of parameters. In this example the value of the log likelihood ratio is 37.74 (see Table 2), which indicates the null hypothesis can be rejected (the  $\chi^2$  value at the 0.05 level of significance is 9.49).

TABLE 1 ESTIMATION RESULTS OF THE ILLUSTRATIVE MODEL

Parameter	Associated Variable	Estimated Value	Standard Error	$t$ Statistic
$a_1$	Total approach volume: $Q_i$	0.00865	0.0027	3.20
$a_2$	Total approach volume: $Q_i$	0.96377	0.0357	1.01 <sup>a</sup>
$a_3$	Average speed: $S_i$	-0.17642	0.0618	2.85
$a_4$	Average speed: $S_i$	0.91425	0.0733	1.17 <sup>a</sup>

NOTE: Number of observations = 40;  $L(0) = -27.726$ ;  $L(\beta) = -8.114$ ;  $-2(L(0) - L(\beta)) = 31.224$ ;  $\rho^2 = 1 - L(\beta)/L(0) = 0.707$ .

<sup>a</sup>Not significant at 95 percent confidence level ( $\alpha_i = 1$  tested).

TABLE 2 REESTIMATED RESULTS OF THE ILLUSTRATIVE MODEL

Parameter	Associated Variable	Estimated Value	Standard Error	$t$ Statistic
$a_1$	Total approach volume: $Q_i$	0.00911	0.0038	2.372
$a_3$	Average speed: $S_i$	-0.17518	0.0649	2.698

Number of observations = 40;  $L(0) = -27.726$ ;  $L(\beta) = -8.856$ ;  $-2(L(0) - L(\beta)) = 37.74$ ;  $\rho^2 = 1 - L(\beta)/L(0) = 0.681$ .

With respect to the goodness of fit, the likelihood ratio index,  $\rho^2$ , generally serves as a good indicator for discrete models, as described above, although other rigorous, yet complex, statistical tests are available (14-16). Owing to the inevitable involvement of various behavioral or unobservable factors, the likelihood ratio index in most discrete models, which may be considered to be well specified, often cannot achieve as high a value as  $R^2$  in successful regressions. In fact, the illustrative example with  $\rho^2$  near 0.7, as shown in Tables 1 and 2, will generally have a reasonably good specification.

### Model Application

Instead of spending time and money in collecting vehicle headways, a traffic engineer could use a model such as this (Equations 9 and 10) to predict the probability of dependent vehicle arrivals under the given traffic conditions and geometric characteristics. The following example illustrates the procedure for estimating this probability. If  $L_i = 400$  ft = 0.0758 mi,  $S_i = 15$  mi per hour,  $Q_i = 150$  vehicles per hr per lane, and  $N_i = 3$  lanes, then

$$SA_i = N_i \cdot L_i = (3)(0.0758) = 0.2273$$

$$\begin{aligned} CS_i &= a_1(Q_i) + a_3(S_i) \\ &= (0.00911)(150)(3) - (0.17518)(15) \\ &= 1.4718 \end{aligned}$$

Accordingly, the probability of incurring correlated arriving vehicles under the above traffic conditions is

$$\begin{aligned} \text{Prob [nonindependent vehicle arrivals} \mid L_i, S_i, N_i, Q_i] \\ &= \exp(CS_i) / [\exp(CS_i) + \exp(SA_i)] \\ &= 4.3571 / (4.3571 + 1.4718) = 0.77 \end{aligned}$$

In other words, one can conclude that the traffic flows in this

road section are highly correlated about 77 percent of the time; thus, the commonly used Poisson distribution is not valid under this scenario. This result should not be unexpected, given the volume and the close signal spacing used for this example. Other distributions that can provide for dependent and independent arrivals should be considered.

### CONCLUSION

The present study has introduced an effective yet economic approach to estimate the degree of correlation among arriving vehicles under given conditions and geometric characteristics. With the proposed technique, traffic professionals can easily determine if the existing delay formulas and other traffic models based on the Poisson distribution are applicable. In particular, the results of this test can indicate when the Poisson assumption may be used.

As the primary focus of the paper is to introduce the modeling methodology and its application, only the simplest case comprising two noninterconnected intersections is considered. This model can be extended to more complex traffic systems if appropriate model variables are included in the specification, e.g., a variable indicating whether a platoon arrives during the red or green time would be necessary when formulating a similar model for a progressive signal system.

The formulation proposed in the example serves only as an illustration. More complete information, as described in previous sections, must be collected, and rigorous statistical procedures applied for the testing of the model specification to determine the most appropriate formulation.

### ACKNOWLEDGMENT

The data used in the illustrative example was collected by Tsai Ying-Chieh.

## REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
2. A. J. H. Clayton. Road Traffic Calculations. *Journal of the Institution of Civil Engineers*, Vol. 16, 1940.
3. M. Beckman, C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation*, Yale University Press, 1956.
4. F. V. Webster. Traffic Signal Settings. *Road Research Technical Paper Nr. 39*, 1958.
5. G. F. Newell. Queues for a Fixed-Cycle Traffic Light. *Annual Mathematical Statistics*, Vol. 31, 1960.
6. A. J. Miller. Settings for Fixed-Cycle Traffic Signals. *Operational Research Quarterly*, Vol. 14, 1963.
7. A. J. Miller. The Capacity of Signalized Intersections in Australia. *Australian Road Research Board Bulletin Nr. 3*, 1968.
8. G. F. Newell. Approximation Methods for Queues with Application to the Fixed-Cycle Traffic Light. *SIAM Review*, Vol. 7, 1965.
9. T. P. Hutchinson. Delay at a Fixed Time Traffic Signal, II: Numerical Comparisons of Some Theoretical Expressions. *Transportation Science*, Vol. 6, 1972.
10. *Signal Operations Analysis Package (SOAP)*. Implementation Package IP-79-9, five volumes. FHWA, U.S. Department of Transportation, 1979.
11. *Analysis of Reduced-Delay Optimization and Other Enhancements to PASSER II-80—PASSER II-84—Final Report*. Report No. TTI-2-18-83-375-1F. Texas Transportation Institute, 1984.
12. R. Herman and R. B. Potts. Single-Lane Traffic Theory and Experiment. *Proc., 1st Symposium on the Theory of Traffic Flow*, ed. R. Herman, Elsevier, 1961.
13. G.-L. Chang and J. C. Williams. Empirical Investigation of Theoretical Delay Models, working paper.
14. J. L. Horowitz. Testing Probabilistic Discrete Choice Models of Travel Demand by Comparing Predicted and Observed Aggregate Choice Shares. *Transportation Research B*, Vol. 19B, 1985.
15. T. S. Breusch and A. R. Pagan. A Simple Test for Heteroskedasticity and Random Coefficient Variation. *Econometrica*, Vol. 47, 1979.
16. J. L. Horowitz. Statistical Comparison of Non-Nested Probabilistic Discrete Choice Models. *Transportation Science*, Vol. 17, 1983.
17. C. F. Daganzo and L. Schoenfeld. *CHOMP User's Manual*. ITS Research Report UCB-ITS-RR-78-7. Institute of Transportation Studies, University of California, Berkeley, 1978.
18. M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis*. MIT Press, 1985.
19. G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

---

*The authors are solely responsible for the content of this paper.*

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

# WEAVSIM: A Microscopic Simulation Model of Freeway Weaving Sections

M. ZAREAN AND Z. A. NEMETH

---

**Intense lane-changing maneuvers at weaving sections create turbulences that often lead to congestion. The study of the dynamics of traffic flow at weaving sections thus has the potential to generate benefits. This paper describes a microscopic simulation model, WEAVSIM, developed for such studies. Freeway Data Collection for Studying Vehicle Interaction, a project of the FHWA, produced data sets at weaving sections and other problem areas to facilitate the study of freeway operations and the enhancement of freeway simulation models. Some of these data sets have been used in testing WEAVSIM. The application of the model was demonstrated by a small-scale simulation study of weaving sections.**

---

Weaving sections are those areas of a highway where two or more traffic streams, temporarily traveling in the same general direction, cross each other. Weaving sections are most frequently found at freeway interchanges. The lane-changing maneuvers performed by the weaving traffic introduce frictions and turbulence into the traffic flow that are not usually experienced on basic freeway sections. Therefore, weaving sections are often the cause of recurring congestion problems. In addition, weaving represents a traffic flow situation that is especially difficult to simulate.

An FHWA research study, entitled Freeway Data Collection for Studying Vehicle Interaction, was undertaken to develop a series of data sets on microscopic vehicular traffic flow for several types of potentially problematic geometric configurations (1). Several sets of data were collected at weaving sections. The development of the data sets involved digitizing vehicle positions from time-lapse aerial photographs. The data consists of a record of lateral and longitudinal positions of vehicles each second over a one-hour period. The data sets are being made available to researchers who are interested in studying a particular aspect of traffic flow. The data on microscopic vehicle movements were expected to be especially useful in enhancing freeway simulation models.

A comprehensive review of the literature on current freeway simulation models (2) reveals that although some general-purpose freeway corridor simulation models, such as INTRAS (3), could be utilized to some extent for the study of traffic operations in weaving sections, no specific purpose simulation model has been designed for a detailed study of weaving sections. This paper describes a microscopic simulation model, WEAVSIM, developed specifically for the study of the dynamics

of traffic flow at weaving sections. The data sets provided by FHWA were utilized in testing the model.

## THE SIMULATION MODEL

WEAVSIM is written in SIMSCRIPT II.5 simulation programming language. SIMSCRIPT II.5 is a free-format, English-like programming language which is readable and understandable even by a nonprogrammer who is primarily interested in the system under study and not necessarily in the computer programming. WEAVSIM is thoroughly commented for easy understanding of the logic. The execution time and computer memory requirements of the WEAVSIM program vary considerably with the input volumes. When implemented on AMDAHL/V8, the average real-time/CPU time ratio was approximately 30:1.

In WEAVSIM, vehicles are generated randomly at the system entry points. Each vehicle behaves as an individual entity having a set of attributes which control its progress through the system. These attributes are assigned either stochastically or deterministically. The model is based on a rational description of the behavior of vehicles as they proceed through weaving sections. At each one second of real time, all vehicles are processed through the system using a car-following algorithm, which governs longitudinal movements, and a lane-changing algorithm, which controls lateral movements. Results from various human factor studies have been utilized in the development of the logic. All vehicles are advanced through the system in accordance with their desired speed and destination, influenced by the immediate environment.

The car-following algorithm is a modified version of the so-called "failsafe" approach developed for INTRAS (2). This approach is based on a combination of the following three concepts:

1. A following vehicle always seeks a desired safe headway behind a lead vehicle, which is a function of vehicle speed, relative speed, and vehicle and driver type.
2. A following vehicle is able to avoid collision even when a lead vehicle undergoes the most extreme deceleration. This constraint is, however, relaxed during lane-changing maneuvers. Vehicles may accept potentially unsafe positions for a short period of time when engaged in the weaving maneuvers.
3. The desired safe headway is inversely proportional to the driver's maximum speed. This means that a fast driver will maintain a smaller lead-headway than a slow driver, assuming both are traveling at the same speed.

---

M. Zarean, Baskerville-Donovan Engineers, Inc., 316 South Baylen Street, Suite 300, Pensacola, Fla. 32501. Z. A. Nemeth, Department of Civil Engineering, The Ohio State University, Columbus, Ohio 43210.

The lane-changing algorithm moves vehicles from one lane to another by first establishing a desire or need for such a move and then searching for and accepting a suitable gap in the adjacent lane. The model assumes that as the ratio of the lane-weaving volume (weaving volume entering each lane) to the total weaving volume increases and, as the weaving vehicles move closer to the exit gore, vehicles become more willing to accept higher risk (i.e., shorter gaps) when engaged in lane-changing maneuvers. The lane changing logic also allows vehicles to look ahead of or behind the adjacent vehicle for appropriate gaps and, if needed, to adjust their speed to improve their position with respect to the available gap. The model performs two types of lane-changing: essential and non-essential. An essential lane-changing is performed by all weaving vehicles as they must change lanes to reach their desired destinations. A non-essential lane-changing is performed if the vehicle wishes to pass a slower vehicle.

In WEAVSIM the system starts empty and idle. Thus, it contends with an initiation bias prior to the steady-state conditions. Based on the observation of average delay and travel time, as a function of simulation time, a "warm-up" period of two hundred seconds was selected for the system to reach the steady-state condition. Buffer lengths are also provided at each end of the weaving section to dissipate the transient at the geometric boundaries. At the upstream end, four hundred feet of the simulated section are treated as the "warm-up" zone. At the downstream end, five hundred feet are used as the "cool-off" zone.

The input modeling of WEAVISM includes the following:

1. Interarrival headways of vehicles are generated using the inverse transformation method of the random deviate generation for a shifted negative exponential distribution with a minimum headway of one second. Interarrival headways are a function of lane volume.
2. Free-flow speeds are sampled randomly from a truncated normal distribution with mean and standard deviations of 60 and 10 mph, respectively.
3. Brake reaction times are generated stochastically from a gamma distribution with mean and standard deviations of 0.745 and 0.073 seconds, respectively.
4. Maximum deceleration/acceleration rates are assigned deterministically based on the vehicle type and speed.
5. Roadway parameters representing the geometry of the simulated section are directly specified.

The standard output of the model includes an echo of the input parameters and statistics on measures of performance describing the operational conditions of the weaving section. Vehicle trajectories, status listing, and a set of statistics on measures of performance are collected at user-specified time intervals. A graphical presentation of these statistics is an optional output of the model.

## TESTING

Testing of WEAVSIM involved comparison of the simulated observations with field data provided by FHWA. Data used for this comparison were collected at the Baltimore-Washington Parkway northbound at I-95 in the Washington, D.C.,

area and at the Harbor Freeway northbound between the Santa Monica Freeway (I-10) and Sixth Street in Los Angeles, California.

Intervals of five minutes were selected from the data sets for comparison. A SIMSCRIPT program was written to read the field data as if they were created by WEAVSIM. To manipulate and analyze the data, this program was properly linked to a statistical analysis system package. This enabled creation of outputs from field data with the same format as those produced by WEAVSIM. This facilitated an easy comparison of the two outputs. The input volume and traffic compositions were also obtained from the data for each time interval and used as input parameters for the simulation model. The geometrics of the model were also adjusted to represent that of the site at which the field data were collected. Although five-minute intervals of data were used, to eliminate possible randomness effect, simulation runs were made for a period of thirty minutes plus a two-hundred-second warm-up time. The outputs were then compared with those obtained from the field data.

The following traffic descriptive parameters were targeted for comparison:

- headway distributions,
- distributions of accepted gaps,
- merging point distributions,
- weaving and non-weaving speed distributions, and
- vehicle trajectories.

The Kolmogorov-Smirnov distribution free test was applied to compare the observed and simulated distributions. The paired *t*-test was conducted to compare mean values of the observed and simulated speeds and headways. Finally, the *F*-distribution was applied for the comparison of variances. A complete description of the simulation model, including its testing, is given elsewhere (3). Some examples of the comparison of the observed versus simulated data are shown in Tables 1, 2, 3, and 4. A sensitivity analysis of the output to variation in the input variables was also performed with satisfactory results.

In addition, model behavior was tested at extreme traffic conditions by introducing a severe speed disturbance to the leader of a platoon and observing the behavior of the following vehicles. Figures 1 and 2 show the introduction of the speed disturbance and the recovery behavior of the platoon.

The testing results indicate overall that the model reproduces behavior of the real-life system reasonably well.

## IMPLEMENTATION

Model implementation is illustrated with a simulation study. The objective of this study is to demonstrate the impact of upstream traffic conditions on some measures of performance (i.e., speed, delay) which describe the operational conditions of the weaving sections.

In one way or another, available analysis and design techniques developed for weaving sections consider the effect of the weaving section length and weaving and non-weaving flow on the operational conditions at these sites. Speeds of vehicles entering the weaving section directly affect the operational conditions within the weaving sections. At the same time, traffic

TABLE 1 COMPARISON OF SIMULATED VS. OBSERVED HEADWAY DISTRIBUTIONS AT THE HARBOR SITE

HEADWAY	lane 1		lane 2		lane 3	
	OBS.	SIM.	OBS.	SIM.	OBS.	SIM.
<= 1.0	0.13	0.07	0.17	0.10	0.21	0.13
<= 2.0	0.57	0.67	0.54	0.60	0.57	0.59
<= 3.0	0.81	0.83	0.78	0.82	0.77	0.81
<= 4.0	0.90	0.87	0.88	0.91	0.87	0.87
<= 5.0	0.94	0.90	0.93	0.94	0.92	0.90
<= 6.0	0.96	0.92	0.95	0.96	0.95	0.93
<= 7.0	0.97	0.94	0.96	0.97	0.97	0.94
<= 8.0	0.98	0.95	0.97	0.98	0.98	0.96
<= 9.0	0.98	0.96	0.97	0.98	0.99	0.96
<=10.0	0.98	0.97	0.97	0.99	0.99	0.97
<=11.0	0.98	0.98	0.98	0.99	0.99	0.98
<=12.0	0.99	0.99	0.99	0.99	1.00	0.98
<=13.0	1.00	1.00	1.00	1.00	1.00	0.99
<=14.0	1.00	1.00	1.00	1.00	1.00	0.99
<=15.0	1.00	1.00	1.00	1.00	1.00	1.00
no. of obs.	302	877	294	950	265	804
MEAN	2.27	2.50	2.39	2.32	2.33	2.68
STD. DEV.	2.44	2.62	2.30	2.19	2.91	2.95
J	0.060		0.070		0.080	
J(crt)	0.091		0.088		0.096	
t	-1.34		+0.47		-1.68	
t(crt)	-1.96	+1.96	-1.96	+1.96	-1.96	+1.96
DIFF(Appr. SE)	-.23(.17)		0.07(.15)		-.35(.21)	
F	0.933		1.050		0.989	
F(crt)	0.827	1.198	0.827	1.198	0.817	1.211

TABLE 2 COMPARISON OF SIMULATED VS. OBSERVED HEADWAY DISTRIBUTIONS AT THE BALTIMORE SITE

HEADWAY	lane 1		lane 2		lane 3	
	OBS.	SIM.	OBS.	SIM.	OBS.	SIM.
<= 1.0	0.32	0.26	0.16	0.12	0.28	0.16
<= 2.0	0.74	0.71	0.76	0.71	0.68	0.49
<= 3.0	0.90	0.82	0.88	0.87	0.86	0.66
<= 4.0	0.95	0.90	0.93	0.93	0.93	0.74
<= 5.0	0.97	0.94	0.95	0.96	0.96	0.80
<= 6.0	0.99	0.96	0.97	0.98	0.98	0.84
<= 7.0	0.99	0.97	0.98	0.98	0.99	0.88
<= 8.0	0.99	0.99	0.99	0.99	0.99	0.90
<= 9.0	1.00	0.99	0.99	0.99	1.00	0.93
<=10.0	1.00	1.00	1.00	1.00	1.00	0.94
<=11.0	1.00	1.00	1.00	1.00	1.00	0.96
<=12.0	1.00	1.00	1.00	1.00	1.00	0.97
<=13.0	1.00	1.00	1.00	1.00	1.00	0.99
<=14.0	1.00	1.00	1.00	1.00	1.00	1.00
<=15.0	1.00	1.00	1.00	1.00	1.00	1.00
* of obs.	158	480	192	544	85	286
MEAN	1.86	2.03	1.74	1.93	1.84	3.46
STD. DEV.	1.29	1.59	1.44	1.26	1.47	3.42
J	0.080		0.050		0.200	
J(crt)	0.125		0.114		0.168	
t	-1.22		-1.73		-4.40	
t(crt)	-1.96	+1.96	-1.96	+1.96	-1.96	+1.96
DIFF(Appr. SE)	-.17(.14)		-.19(.11)		-1.62(.37)	
F	0.805		1.147		0.433	
F(crt)	0.768	1.280	0.786	1.254	0.697	1.388



TABLE 3 COMPARISON OF SIMULATED VS. OBSERVED SPEED DISTRIBUTIONS AT THE HARBOR SITE

SPEED mph.	WEAVING		NONWEAVING	
	OBS.	SIM.	OBS.	SIM.
<= 15.0	0.00	0.00	0.00	0.00
<= 20.0	0.09	0.16	0.12	0.07
<= 25.0	0.42	0.56	0.46	0.35
<= 30.0	0.81	0.91	0.87	0.61
<= 35.0	0.98	1.00	1.00	0.94
<= 40.0	1.00	1.00	1.00	0.99
<= 45.0	1.00	1.00	1.00	1.00
<= 50.0	1.00	1.00	1.00	1.00
# of obs.	112	339	113	361
MEAN	26.07	24.29	25.44	27.88
STD. DEV.	4.59	4.06	4.61	5.52
DIFF(Appr. SE)	1.78(.46)		-2.44(.57)	
J	0.140		0.260	
J(crt)	0.148		0.147	
F	0.990		1.289	
F(crt)	0.724	1.347	0.719	1.358

TABLE 4 COMPARISON OF SIMULATED VS. OBSERVED SPEED DISTRIBUTIONS AT THE BALTIMORE SITE

SPEED mph.	WEAVING		NONWEAVING	
	OBS.	SIM.	OBS.	SIM.
<= 15.0	0.00	0.00	0.00	0.00
<= 20.0	0.00	0.01	0.00	0.01
<= 25.0	0.01	0.08	0.04	0.06
<= 30.0	0.23	0.27	0.12	0.17
<= 35.0	0.55	0.56	0.38	0.42
<= 40.0	0.73	0.84	0.76	0.87
<= 45.0	0.88	0.97	0.88	0.99
<= 50.0	0.96	1.00	0.98	1.00
<= 55.0	0.97	1.00	0.99	1.00
<= 60.0	1.00	1.00	1.00	1.00
<= 65.0	1.00	1.00	1.00	1.00
# of obs.	77	239	170	496
MEAN	35.92	33.87	43.70	39.90
STD. DEV.	7.20	6.05	7.03	5.36
DIFF(Appr. SE)	2.05(.83)		3.80(.51)	
J	0.090		0.110	
J(crt)	0.178		0.121	
F	1.013		1.310	
F(crt)	0.682	1.416	0.775	1.271

disturbances in the weaving section have an impact on arrival speeds of vehicles entering the weaving sections. These impacts, although generally recognized, have not been addressed by any of the current analysis and design procedures.

WEAVSIM is used here to demonstrate the impact of upstream traffic conditions, represented by speed, on operating conditions within the weaving section.

When two traffic streams merge, the arrival speeds of the two streams are not always compatible. This is the case when

the weaving section is formed by an on-ramp which is followed by an off-ramp with a continuous auxiliary lane between the ramps. The arrival speed of the ramp vehicles is influenced by the restrictions of the ramp geometrics and is usually lower than the speed of the freeway vehicles. This difference in arrival speeds directly influences the operational conditions of weaving sections. This study investigates the effect of the difference in arrival speeds of the two merging traffic streams on the speed and delay in the weaving section.

A set of four dependent variables which are indicators of the operational conditions and five independent variables which control the operation through weaving sections has been selected for this study. The dependent variables are weaving and non-weaving speeds and delays. Delay is computed as the difference between the ideal travel time at free-flow speed and the actual travel time for each vehicle. The independent variables are weaving section length, freeway volume, ramp volume, proportion of weaving vehicles to total volume, and difference in arrival speeds (SPDDIF). Three levels of each variable were selected. Using various combinations of the independent variables, a total of 243 experiments were performed. Each simulation experiment was run for a period of forty-five minutes plus a warm-up period of two hundred seconds. Using batch means, with batches of size fifteen minutes, three independent sets of output were obtained from each experiment. These outputs were then subjected to the following multiple regression analysis involving first and second order parameters, as well as one-way interactions:

$$MOP = \beta_0 + \sum_{i<1}^5 \beta_i X_i + \sum_{i<1}^4 \sum_{j<1}^5 \beta_{ij} X_i X_j + \sum_{i<1}^5 \alpha_i X_i^2$$

$$MOP = \text{Predicted measure of performance,}$$

$$\beta_0, \beta_i, \beta_{ij} = \text{Estimates of the constants, and}$$

$$X_i, X_j = \text{Input parameters.}$$

Finally, a stepwise regression approach was employed to derive the best fitting model for each individual measure of performance. The effect of the difference in arrival speeds was illustrated by computing the weaving and non-weaving speeds and delays for various values of this parameter using the regression models. The results are summarized in Table 5. The direct impact of difference in arrival speeds on weaving speeds and delays within the section is demonstrated, with more pronounced effect on weaving traffic than on non-weaving traffic. Values of the weaving and non-weaving speeds computed using the new *Highway Capacity Manual* (HCM) procedures (4) are also shown in this table. Although no statistical test was performed, the values predicted by the model seem to be compatible with those produced using the HCM procedures.

## SUMMARY

WEAVSIM, a microscopic simulation model of freeway weaving sections was developed, using SIMSCRIPT II.5 simulation programming language. It allows the simulation of traffic flow through a three-lane weaving section of any length and under varying traffic volume conditions. Operational conditions are represented by speed and delay. The intent of this study was neither the development of new freeway weaving

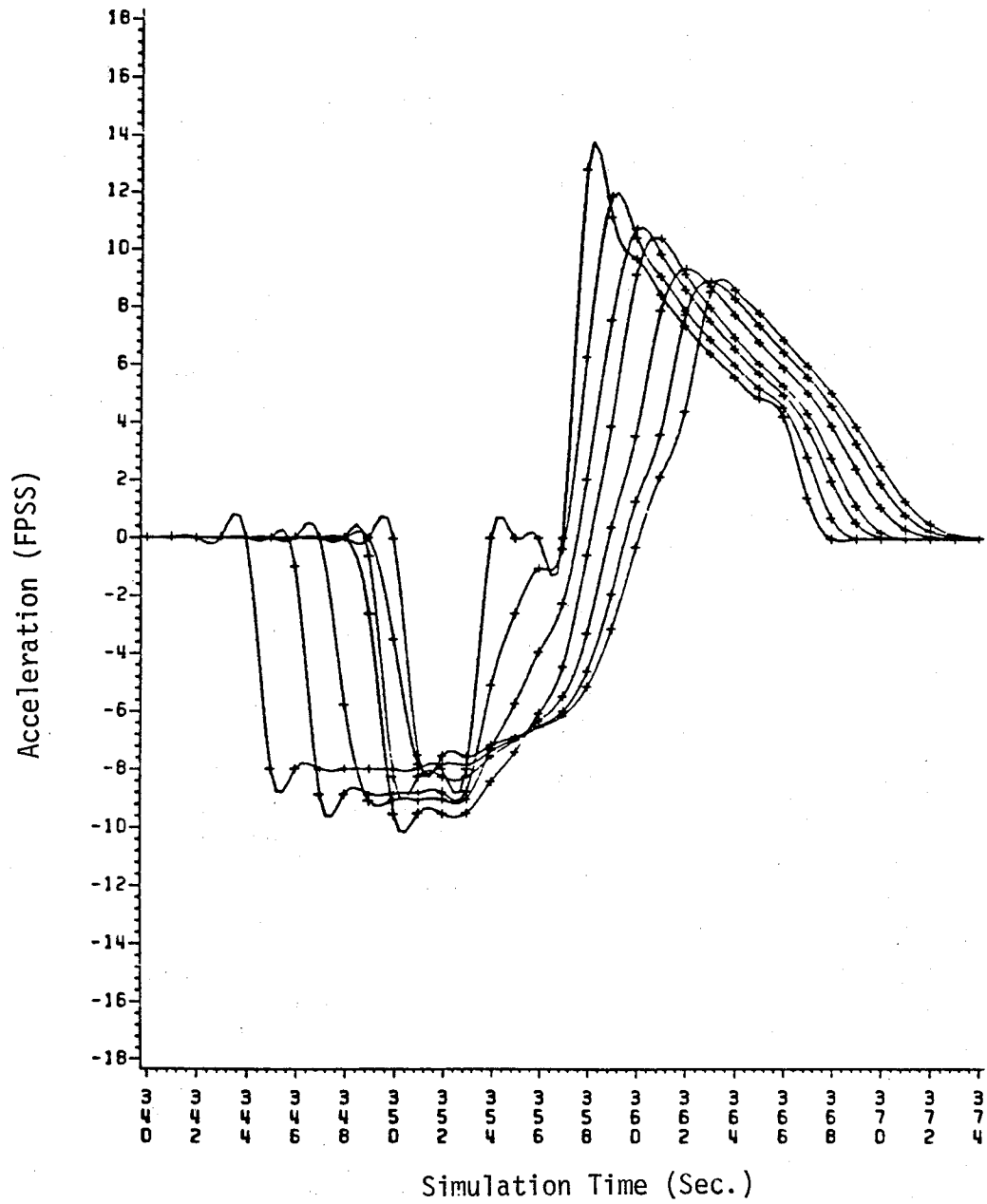


FIGURE 1 Acceleration changes during a speed disturbance.

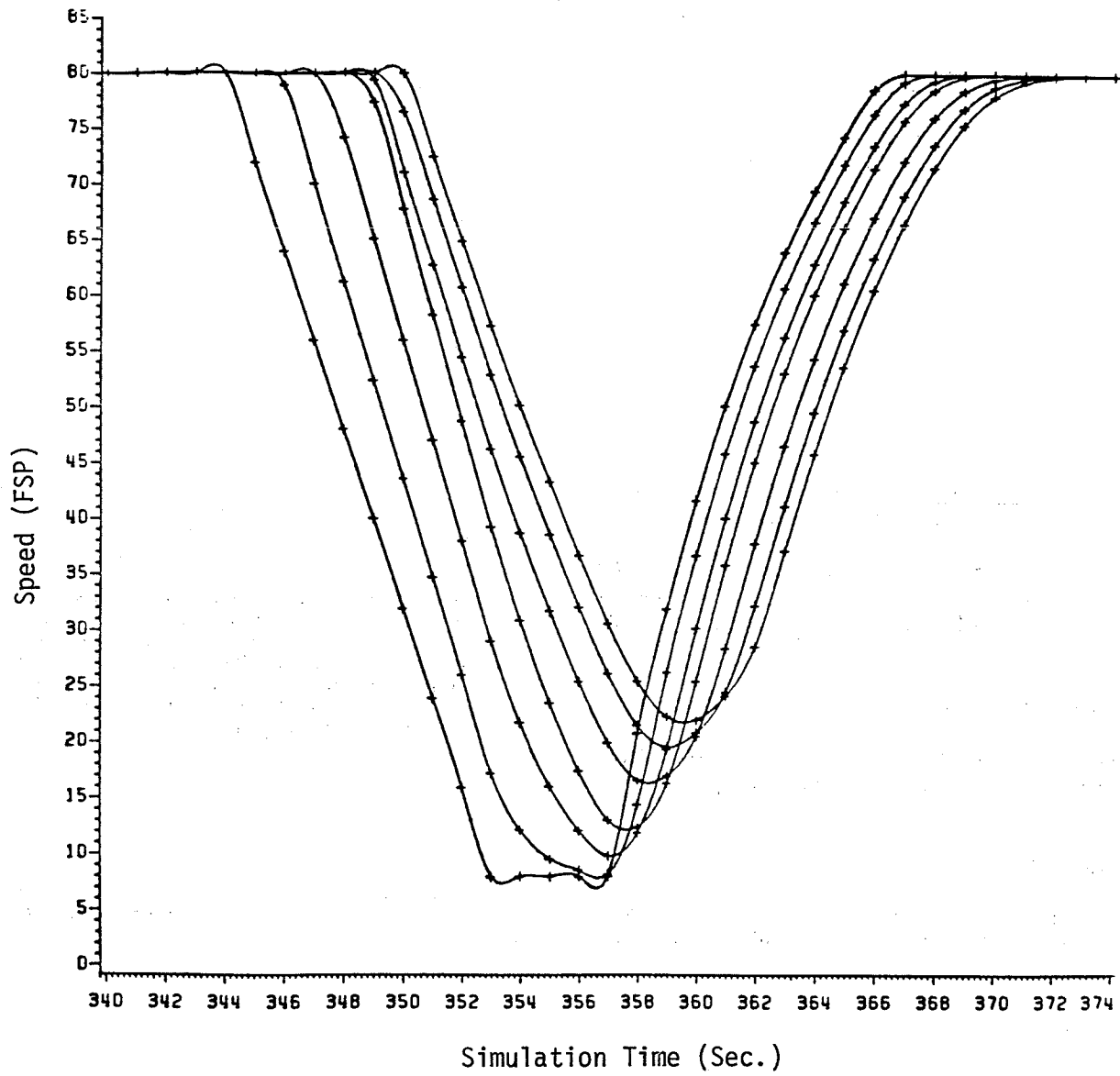


FIGURE 2 Platoon recovery during a speed disturbance.

TABLE 5 SPEEDS AND DELAYS FOR VARIOUS VALUES OF SPEED DIFFERENCE

SPDDIF	SPEED (MPH)		DELAY (SEC./VEH./MILE)	
	WEAVING	NON-WEAVING	WEAVING	NON-WEAVING
0	37.07	40.86	24.89	17.89
10	32.78	38.49	29.19	20.08
20	28.11	36.12	34.18	22.35
HCM	34.50	42.00	N/A	N/A

CONDITION:

Weaving Section Length	= 1000 FT.
Ramp (Minor Entrance) volume	= 794 VPH
Mainline (Major Entrance) Volume	= 2483 VPH
Proportion of Weaving Vehicles	= 0.55

Mainline arrival speeds were sampled from a truncated normal distribution with mean and standard deviations of 60 mph and 10 mph, respectively.

Ramp arrival speeds were sampled from a truncated normal distribution with mean and standard deviations of (60 mph - SPDDIF) and 10 mph, respectively.

section design procedures nor the modification of current procedures, but to develop a tool which might provide the means for such endeavor.

The availability of microscopic data sets from the FHWA permitted the testing of the model. The results of the test indicate that the model is capable of producing realistic results.

#### ACKNOWLEDGMENT

The authors are grateful to Stephen Cohen of the FHWA for providing the data sets which made this study possible.

#### REFERENCES

1. S. A. Smith and M. E. Roskin. Creation of Data Sets to Study Microscopic Traffic Flow in Freeway Bottleneck Sections. In

*Transportation Research Record 1005*, TRB, National Research Council, Washington, D.C., 1985.

2. M. Zarean. *Development of a Simulation Model for Freeway Weaving Sections*. Ph.D. dissertation. The Ohio State University, Columbus, 1987.
3. D. A. Wicks and E. R. Lieberman. *Development and Testing of INTRAS, A Microscopic Freeway Simulation Model, Volume 1: Program Design, Parameters Calibration and Freeway Dynamics Component Development*. Report RD 80/106. FHWA, U.S. Department of Transportation, Washington, D.C., 1980.
4. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.

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

# Effect of Weather on the Relationship Between Flow and Occupancy on Freeways

FRED L. HALL AND DEANNA BARROW

The relationship between flow rates and roadway occupancies on freeways has been investigated in several recent papers. However, all the data in those investigations have come from ideal conditions. The purpose of this paper is to investigate the same relationship under adverse weather conditions. An earlier study found that rain reduced freeway capacity, but it did not address the effects of rain on the nature of the function relating the two variables. Three possible effects on the function are investigated, and it is found that, in essence, the slope of the line relating flow to occupancy (in the uncongested regime) decreases as weather conditions deteriorate. In other words, different parameters are needed for the function to describe the relationship under different weather conditions. This finding has some important implications for efforts to develop a new incident detection algorithm for freeways, based on the nature of the flow-occupancy relationship. This paper addresses the effect of bad weather on freeway traffic operations. In particular, does rainy weather change the nature of the relationships among speed, flow, and roadway occupancy? Although it might be objected that the existence of such an effect is obvious and hardly merits detailed attention, it turns out to be important to develop a quantitative description of the effect for use with an automatic incident detection logic. This paper begins by briefly describing the background for the incident detection logic, in order to explain the rationale for developing a solid quantitative treatment of the effect of bad weather. The second section describes the data that were used for the analysis, with particular reference to weather conditions. The third section, dealing with the analysis, includes both a discussion of the methods used and a presentation of the results. The fourth section presents the conclusions that have been drawn from the investigation.

Recently, Navin (1) and Hall (2) proposed models of freeway operations based on catastrophe theory. One practical consequence of this proposed model is that it provides the possibility for a new logical basis for incident detection, described by Persaud and Hall (3), and suggested earlier by Athol (4). However, for that logic to function effectively, it is first necessary to ensure that the theoretical picture encompasses the full range of operating conditions that are likely to be encountered. Jones et al. (5) suggested that capacity is reduced during rain, although their functions appeared to fit rather poorly at capacity. If their functions are correct, however, this finding could arise in several ways, each of which implies a different consequence for the theoretical picture of traffic operations as represented by catastrophe theory. Athol (4) sketched one of these ways, but it appears to be only an impression, not fully analyzed.

The catastrophe theory model portrays freeway operations data as falling on a partially folded surface. Figure 1 portrays the general shape of this surface, the location of freeway data on it, and the two-dimensional projections from that surface, in the form of speed-flow, speed-occupancy, and flow-occupancy graphs. This model offers a number of improvements over the more conventional model, including the ability to account for the sudden jump in speeds that takes place during transitions to and from congested operations. In addition, the model makes clear that operations do not have to pass through capacity flows in moving to or from congested conditions.

One important empirical finding in the context of the catastrophe theory model is that uncongested operations occur fairly close to the "edge" on the upper fold; that is, small increases in occupancy for constant volume cause operations to "drop over the edge" into congestion. Although this observation needs additional confirmation, it appears to hold true for two different sets of Ontario data and holds the key to the automatic incident detection logic proposed by Persaud and Hall (3). That logic depends on being able to specify the functional form for uncongested operation and then identifying departures from the general pattern that would indicate movement to congestion. The advantage of this logic is that comparison with other locations would not be needed and detection of incidents could thereby be speeded up.

However, for such a logic to be practical, it has to cover more than just good weather days. Before this paper, all the analyses of traffic data in the context of catastrophe theory were intentionally limited to data collected under ideal conditions. From the results of Jones et al. (5), it is clear that capacity is affected. It is not clear whether the underlying relationships among the three variables are also affected, or whether data simply arise over only part of the normal range. In terms of the flow-occupancy curve (which earlier studies using the catastrophe theory model found to have an inverted V shape), three plausible situations could give rise to the results of Jones et al. First, the slope of the lines making up the inverted V may decrease, so that flow is lower at any occupancy. This situation would be a natural result of lower speeds and is the picture sketched by Athol (4). Second, it may be that the location of the function is unchanged at low flows but that for higher flows the slope is decreased. And third, it may be that the location of the line is unchanged but that operations simply do not achieve the higher flow rates during rainy weather. In this case, the data might tend to show an inverted U shape rather than the sharper point of the inverted V.

The proposed incident detection logic would be simplest to implement if the possibility holds, but whether it holds can

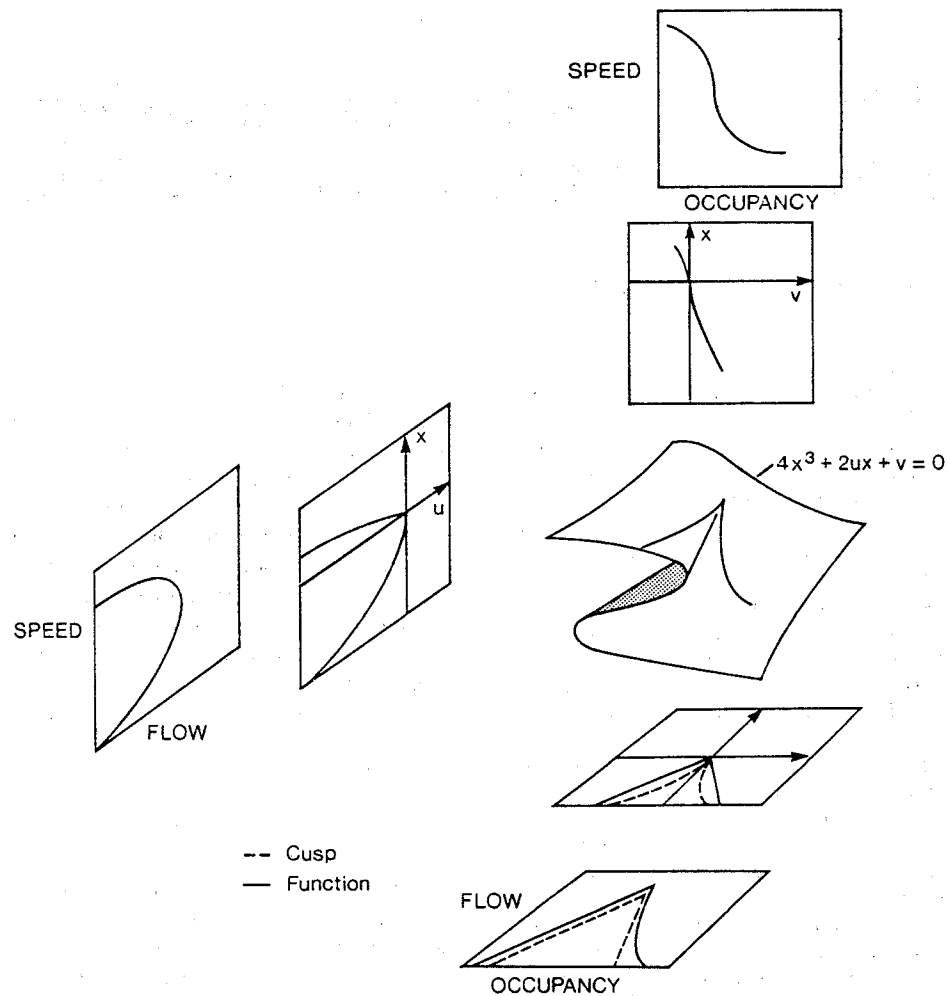


FIGURE 1 Catastrophe theory surface, with a representation of a traffic flow function, projections onto the two-dimensional planes from catastrophe theory, and transformed versions of those projections into traffic operations relationships (2).

be ascertained only by rejecting the other possibilities. Regardless of which one is the case, it is obviously necessary to resolve this question before attempting to implement an automatic incident detection system based on this logic. Similar questions arise for geometric issues, one of which (the effect of grade) was examined by Persaud and Hall (6).

#### DATA

The traffic data were obtained from the Burlington Skyway Freeway Traffic Management System (FTMS), which encompasses approximately 10 km on the Queen Elizabeth Way (QEW) near Hamilton, Ontario. Three aspects of the data acquisition are described: first, the selection of appropriate locations in the FTMS; second, the weather data and selection of appropriate days from the record for analysis; and third, the choice of variables to be used.

#### Locations

The Burlington Skyway FTMS currently includes 12 data collection stations, 6 northbound and 6 southbound, in addition

to closed circuit television and changeable message signs. Several of the 12 stations are located on significant grades, either on the structure over the ship canal connecting Hamilton Harbour to Lake Ontario or in the vicinity of overpasses for city streets. Because it seemed likely that grades would have their own effect on the underlying relationships at issue in this study (see Persaud and Hall (6)), it was necessary to select only locations on relatively level sections of the roadway. Two stations, northbound 7 and northbound 12, met these conditions. Station 7 was selected as the location for the extraction of data from the FTMS records.

#### Weather Data

The *Monthly Meteorological Summary* (7) for the Hamilton Airport weather station, produced by Environment Canada, provided a good basis for determining weather statistics. The monthly reports contained hourly data, including measures of temperature, wind direction, wind speed, and precipitation.

August 1986 and November 1986 were chosen for analysis. Two types of days were selected. Days having prolonged periods of rain, snow, or freezing rain were used to identify the effect

of adverse weather conditions on traffic. Days having no precipitation occurrences were used for comparison.

Because the FTMS site is located about 20 km (12 miles) from the Hamilton Airport weather station, there was some question whether the summary of weather conditions recorded at the airport accurately reflected the weather conditions at the Burlington Skyway. The Ontario Ministry of Transportation and Communications (MTC) kept detailed road condition reports for the winter months, mid-November to mid-April. However, for the remainder of the year, the only weather information available at FTMS was a handwritten log kept when the weather had a direct effect on traffic. We did not wish to rely on the FTMS log because there might well have been times when the weather had not affected the traffic, with the result that relying on the log would give a biased picture of the effect. Hence the *Monthly Meteorological Summary* (7) was the primary source for determining weather conditions at the Burlington Skyway.

Of the 31 days in August, 4 days had consistent reports of rain throughout the day. The 4 days (August 1, 6, 7, and 23) had 8 or more hours of rainfall each. Limiting the analysis to these days was intended to minimize the chance that the rainfall at Hamilton Airport was the result of a localized summer thunderstorm that may not have affected the skyway. (Extended periods of rainfall almost certainly indicate widespread rain.)

To assist in determining the road conditions for the month of November 1986, a comparison was made between MTC's independent data on road conditions and Environment Canada's monthly weather summary. In November 1986, 14 days were recorded by Environment Canada as having some form of precipitation at the airport. Comparison with MTC's records showed only 2 days with wet or slippery road conditions on the skyway: November 20 had continuous snowfall, poor visibility, and slippery road conditions; and November 26 had continuous rainfall and wet road conditions but above-freezing temperatures. Unfortunately, when the November 26 data were retrieved, no usable data were found, apparently because of detector malfunctions. Consequently, all the rainy weather data for the analysis were obtained from August.

It was essential to analyze several days when weather conditions were ideal. For consistency, the clear weather data were also extracted from August. Three days were chosen (August 4, 20, and 22), all of which had reports of no precipitation in Environment Canada's monthly weather summary.

## Variables

After the locations and the days for analysis were selected, it was necessary next to select the variables. At each station, there are pairs of speed measuring detectors in each lane. The FTMS records upstream and downstream flow rates, upstream and downstream occupancies, mean speed, and mean vehicle length for 30-second intervals, 24 hours a day, 7 days a week for each detector pair. Flow rates and occupancies were selected as the variables to work with. The decision to use occupancies rather than to calculate densities was made in earlier work and was discussed in Hall et al. (8). The two main advantages for the present work were that occupancies is the variable provided by most freeway management systems (and therefore most familiar to those managing such systems),

and that it provides a tighter fit against flow than does density. The flow-occupancy relationship was selected for analysis over the other two relationships because the sharply peaked nature of the flow-occupancy relationship makes it easier to separate congested and uncongested operations.

## ANALYSIS

Several steps were followed in the analysis. The first was a simple visual inspection of scatterplots of the data, to see whether they conformed generally to the picture developed earlier using 5-minute data (8, 9). Because the data did conform, the second step was to remove the congested data from the file, in order to concentrate on the relationship for uncongested operations. The third step was to fit functions to those data, for each day separately and then for combinations of days. Several procedures were used in this step. A fourth step was to look for differences only in the high flow ranges, by fitting functions to a restricted range of occupancy values. In these ways, all three possibilities raised earlier were investigated.

Inspection of data plots for flow versus occupancy indicated general similarities between the rainy weather (fig. 2) and clear weather data (fig. 3), as well as with plots based on 5-minute averages (fig. 4). The left-hand, or uncongested, side of the curve is represented by a tight cluster of points, whereas the right-hand side is represented by a large scatter of points. This correspondence in the general pattern allowed for a more detailed, statistical comparison. Had the overall patterns been dissimilar, there would have been no need to proceed.

Maximum flow rates are higher in figures 2 and 3 than in figure 4 because in all cases the flow rates are calculations of hourly values, based on short interval observations. Figures 2 and 3 represent 30-second observations, whereas figure 4 is based on 5-minute counts. In fact, the maximum flows observed in figures 2 and 3 are not sustainable for more than 30 seconds. Detailed inspection of the data shows that no two consecutive observations have such high flow rates. Despite the volatility of the 30-second data, we have chosen to analyze the data at that level of detail, in order for this analysis to contribute to an efficient incident detection logic.

For two reasons it was decided to limit the analysis to the uncongested portion of the data. First, the rationale for this investigation arises from the need to identify the limits on uncongested operation, to detect incidents that represent departures. Hence the uncongested data are of primary interest. Second, the scatterplots suggest that this portion of the data can be depicted sensibly by a line. The congested data, on the other hand, are so widely scattered as to suggest that it would be misleading to represent them by a single line. Further, if the underlying relationship between these two variables is an inverted V, as has been suggested (8, 9), then the two arms can be fitted separately. Indeed, to include the congested data may well distort the relationship identified for the uncongested portion of the curve.

To ensure that only uncongested data were included in the analysis, the following procedures were used. First, periods of congestion were identified on the basis of speeds. The definition of congestion was based on a speed change of more than 18 km/h over a 30-second interval. The period of congestion begins with a sudden drop in speeds and ends with a

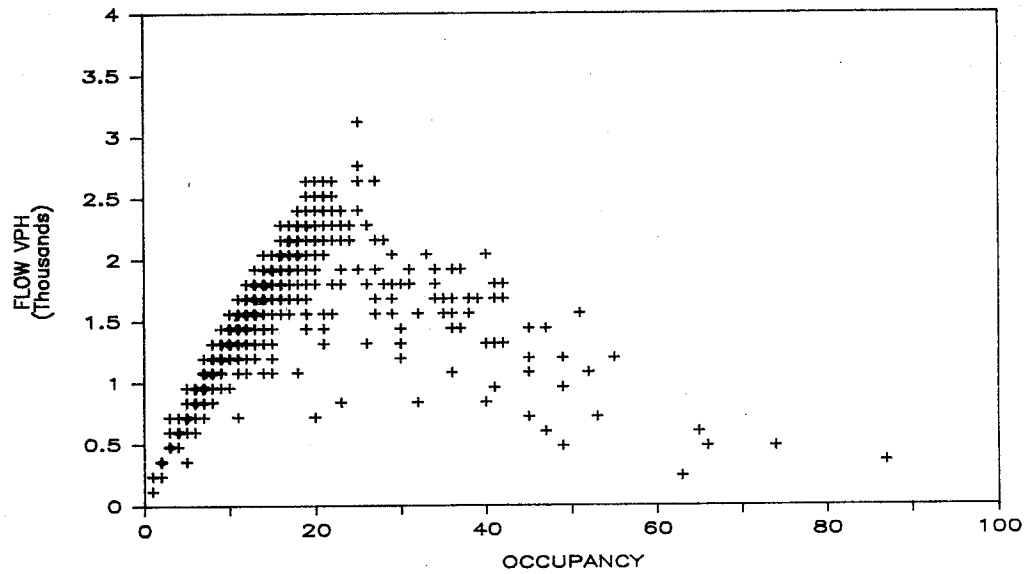


FIGURE 2 Flow rates versus occupancy, based on 30-second data, for rainy conditions, August 7, median lane, station 7, Skyway FTMS.

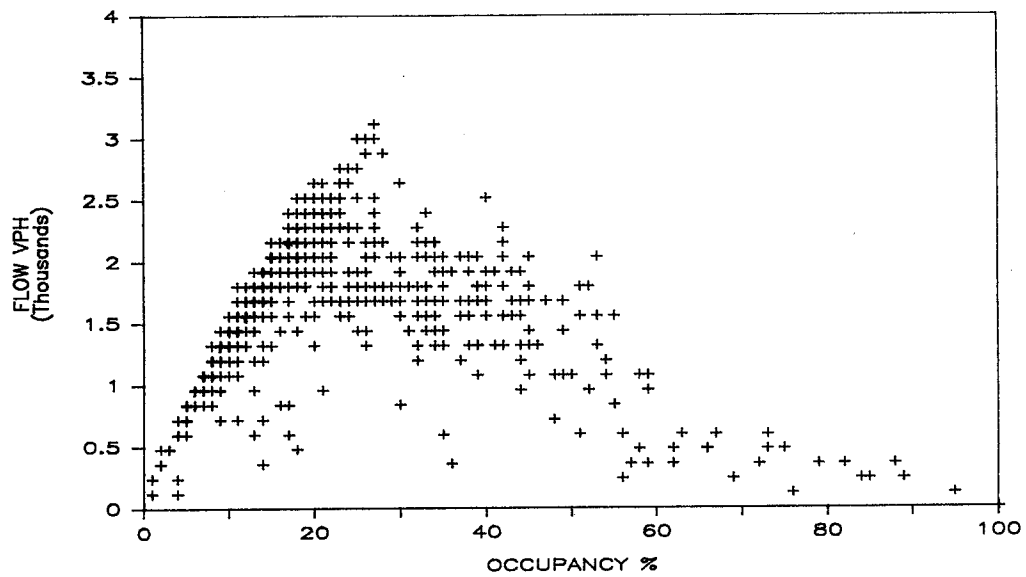


FIGURE 3 Flow rates versus occupancy, based on 30-second data, for clear weather, August 4, median lane, station 7, Skyway FTMS.

jump of this magnitude back from low to high speeds. At the stations of interest on the skyway, there are no physical bottlenecks that would cause recurrent congestion. Instead, congestion occurs as a result of an incident. In order to be sure to remove all data that might represent transitions between congested and uncongested operations, we also deleted from consideration 5 minutes of data before the drop in speed and 5 minutes of data after the speed jump. Any interval containing missing data for one or more variables was also deleted on the assumption that the variables that were reported for the interval also may have been affected by the temporary malfunction in the detector. This screening of the data began at 6 A.M. on each day and continued until roughly 500 data points were obtained for each day, or until 6 P.M.

Functions were fitted to the uncongested data in order to test the first possibility raised earlier, namely, that the slope of the relationship is reduced by bad weather. The approach taken for the curve fitting was as follows. First, a functional form had to be selected. Then equations were fitted to each day's data separately, using all of the available uncongested data. Tests were run with dummy variable regressions to see if the observable differences in coefficients were significant. On the basis of these results, it was recognized that it was necessary to draw a sample from each day's data, in order to meet one of the assumptions of regression analysis more closely, namely that of a uniform distribution of observations across the independent variable. On the basis of this sample, all of the equations were reestimated.



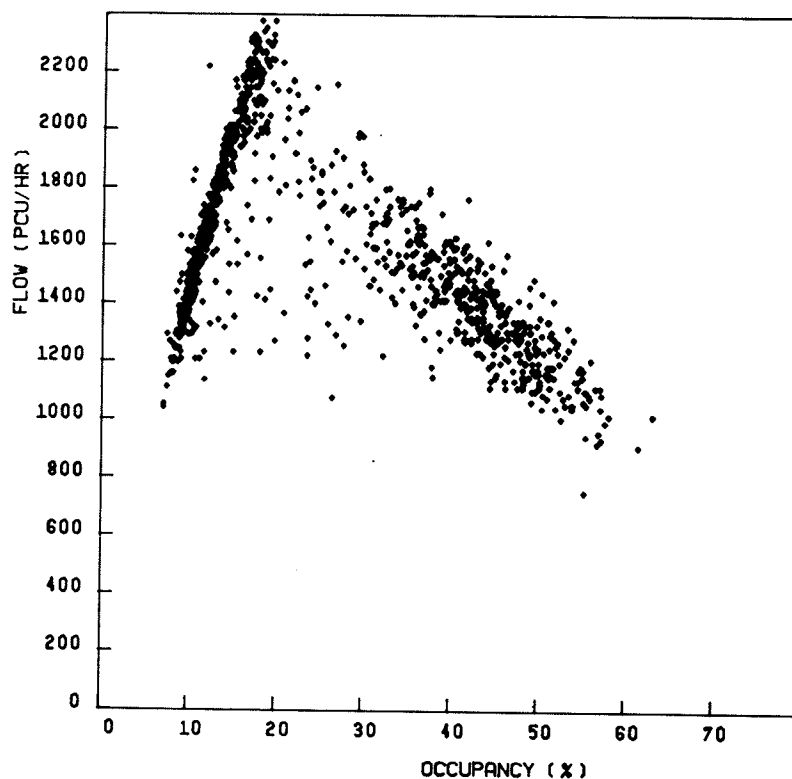


FIGURE 4 Flow rates versus occupancy, based on 5-minute data, for clear weather, 45 days, late 1979, median lane, station 4, QEW FTMS at Highway 10 (8).

In the process of extracting the uncongested data from the larger set, it became obvious that the critical occupancy (i.e., that at which maximum flows are achieved) is lower for the rainy weather (about 28%) than for the clear weather (about 30%). This was not tested mathematically, but it is worth noting. For the regression analyses, both sets are restricted to values less than or equal to 28% for comparability. Several tests were run to ensure that this limitation did not change the nature of the function for the clear weather days. (Although critical occupancy is an important parameter in some incident detection approaches, it appears to be a more conservative indicator than is necessary and is therefore not pursued in this analysis.)

From the visual inspection, two functional forms seemed plausible: a power function and a linear one. The linear model was rejected because it would not necessarily go through the origin, which is mandatory for an equation representing flow and occupancy. Consequently, the function used was

$$\text{flow} = b_0 * \text{occupancy}^{b_1}$$

and the model actually estimated was

$$\ln(\text{flow}) = a + b_1 * \ln(\text{occupancy})$$

where  $a = \ln(b_0)$ .

The results in table 1 based on the full set of data show that all of the equations are quite similar, and there is indeed a fair amount of overlap among the  $b_1$  coefficients for the two weather categories. These results also show that in each weather category there is one outlier from the general cluster of  $b_1$  values.

The obvious next step is to test whether the difference in equations within one weather condition is significant. If it is, then there is little point in looking for significant differences between weather conditions. Initial investigation of this issue showed that before it could be properly resolved, the data needed some further treatment. The observations are clearly not uniformly distributed across the range of densities (see figs. 5 and 6). This distribution is contrary to one of the assumptions of regression analysis and might well cause misleading results when statistical inference is important. To overcome this possible problem, it was necessary to sample from within the available data.

A random sampling procedure was carried out on the full data set for each day. From the full set, 10 flows were randomly selected, without replacement, at each occupancy for each day. When there were fewer than 10 flows at any occupancy, as is the case with the extreme occupancies, all were retained to ensure as close to a uniform distribution as possible without drawing too small a sample.

The equations for each day were then reestimated on the basis of the sampled data (table 2). The most noticeable point is that the equation for what had been the outlying day has been changed and is no longer the farthest from the group average.

To test whether there are significant differences among the days in one condition, a dummy (i.e., 0, 1) variable was introduced in table 3, and an expanded equation analyzed:

$$\ln(\text{flow}) = a + b_1 * \ln(\text{occ}) + b_2 * D_1 + b_3 * D_1 * \ln(\text{occ})$$

where  $D_1$  is 1 for the extreme day (August 23 for rainy weather;

TABLE 1 REGRESSION ANALYSIS ON UNCONGESTED DATA SETS

DAY	COEFFICIENTS (T-RATIO)		R <sup>2</sup>	# OF DATA POINTS
	A	B1		
RAINY WEATHER				
1 AUG. 1986	5.411 (204.6)	0.7768 (74.60)	0.942	342
6 AUG. 1986	5.342 (217.5)	0.8064 (79.78)	0.928	498
7 AUG. 1986	5.325 (317.3)	0.8124 (115.8)	0.964	500
23 AUG. 1986	5.341 (445.6)	0.7969 (137.8)	0.972	550
CLEAR WEATHER				
4 AUG. 1986	5.386 (464.1)	0.8000 (142.6)	0.978	456
20 AUG. 1986	5.349 (276.5)	0.8308 (105.9)	0.954	547
22 AUG. 1986	5.364 (346.5)	0.8271 (128.1)	0.971	495

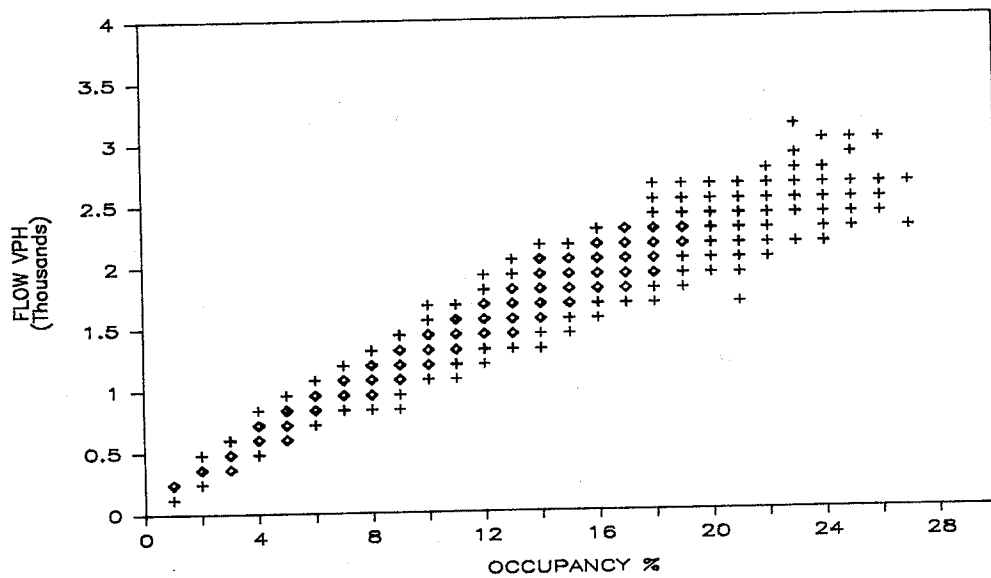


FIGURE 5 Uncongested flow-occupancy data, for all four rainy days, showing areas of heavier concentration of data.

August 22 for good weather), and 0 otherwise. In such an equation, the coefficients for the condition represented by the dummy variable can be found by summing the relevant coefficients. For example, in the equation for the rainy days, the pooled estimate for the three days is

$$\ln(\text{flow}) = 5.37 + 0.789 * \ln(\text{occ})$$

and the estimate for the extreme day, August 23 can be found, setting the dummy variable equal to 1, as

$$\ln(\text{flow}) = (5.372 - 0.049) + (0.7891 + 0.0198) * \ln(\text{occ})$$

which is exactly the equation for August 23 given in table 2.

If the extreme day requires a separate equation, either  $b_2$  or  $b_3$  or both will be significant. Because these coefficients have been estimated on the basis of the full sample size for that day (207 points in the case of August 23), the sample is large enough that the Student's  $t$  distribution can be approximated by the normal distribution. Hence the critical value for a 5% confidence level is 1.96. The results (table 3) show that neither of the extreme days requires an equation that is significantly different from the mean of the others in the group.

The final step was to run a regression using three dummy variables, one for each of the extreme days as just identified and one for the weather condition itself. This approach would

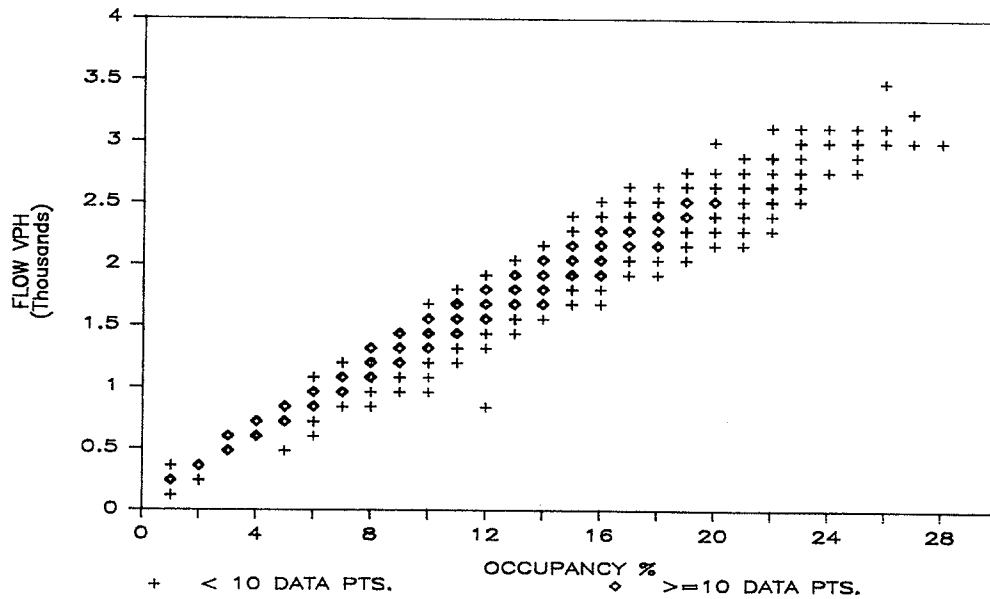


FIGURE 6 Uncongested flow-occupancy data, for all three clear days, showing areas of heavier concentration of data.

TABLE 2 REGRESSION ANALYSIS ON SAMPLED DATA SETS

DAY	COEFFICIENTS (T-RATIO)		R <sup>2</sup>	# OF DATA
	A	B1		
RAINY WEATHER				
1 AUG. 1986	5.397 (195.69)	0.7813 (71.25)	0.963	197
6 AUG. 1986	5.370 (156.0)	0.7884 (56.92)	0.939	211
7 AUG. 1986	5.354 (165.1)	0.7953 (61.03)	0.947	210
23 AUG. 1986	5.323 (206.8)	0.8089 (72.22)	0.962	207
CLEAR WEATHER				
4 AUG. 1986	5.381 (326.3)	0.8024 (112.19)	0.984	210
20 AUG. 1986	5.369 (249.5)	0.8256 (95.43)	0.977	219
22 AUG. 1986	5.411 (307.2)	0.8087 (109.7)	0.983	213

show clearly whether the difference between weather conditions is more important than the difference within one condition. The *t* statistics for the resulting equation show that only one of the coefficients involving a dummy variable is significant—the coefficient on the logarithm of occupancy for rainy days. Thus this segment of the analysis supports the hypothesis that different functions are needed for the two weather conditions and that there is no significant variation within one weather condition. Because separate equations are needed, they were estimated individually, without any dummy variables for separate days. The resulting functions and data are shown in figure 7. The equations found were, for clear weather days,

$$\ln(\text{flow}) = 5.385 + 0.81371 * \ln(\text{occ})$$

and for rainy weather days,

$$\ln(\text{flow}) = 5.352 + 0.7969 * \ln(\text{occ})$$

The possibility remains that the functions are different not because of an overall shift in the data but because of a shift in only one part of the range. There appears to be some visual support for this for the rainy-days data in figure 7, in that for occupancies above 20% the data may not be evenly distributed about the line but instead lie predominantly below it. Data for the clear days, however, seem uniformly distributed about the line.

TABLE 3 REGRESSION ANALYSIS ON SAMPLED DATA SET USING DUMMY VARIABLES

RAINY WEATHER		EXTREME DAY 23 AUG. 1986	
D1 = 1 extreme day = 0 remaining rainy weather days			
$\ln(\text{flow}) = 5.372 + 0.7891 \cdot \ln(\text{occ}) - 0.049 \cdot D1 + 0.0198 \cdot D1 \cdot \ln(\text{occ})$			
(t-statistic) (279.17) (102.36) (-1.63) (1.56)			
CLEAR WEATHER		EXTREME DAY 22 AUG. 1986	
D2 = 1 extreme day = 0 remaining clear weather days			
$\ln(\text{flow}) = 5.37 + 0.816 \cdot \ln(\text{occ}) + 0.0395 \cdot D2 - 0.0076 \cdot D2 \cdot \ln(\text{occ})$			
(t-statistic) (408.47) (148.98) (1.74) (-0.80)			
COMBINED RAINY AND CLEAR WEATHER DATA			
D1 = 1 extreme rainy weather day D2 = 1 extreme clear weather day D3 = 1 all rainy weather days = 0 all clear weather days			
$\ln(\text{flow}) = 5.37 + 0.789 \cdot \ln(\text{occ}) - 0.0491 \cdot D1 - 0.0001 \cdot D3 + 0.0395 \cdot D2$			
(t-statistic) (315.13) (115.54) (-1.84) (-0.00) (1.40)			
$+ 0.0198 \cdot D1 \cdot \ln(\text{occ}) + 0.0272 \cdot D3 \cdot \ln(\text{occ}) - 0.0076 \cdot D2 \cdot \ln(\text{occ})$			
(1.76) (2.82) (-0.64)			

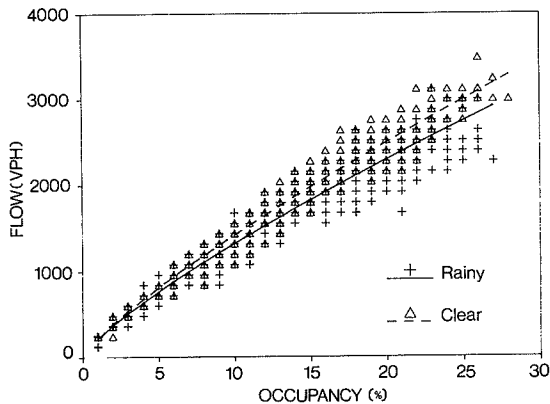


FIGURE 7 Overlaid sampled flow-occupancy data for both clear and rainy conditions, with regression lines.

To test whether the differences occur only at high flows, additional dummy variable regressions were run (using only the one dummy variable related to weather). These were done sequentially, eliminating all data for occupancies above a given value, starting with 27% and decreasing the threshold 1% each time, until there remained no significant difference due to weather. Only for occupancies below 8% did the difference between the two equations disappear entirely. Because that result meant that the difference in function is found over two-thirds of the data range, we deemed it more appropriate to use two separate equations for the full range rather than attempt to differentiate them only above 8%.

It is clear from the comparison of figure 2 with figures 3 and 4 that the third possibility, that of a U-shaped function for rainy days, is not supported. Consequently, it appears

from the analysis of the rainy-days data that there is a clear shift downward during adverse weather conditions, in the function representing uncongested flow-occupancy data.

To test this result when carried to an extreme, we also investigated the one November day (the 20th) with severe weather conditions (snow, poor visibility, and slippery roads), attempting to analyze it the same way. The most difficult task proved to be separating the congested and uncongested data. Speed drops of the magnitude used for the earlier data were not found here, undoubtedly because uncongested speeds were so much lower. The difficulties can be seen in the flow-occupancy plot of these data (figure 8). Consider the left-hand side data—i.e., the cluster around the diagonal from (0, 0) to (20, 2000). It seems likely that some of the data on the lower part of this cluster represent operations in the early stages of congestion. Including speed as a decision variable did not make the distinction any more obvious. After trying a number of decision criteria, we decided to use the simple one that speeds greater than or equal to 70 km/h represented uncongested flow, simply for comparison purposes. The best fit line to those data is

$$\ln(\text{flow}) = 4.9868 + 0.8883 \ln(\text{occ})$$

with an  $R^2$  of 0.946, based on 367 observations. Because of the arbitrariness of this decision criterion, the difference between this equation and those for clear or rainy days was not tested statistically. However, a visual inspection (fig. 9) shows that the estimated flow-occupancy line for the snowy conditions falls below the other two, as might be expected. Note also that to the extent that the 70 km/h criterion has excluded points which in fact represent uncongested operations, the line for snowy conditions is closer to the other two than it really should be.

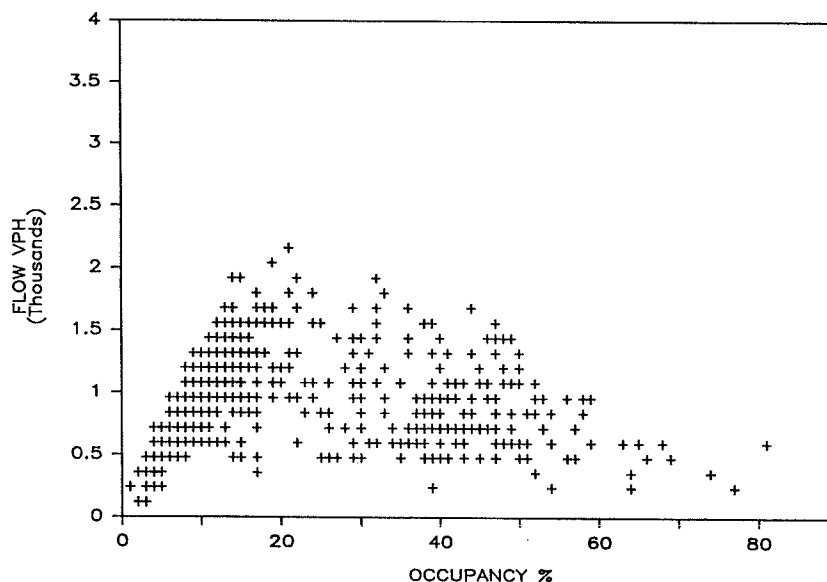


FIGURE 8 Flow rates versus occupancy, based on 30-second data, for snowstorm conditions, November 20, median lane, station 7, Skyway FTMS.

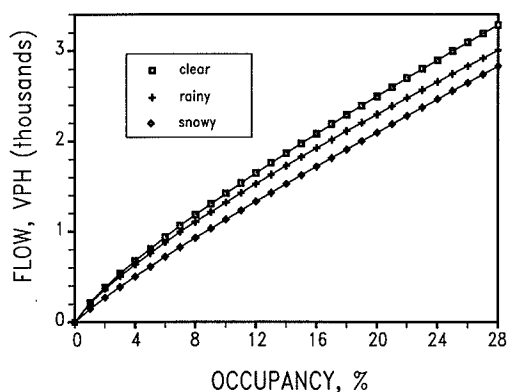


FIGURE 9 Regression estimates of the uncongested flow-occupancy curves for three weather conditions.

## CONCLUSIONS

The immediate conclusions from this analysis seem clear: adverse weather affects the flow-occupancy function by reducing the slope of the curve that describes uncongested operation; as a result, maximum flows are also reduced during such conditions. Neither of these results comes as any great surprise. Indeed, Athol (4) suggested this kind of picture more than 20 years ago. Nevertheless, these results have helped to clarify the nature of the changes.

The implications of these results for the incident detection logic described by Persaud and Hall (3) appear to require a more complicated logic than might have been needed otherwise. Because bad weather will cause a downward shift in the flow-occupancy line, the incident detection trigger cannot rely on these two variables alone but must simultaneously include speed. Yet as conditions worsen (e.g., the November 20 snowstorm), speeds naturally decline until it is difficult to identify speed drops.

One possibility is that the proposed incident detection logic

will be of most benefit for climates that are more temperate than Ontario's, or for fair weather only. An alternate possibility is that an adaptive logic could be developed that builds the "expected" picture of operations anew every few hours or on demand. These, however, are the consequences of the results. The conclusion itself seems clear enough: bad weather affects the flow-occupancy relationship by pushing down the tip of the inverted V and making it flatter.

## REFERENCES

1. F. P. D. Navin. Traffic Congestion Catastrophes. *Transportation Planning and Technology*, Vol. 11, 1986, pp. 19-25.
2. F. L. Hall. An Interpretation of Speed-Flow-Concentration Relationships Using Catastrophe Theory. *Transportation Research A*, Vol. 21A, 1987, pp. 191-201.
3. B. Persaud and F. L. Hall. Catastrophe Theory and Patterns in 30-Second Freeway Traffic Data—Implications for Incident Detection. Presented at the Annual Meeting, Transportation Research Board, 1988.
4. P. Athol. Interdependence of certain operational characteristics within a moving traffic stream. *Highway Research Record 72*, HRB, National Research Council, Washington, D.C., 1965, pp. 58-87.
5. E. R. Jones, M. E. Goolsby, and K. A. Brewer. The environmental influence of rain on freeway capacity. *Highway Research Record 321*, HRB, National Research Council, Washington, D.C., 1970, pp. 74-82.
6. B. Persaud, and F. L. Hall. Effect of Grade on the Relationship Between Flow and Occupancy on Freeways. *Transportation Research Record 473*, TRB, National Research Council, Washington, D.C., 1988, pp. 33-38.
7. Environment Canada, Atmospheric Environment Service. *Monthly Meteorological Summary at Hamilton Airport, Ontario*; August 1986 and November 1986.
8. F. L. Hall, B. L. Allen, and M. A. Gunter. Empirical analysis of freeway flow-density relationships, *Transportation Research A*, Vol. 20A, 1986, pp.197-210.
9. F. L. Hall, and M. A. Gunter. Further analysis of the flow-concentration relationship. *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986, pp. 1-9.

# Analysis of Platoon Dispersion with Respect to Traffic Volume

KARSTEN G. BAASS AND SERGE LEFEBVRE

Platoon dispersion depends on friction. The authors hypothesize that this friction can be external and internal and that the internal friction depends on traffic density and volume. Platoon dispersion should consequently be influenced by traffic density and volume. This hypothetical relationship was confirmed by limited field observations and simulations. Platoon dispersion increases with increasing volume up to a maximum value, which is observed at lower volumes than capacity. Dispersion decreases with further increasing volume and becomes nearly zero at volumes near capacity. If further studies confirm the observed relationship, platoon dispersion could be calculated directly from densities or volumes and travel times, applied to Robertson's model on a link by link basis, thus improving the results of traffic signal coordination models.

Traffic signals have the effect of grouping vehicles into platoons that leave the intersection under saturated flow conditions. As the platoon moves down the arterial it tends to spread out over time and space. This phenomenon of platoon dispersion is an important factor in traffic signal coordination. Robertson's model allows one to calculate the histogram of a platoon at different points downstream from an intersection. The basic equation that determines the degree of dispersion is

$$F = \frac{1}{1 + \alpha\beta t} = \frac{1}{1 + Kt}$$

The value of dispersion  $K$  (a capital letter is used here to avoid confusion with the density  $k$ ) in this equation depends on the type of road under study and on the degree of friction. McCoy et al. (1) described platoon dispersion in lower friction environments, and the TRANSYT 7F manual allows one to choose a  $K$  value depending on several degrees of friction ranging from low to high as in figure 1.

Figure 1 shows that platoon dispersion increases with friction. It would be useful to have a more quantitative measure of friction that could take explicitly into account an independent variable like service volume or volume divided by saturation flow. Service volumes or saturation flows would be established considering the many independent variables such as corridor width, percentage of heavy vehicles, and so on, and would be particular to a given arterial type. If friction could then be explained by service volumes, the basic dispersion equation would become a function of service volumes and travel times. Platoon dispersion could then be integrated on a link basis into the traffic signal coordination model.

The aim of this paper is to describe the concepts of such a

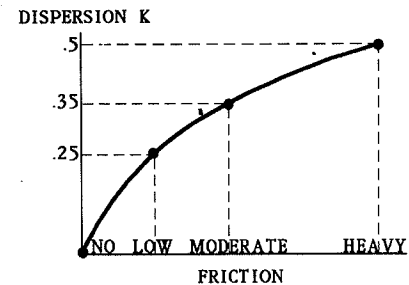


FIGURE 1 Platoon dispersion as a function of friction (TRANSYT 7F).

relationship based on field observations, microscopic simulations, and theoretical considerations.

## CONCEPT AND THEORETICAL CONSIDERATIONS

One can easily observe that platoon dispersion is a function of the physical features of the highway facility and of the interaction between vehicles in the traffic stream. The friction concept introduced by May (2) in 1959 is very useful in the analysis and explanation of platoon dispersion. Friction results from two phenomena: internal friction between moving vehicles and external friction related to roadway features. External and internal friction are implicitly considered in the concept of level of service, which takes into account such variables as number of lanes and their widths, vertical and horizontal alignment, and the location of the arterial in the city.

The internal friction reflects the individual driver's comfort, safety, and freedom to maneuver; the greater the traffic volumes, the more adverse the effects of friction. Conceptually, when volume and density are zero, friction also is zero and increases with increasing volume and density up to a certain maximal value, which occurs at less than the maximal volume (level of service  $E$ ). As volume increases, friction diminishes and approaches zero at capacity, when every vehicle is moving at nearly the same speed. Thereafter, friction again increases with increasing density and should be maximal (at least theoretically) at jam density. Figure 2 illustrates this conceptual relationship.

The purpose of the following paragraphs is to show whether there is a rational explanation of this interrelationship to be found in the different phases of driver behavior as volume and density increase. May (3) and Leutzbach (4) distinguished three zones of different driver behavior. May identified the three zones on the volume-density curve: constant speed, constant volumes, and constant rate of change of volumes with

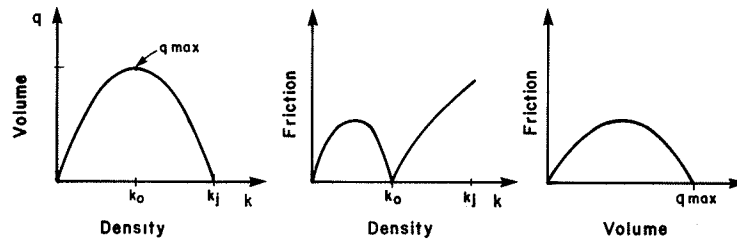


FIGURE 2 Conceptual relationship between friction, density, and volume.

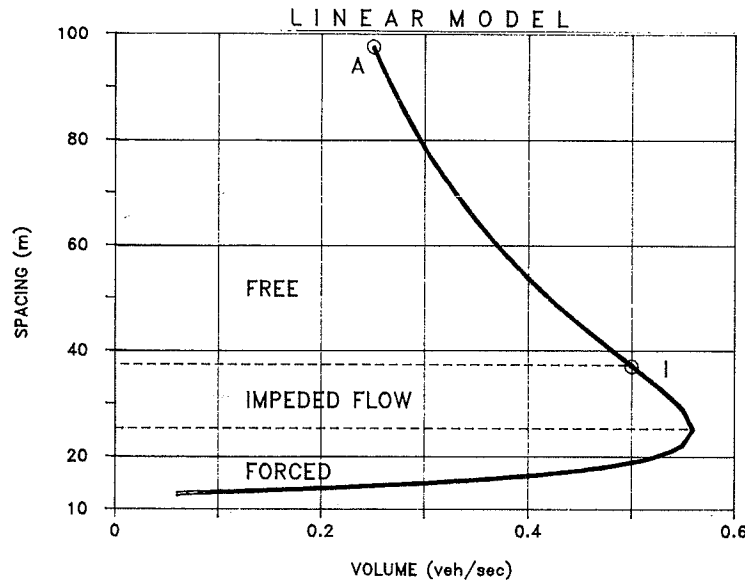


FIGURE 3 Volume-spacing relationship (linear model).

density. In the following section, an explanation for this division into three zones is attempted and at the same time a rational explanation for friction and corresponding platoon dispersion is given.

In order to analyze driver behavior with respect to volume changes, let us study the volume-spacing relationship (fig. 3).

For the purpose of example, this figure represents the linear macroscopic traffic model with a free speed of  $v_f = 28$  (m/s) and a jam density of  $k_j = 0.08$  (veh/m). Attentive analysis of this figure reveals that there is a point of inflection in the curve at a certain spacing; that is, the driver changes behavior at this point. The situation becomes much clearer when one analyzes the rate of change of spacing with respect to a uniform increase of volume. This relation is shown in figure 4.

Starting from a point A on the free-flow side of the curve, the rate of change of spacing decreases faster and faster with increasing volume up to a point, whereas the rate of change remains nearly constant (at the point of inflection I). For more clarity and in order to better understand the changing behavior of the driver, let us consider figures 5 and 6, which represent the rate of change of speed and headway with respect to volume.

Near the point I, the rate of change of spacings is constant while the rate of change of speeds is increasing, and the rate of change of headways is nearly constant. Because headways remain nearly constant at volumes near the capacity, spacings and speeds will change rapidly when volumes increase beyond the point of inflection. At volumes higher than those observed

at the point of inflection ( $q_i$ ) drivers behave differently with respect to the car they follow than at lower volumes. In figure 4 one can distinguish three zones: free flow, impeded flow, and forced flow. This change in driver behavior influences platoon dispersion. Dispersion increases with increasing volumes up to or near the point of inflection and then decreases to near zero value at capacity. The notion of platoon dispersion at densities higher than those observed at capacity seems to be meaningless.

### MATHEMATICAL INVESTIGATION

The location of the point of inflection on the volume-density curve was studied in more detail for the macroscopic traffic stream models defined by Drew (5). In his notation,  $n = -1$  represents the exponential model,  $n = 0$  the parabolic model, and  $n = 1$  the linear model.

The formulae relating spacing  $s$  to volume  $q$  for these models are as follows. For  $n = -1$ ,

$$q = \frac{1}{s} u_m [\ln(sk_j)]$$

$$\frac{dq}{ds} = \frac{1}{s^2} u_m [1 - \ln(sk_j)]$$

$$\frac{dq}{ds^2} = \frac{1}{s^3} u_m [2\ln(sk_j) - 3]$$

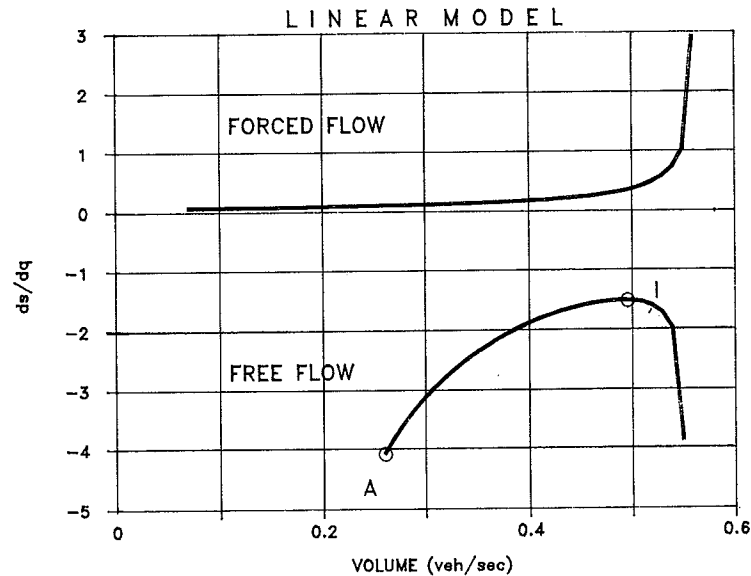


FIGURE 4 Rate of change of spacing versus volume.

For  $n \geq -1$ :

$$q = \frac{u_f}{s} [1 - (1/sk_j)^{(n+1)/2}]$$

$$\frac{dq}{ds} = u_f \left[ \frac{(n+3)}{2} s^{-(n+5)/2} (k_j)^{-(n+1)/2} - s^{-2} \right]$$

$$\frac{dq}{ds} = u_f \left[ 2s^{-3} - \frac{(n^2 + 8n + 15)}{4} (k_j)^{-(n+1)/2} s^{-(n+7)/2} \right]$$

Table 1 gives a comparison between the spacings  $sI$  at the point of inflection and the spacing  $sQ$  at maximal volume as a function of the jam spacing  $s_j$ . Spacings at the point of inflection or of changing driver behavior are 1.5 times longer than the spacings at capacity, and the volume is 10% less than the maximum volume.

It is interesting that the point of inflection is very near or at the same point as the maximum kinetic energy described by Drew (6). Rewriting the energy equation as a function of spacing we have, for  $n = -1$ ,

$$E = \frac{(u_m)^2}{s} [\ln(sk_j)]^2$$

For  $n \geq -1$ ,

$$E = \frac{(u_f)^2}{s} [1 - 2(sk_j)^{-(n+1)/2} + (sk_j)^{(n+1)}]$$

The curve relating spacing to kinetic energy is given, as an example, for the linear model in figure 7.

A comparison of spacings at the point of maximum kinetic energy and at the point of inflection is given in table 2.

Drew et al. (6) stated that the kinetic energy of a traffic stream is equal to the sum of energies of its constituent particles and will be a maximum when internal friction caused by vehicular interaction is a minimum. In the case of the linear model, the point of inflection and maximum energy or theoretical minimum of internal friction fall together. One can conclude that there is a point where driver behavior changes

with respect to spacing, and this point occurs at lower densities and at lower volumes than the capacity. When the volume approaches a certain value, drivers reduce their spacings less and less in reaction to a unit increase of volume. At the point of inflection this tendency is sharply reversed. The actual location of this point of change of driver behavior depends on the type of facility under study and on the model chosen. One might compare this point of inflection to the transition between laminar and turbulent flow in hydraulics.

#### OTHER CRITERIA FOR REPRESENTING INTERNAL FRICTION

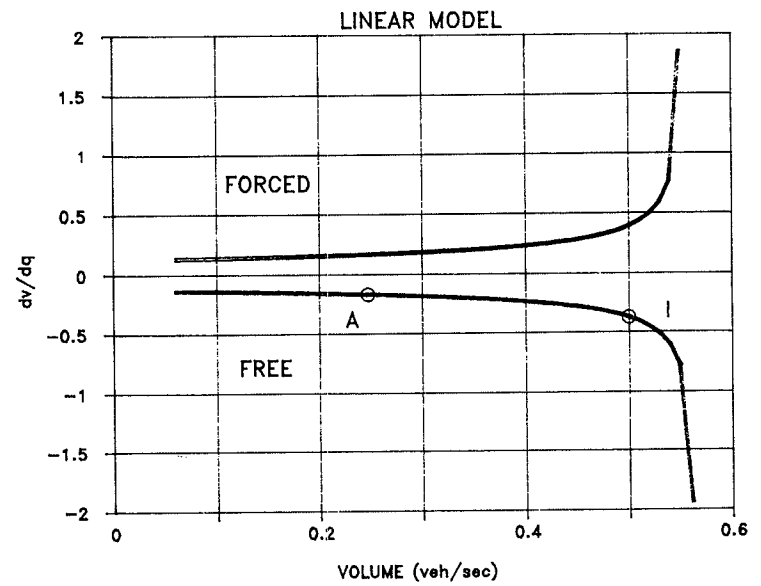
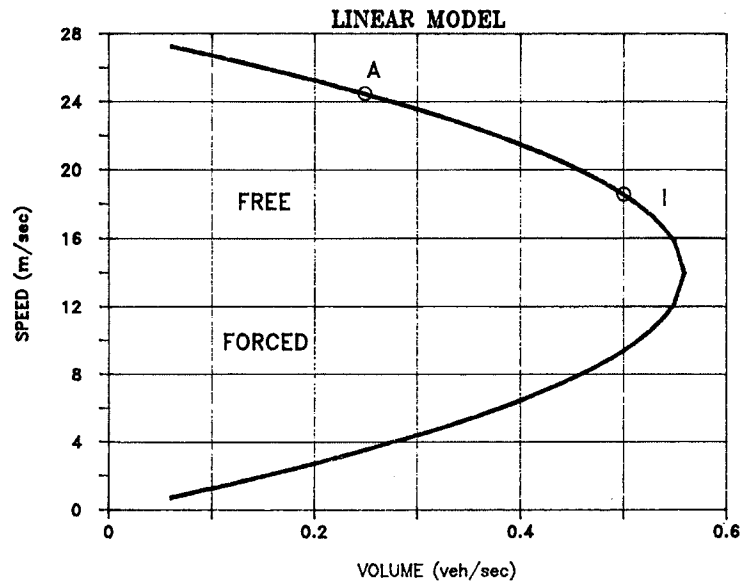
Lee et al. (7) developed a criterion for internal energy or friction based on observations. They found that acceleration noise, as proposed by Drew, does not accurately represent internal friction or the internal energy of the traffic stream. Acceleration noise, at least in their observations, did not show a clear relationship to increasing density. They proposed as a good measure for internal friction the coefficient of variation of speed  $\sigma/v$ , also called the constant of diffusion. As spacings decrease, this measure increases up to a point (the point of maximum energy or the point of inflection); then it decreases to nearly zero at capacity. Figure 8 shows this indicator for the values found by Lee et al. as a function of density.

Although there was no indication in their paper about the volume-density relationship, one can see that the relationship follows the one postulated in the preceding paragraphs for the relation between dispersion and volume.

The measure of friction developed by Helly et al. (8) is the mean velocity gradient, which represents the acceleration noise divided by the average velocity. The measure seems to have some merits, but there are not enough observations available to confirm them.

In the light of what was said, one can hypothesize that dispersion depends on service volumes, or, stated more simply, on volumes divided by the saturation flow. In order to confirm the hypothesis stated in the preceding paragraphs, a





**FIGURE 5** *Left:* speed-volume relationship. *Right:* relationship of rate of speed change to volume.

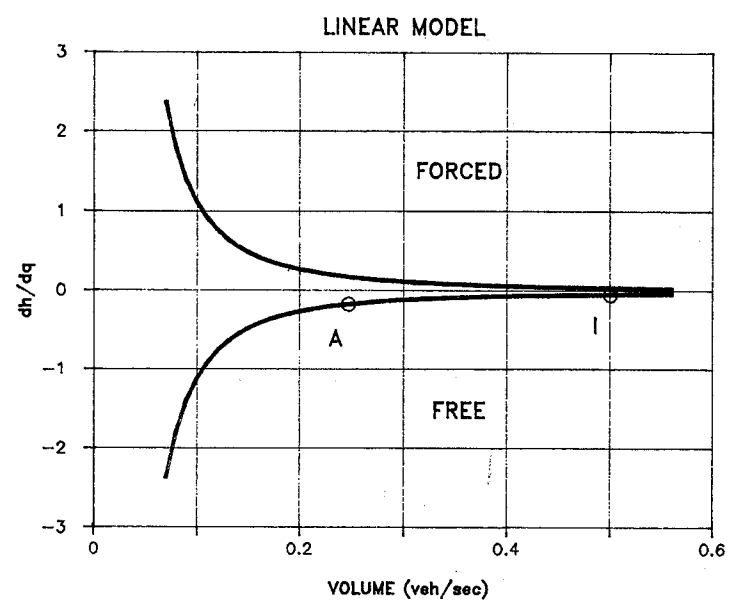
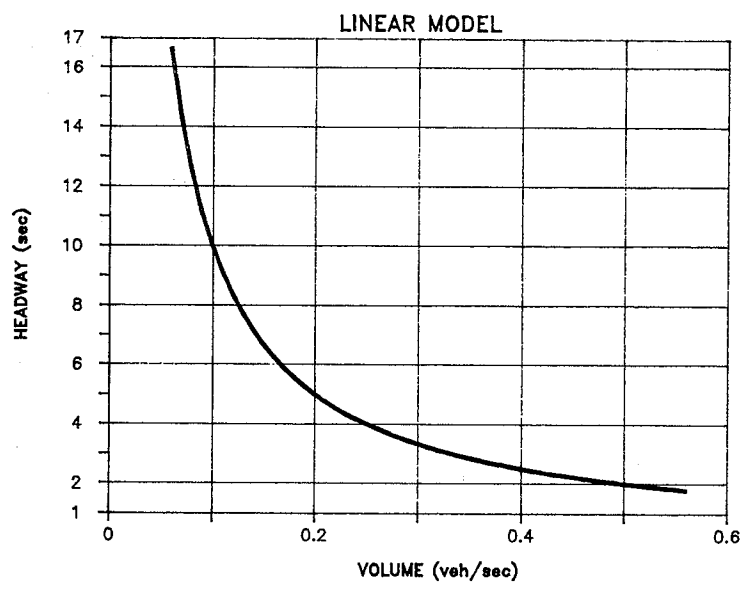


FIGURE 6 Left: headway-volume relationship. Right: relationship of rate of headway change to volume.

TABLE 1 FLOW AND SPACING AT POINT OF INFLECTION AND AT CAPACITY

Model	n	Spacing at point of inflection	point of max. q	$s_I/s_Q$	Volume at point of inflection	point of max. q	$q_I/q_Q$
exponential	n=-1	$4.48s_j$	$2.72s_j$	1.65	$0.33u_m k_j$	$0.37u_m k_j$	0.91
parabolic	n=0	$3.52s_j$	$2.25s_j$	1.56	$0.13u_f k_j$	$0.15u_f k_j$	0.9
linear	n=1	$3s_j$	$2s_j$	1.5	$0.22u_f k_j$	$0.25u_f k_j$	0.9

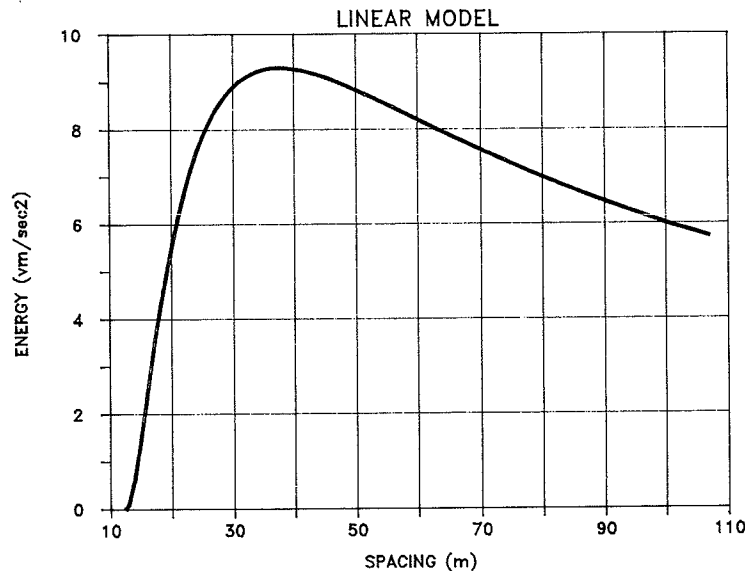


FIGURE 7 Spacing versus kinetic energy for the linear model.

TABLE 2 COMPARISON OF SPACING AT POINT OF INFLECTION AND AT POINT OF MAXIMUM KINETIC ENERGY

Model	n	Spacing at point of inflection	max. E	$s_I/s_E$
exponential	n=-1	$4.48s_j$	$7.39s_j$	0.61
parabolic	n=0	$3.52s_j$	$4s_j$	0.88
linear	n=1	$3s_j$	$3s_j$	1

field study was carried out on two straight and level three-lane urban arterials located in an intermediate-type area in the city of Montreal.

### FIELD STUDY

The arterials were chosen so as to minimize interferences and marginal and intersectional friction. The distance to the downstream intersection was about 800 meters so that the traffic flow was unaffected by the downstream traffic control devices. The observations were made at the intersection (stop line plus the width of the cross street) and at 100 meters and 200 meters from that point. Both arterials had high volumes during peak hours, with platoons ranging from 16 to 90 vehicles. The small platoons were observed at the beginning and at the end of

the peak hours. Approximately 150 platoons were observed at each arterial.

These observations were carried out with three hand-held microcomputers (Tandy 200) with external disk drive and synchronized internal clocks. A program written in BASIC allowed the identification of buses, trucks, and passenger cars. The five keys that were activated on the keyboard were identified by stickers as B (bus), T (truck), P (passenger), CB (beginning of the cycle), and CE (end of the cycle). The data on vehicle types, passage times, and beginning and end of the cycle were stored on the portable disk drive. This procedure was extremely efficient. Only 2% of all platoons had to be eliminated in the validation phase, mostly because of data-saving problems. The commercially available hand-held computers with standard keyboard can be programmed for any

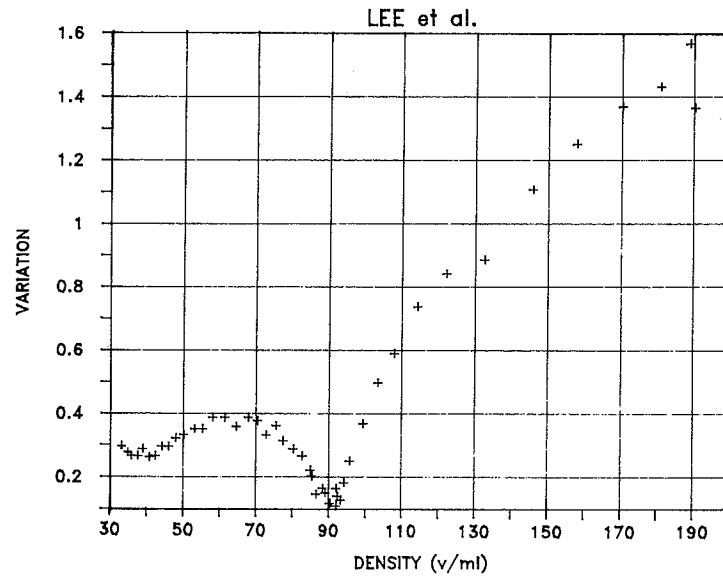


FIGURE 8 Diffusion constant versus density (7).

TABLE 3 NUMBER OF OBSERVATIONS AT TWO STUDY LOCATIONS

no of platoons	Papineau Ave		Laurentian Bld	
	observed	eliminated	observed	eliminated
	338	7	130	4
	platoon size	no of platoons	platoon size	no of platoons
	16	22		
	19	54		
	27	33		
	36	60		
	43	13	46	7
	50	13	55	12
	54	26	60	17
	61	53	67	19
	65	20	73	27
	68	27	75	12
	74	10	78	18
	76	7	81	18

data-acquisition problem by means of simple BASIC programs (see Baass (9) and Bonsall (10)). Even with the Tandy 200's small memory of 32K, this new technology provides an efficient way of acquiring traffic data.

The data were then transferred easily into an IBM PC-AT, where they were treated directly through a spreadsheet (Symphony). Table 3 shows how the validated platoons were classified into groups of similar sizes and average platoons were derived for each platoon size. Figure 9 shows such a platoon at 100 meters and at 200 meters from the stop line.

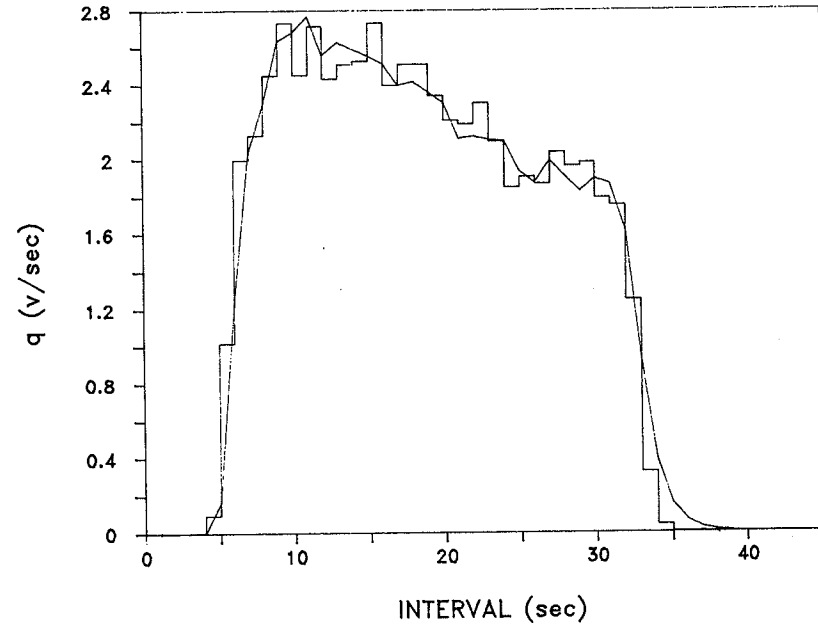
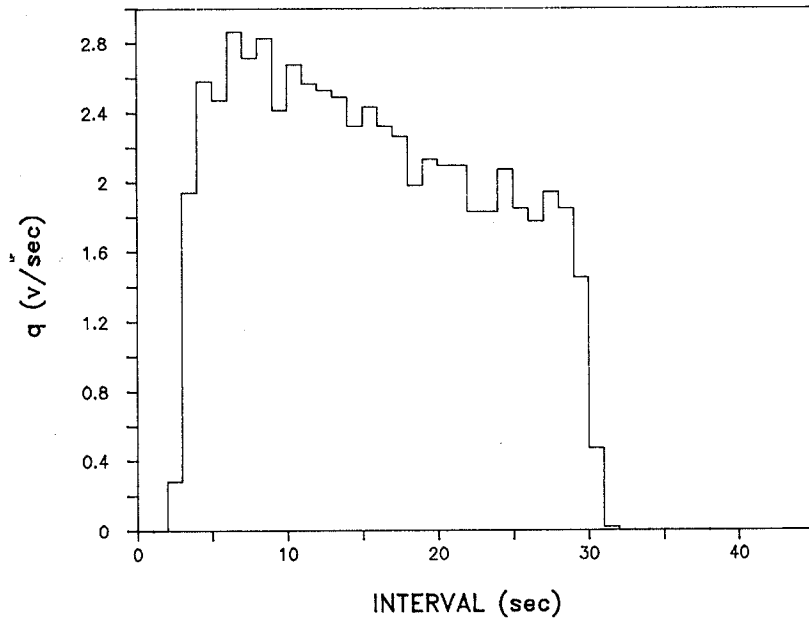
A special procedure designed on the spreadsheet was then used to simulate the platoon dispersion based on the Robertson model for different trial values of  $K$ . The parameter  $\beta$  was considered as a constant value of 0.8 in all calibrations because of the complexity of a dual calibration of  $\alpha$  and  $\beta$  at the same time (see Axhausen et al. (11)). The best fitting  $K$

was obtained for each average platoon using an indicator similar to the  $\chi^2$  by comparing observed and simulated platoon histograms. Therefore the best  $K$  values remained nearly constant for all stations.

The best  $K$  values that were obtained in this way are given in table 4, and the relations between  $K$  and the platoon size are given in figure 10. Lefebvre (12) described the experiment and the calibration procedure.

Calculating the saturation flow or level of service  $E$  at the study sites and using the cycle lengths observed permitted the drawing of figure 11, which gives the best  $K$  with relation to the volume/capacity ratio.

In both the observed cases, the observations and calibrated  $K$  values came close to the hypothesized relationship. Further analysis of the data is necessary to explain the relation between cycle duration, platoon-size, and platoon dispersion  $K$ .



**FIGURE 9** Histogram of a large platoon (68 vehicles) at 100 meters (*left*) and 200 meters (*right*) from the stop line.

TABLE 4 RESULTS OF CALIBRATION OF K IN MODEL OF ROBERTSON

Papineau Ave		Laurentian Bld	
platoon size	best K	platoon size	best K
16	0.17		
19	0.325		
27	0.35		
36	0.405		
43	0.39	46	0.13
50	0.33	55	0.18
54	0.325	60	0.16
61	0.315	67	0.17
65	0.31	73	0.13
68	0.32	75	0.13
74	0.26	78	0.1
76	0.255	81	0.13

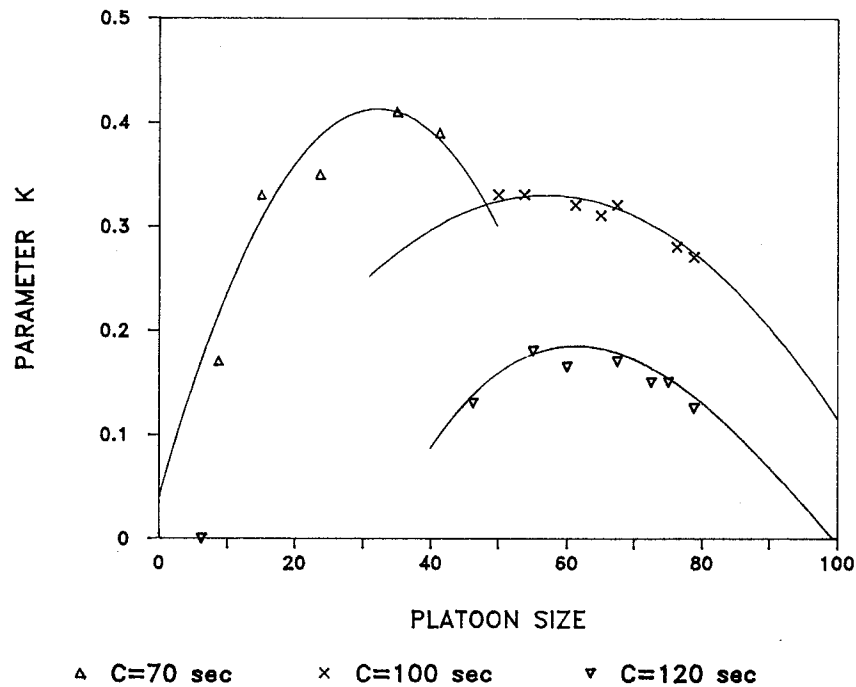


FIGURE 10 Parameter K for the two study sites in relation to platoon size.

### MICROSCOPIC SIMULATION

Microscopic simulations were carried out on a hypothetical three-lane arterial in order to further verify the postulated relationship. This arterial was chosen because it differed from the arterial streets on which the field study was conducted. Manar (13) developed a microscopic traffic simulation model based on the car-following model described by Fox et al. (14). Manar introduced into the existing program multilane simulation with lane-changing behavior (see Seddon (15)) and increased the platoon sizes to up to 110 vehicles on three lanes.

Different sizes of average platoons were obtained with the simulation model. The histograms describing the platoons were

calculated for the average platoons at the stop line and at 100 meters and 200 meters from the stop line. Platoon sizes in different runs are slightly different because this is also a random variable in the model. The 13 different platoons were then treated in the same way as the observed platoons; the best K values obtained in this way are presented in table 5.

These values are represented in figure 12 as a function of platoon size. The shape of the curve is similar to the one obtained in the field study (fig. 11); the different characteristics chosen in the simulation account for the different maximum values of K.

The simulation program produces several useful outputs. One of these is the macroscopic relationship derived from the car-following model. Figure 13 gives an example for this out-

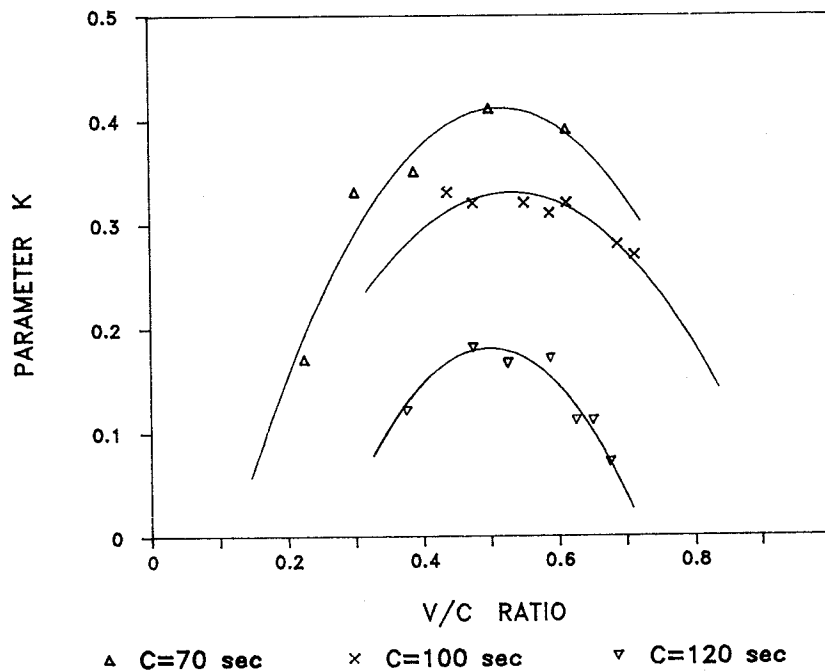


FIGURE 11 Parameter  $K$  for the two study sites in relation to volume/capacity ratio.

TABLE 5 BEST  $K$ -VALUES  
OBTAINED BY SIMULATING  
DIFFERENT PLATOON SIZES

Simulation	
platoon size	best $K$
22	0.23
28	0.14
33	0.20
38	0.17
43	0.26
48	0.31
53	0.27
62	0.48
70	0.63
75	0.70
84	0.74
87	0.48
90	0.42

put, which could be used for further analysis of the platoon dispersion.

Other important outputs are several statistics that were calculated to help in the interpretation of the simulated platoons. One of these is the number of lane changes (from left to right and from right to left). Figure 14 illustrates this variable for the simulated platoons.

As volumes increase, drivers lose the ability to maintain their desired speeds. Drivers have to reduce their speed to match the slower vehicles or, if possible, change lanes. The lane-changing possibility diminishes after a maximum value if volumes are further increased. The maximum value may

be interpreted as the point at which drivers become influenced by the presence of vehicles ahead of them. The number of overtaking and passing opportunities can be viewed as an index for friction. Even if the 1950 Highway Capacity Manual (16) gives only an example for two-lane highways, the form of the relationship in figure 15, which is reproduced here from the manual, may also be valid in multilane arterial analysis.

Two other variables produced by the simulation model are of interest—the velocity gradient defined by Helly (8) and the acceleration noise. The values of these variables are illustrated in figures 16 and 17.

The simulated platoons behaved in a way that corresponds

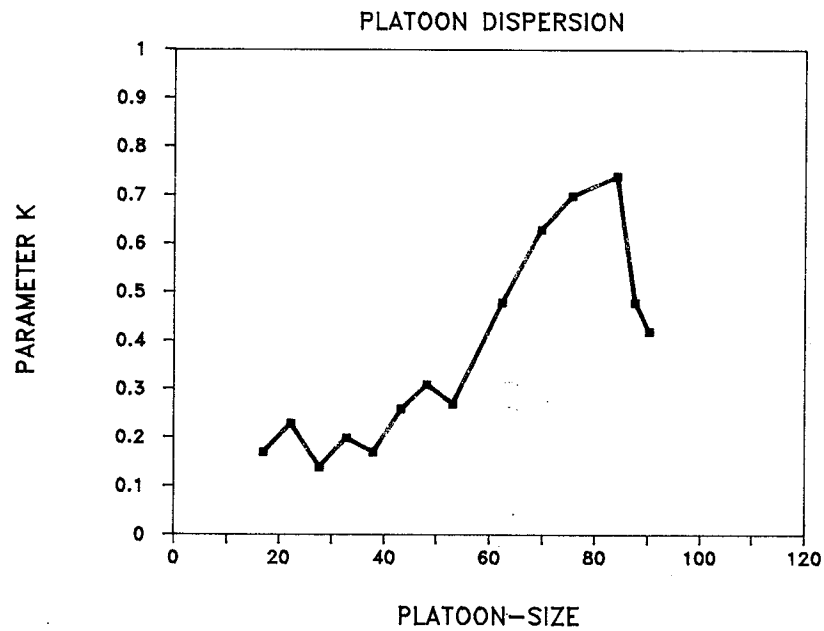


FIGURE 12 Parameter  $K$  for the simulated platoons.

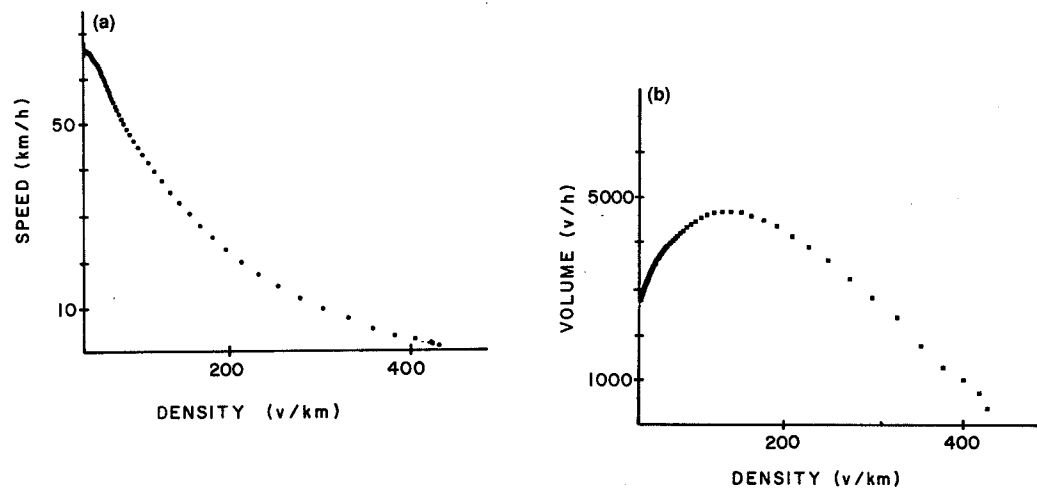


FIGURE 13 Macroscopic relationship derived from car-following model.

fairly well to the proposed relationship between platoon dispersion and traffic volume; these variables show an increase as volume increases, followed by a decrease to a minimum value near capacity.

## CONCLUSION

The degree of platoon dispersion depends on internal and external friction. The external friction is contained in the saturation flow concept. Because the internal friction depends on traffic volumes, the platoon dispersion also has to depend on service volumes. Observed and simulated platoons showed low dispersion at low traffic volumes. External friction being the same, dispersion increases as volumes increase, whereby the actual value of  $K$  depends on the kind of arterial studied. Dispersion increases further up to a maximum value as traffic volumes

increase. This point of maximum dispersion is observed at volumes of 0.6 to 0.8 of the capacity and is near the point of maximum energy. The diminishing dispersion after further volume increases can be explained by changing driver behavior at or near the point of inflection in the volume-spacing curve.

Further study should be devoted to more platoon observations at different locations and for different lane numbers and cycle lengths in order to define (if possible) a typical curve relating service volumes to dispersion. This relationship could have the following form, where  $q$  represents the volumes:

$$F = \frac{1}{1 + (a + bq - cq^2)t}$$

We would thus be able to consider platoon dispersion link by link and the results of the traffic signal coordination models would be improved.



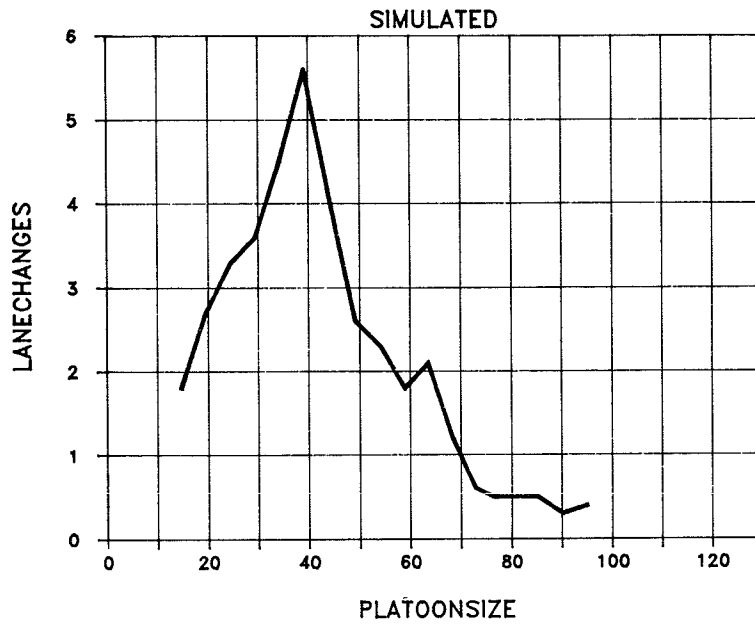


FIGURE 14 Number of lane changes versus platoon size.

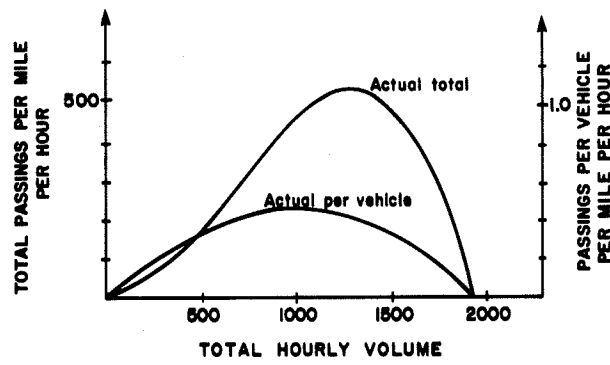


FIGURE 15 Passing opportunities on a two-lane highway (16).

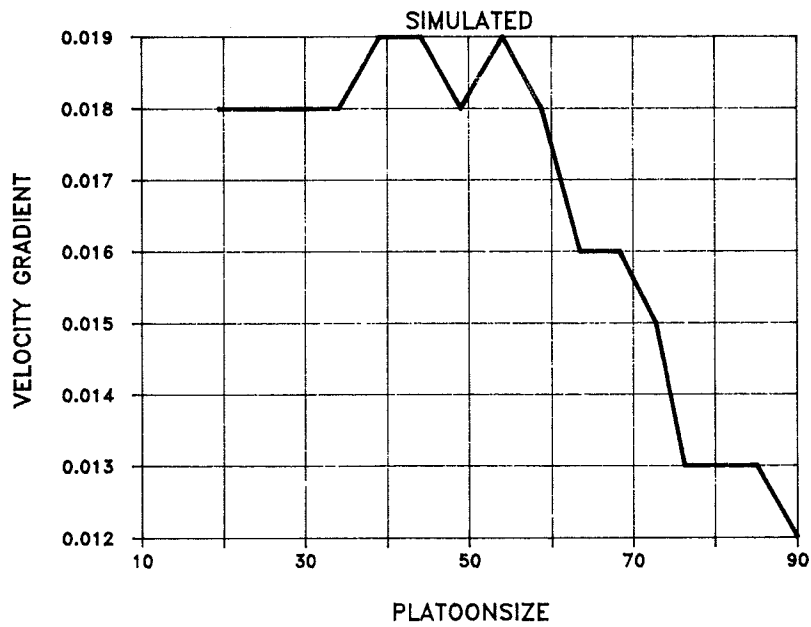


FIGURE 16 Velocity gradient.

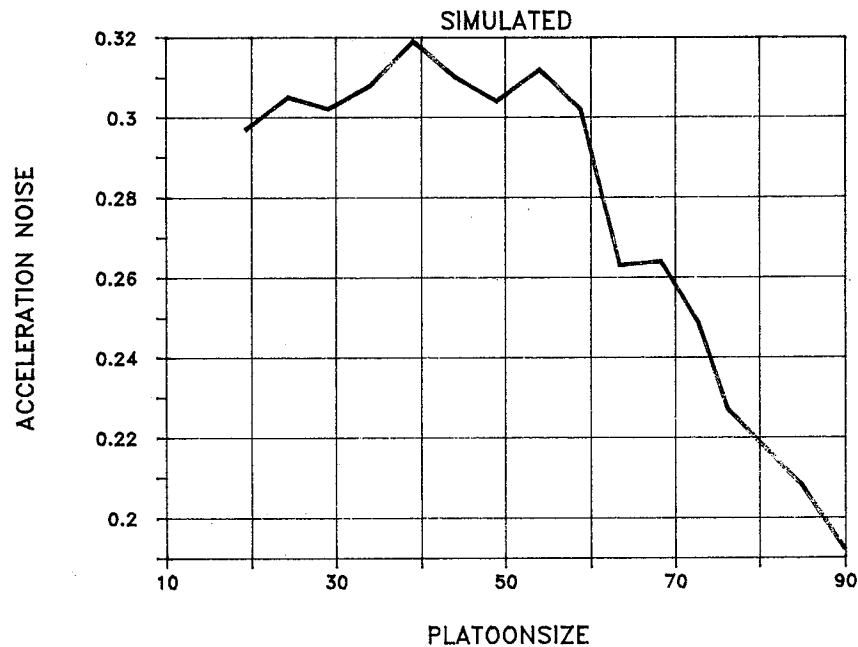


FIGURE 17 Acceleration noise.

#### REFERENCES

1. P. T. McCoy, et al. Calibration of TRANSYT platoon dispersion model for passenger cars under low friction traffic conditions. *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983.
2. A. May. A friction concept of traffic flow. *HRB Proc.*, Vol 38., 1959, pp. 493-510.
3. A. May et al. Development and evaluation of Congress Street Expressway Pilot Detection System. *Highway Research Record 21*, HRB, National Research Council, Washington, D.C., 1963, pp. 48-68.
4. W. Leutzbach. *Einfuehrung in die Theorie des Verkehrsflusses*. Springer Verlag, Berlin, 1972.
5. D. R. Drew. *Traffic Flow Theory and Control*. McGraw Hill, New York, 1968.
6. D. R. Drew, et al. Freeway level of service as influenced by volume and capacity characteristics. *Highway Research Record 99*, HRB, National Research Council, Washington, D.C., 1965, pp. 1-47.
7. J. Lee et al. Internal energy of traffic flows. *Highway Research Record 456*, HRB, National Research Council, Washington, D.C., 1965, pp. 40-49.
8. W. Helly, et al. Acceleration noise in a congested signalized environment. In *Vehicular Traffic Science* (L. C. Edie, ed.), Elsevier, New York, 1967.
9. K. Baass et al. Utilisation de micro-ordinateurs commerciaux pour la saisie des données en circulation. *Routes et Transports XVII-1*. January 1987.
10. P. W. Bonsall et al. Portable data capture devices for transport sector surveys. *Traffic Engineering and Control*, April 1987, pp. 216-223.
11. K. Axhausen et al. Some measurements of Robertson's platoon dispersion factor. Presented at the 66th Annual Meeting of the Transportation Research Board, Washington, D.C., 1987.
12. S. Lefebvre. Analyse de la dispersion des pelotons en fonction du débit. M.A.Sc. thesis. Ecole Polytechnique de Montréal, 1987.
13. A. Manar. Modele de simulation pour la dispersion des pelotons. M.A.Sc. thesis. Ecole Polytechnique de Montréal, 1987.
14. P. Fox et al. Safety in car following. Newark College of Engineering, Newark, New Jersey, 1967.
15. P. Seddon. A practical and theoretical study of the dispersion of platoons of vehicles leaving traffic signals. Ph.D. thesis. University of Salford, England, 1971.
16. U.S. Department of Commerce. *Highway Capacity Manual 1950*. Bureau of Public Roads, Washington, D.C., 1950.

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

# Analysis of Traffic Flow at Complex Congested Arterials

PANOS G. MICHALOPOULOS

**A dynamic macroscopic methodology for analyzing traffic flow at congested intersections and arterial streets is presented in this paper. This includes complexities frequently encountered in practice such as turning and optional lanes, sinks and sources, and spillback effects. The dynamics of interrupted flow under both sign and signal control are treated in an integrated fashion by considering the coupling effects of the main and side street flows. Compressibility is inherently included in the state equations employed; therefore, dispersion and compression characteristics are built into the analysis. The proposed methodology is particularly applicable to severely congested networks where spillbacks must be taken into account. Implementation of the state equations is performed numerically; this allows inclusion of stochastic and heuristic aspects for treating phenomena difficult to deal with macroscopically. The modelling and numerical treatment employed is easily implementable to microcomputers.**

Modelling of interrupted traffic flow at signalized intersections and arterial streets is frequently needed in traffic engineering practice for simulation and control or simply for analyzing a situation during the planning and design stages. More often than not this is accomplished in a macroscopic fashion to minimize the computations and obtain a better understanding of the overall behavior of the process being analyzed; this is particularly the case as the size of the network increases. However, despite the attractiveness of macroscopic models, rigorous and comprehensive dynamic treatment of complexities most frequently encountered in practice is still lacking. Such complexities include turning or optional lanes, spillback effects, unsignalized side streets merging to or diverging from the main stream flow, macroscopic treatment of actuated signals, and friction effects. The problem is more acute at high volume streets where two-dimensional modelling is needed for describing the formation and dissipation of queues and spillbacks in time and space and for taking into account the effects of downstream disturbances on upstream flow.

In view of the need for improved dynamic macroscopic modelling of congested interrupted flow, a new approach is presented in this paper. The modelling employed is two dimensional (time and space), takes compressibility into account (dispersion and compression), and is particularly recommended for medium to heavy flow situations. Moreover, it considers the interactions of side streets and turning lanes and the effects of compression and spillbacks caused by downstream congestion. Finally, it applies to sign and signal control and to isolated intersections and arterials. The proposed

methodology is numerical, i.e., it proceeds by discretizing the time-space domain in small increments and should be more intensive computationally than existing one-dimensional models used in practice. However, it is perfectly suitable for microcomputer implementation because it requires substantially less memory and computational effort than microscopic modelling.

The proposed methodology could be called "mesoscopic" as it allows inclusion of statistical and heuristic aspects for treating certain phenomena in a manner similar to a microscopic approach.

In what follows, a brief theoretical background on the state equations employed is presented with a general numerical solution method for their implementation. This is followed by the treatment of simple isolated signalized intersections. Due to space limitations, only a summary of details such as treatment of sinks and sources, exclusive turning lanes, and traffic actuated control is presented here. Extension to coordinated intersections and implementation to exemplary situations are included to demonstrate the applicability of the proposed modelling.

## BACKGROUND

Macroscopic analysis of traffic stream flow requires estimation of the three basic variables, i.e., density ( $k$ ), speed ( $u$ ), and flow rate ( $q$ ), on every point of the roadway at all times during the analysis period. From these variables the measures of effectiveness (delays, stops, total travel, and total travel time) can be derived. The most rigorous approach for calculating  $q$ ,  $k$ , and  $u$  in both time and space is by employing the law of conservation with an equilibrium speed-density or flow-density relationship to take compressibility (compression and dispersion) into account (1). An even more sophisticated treatment is to also consider acceleration and inertia of a traffic mass (2, 3), but although this approach is plausible, it has not shown significant advantages at congested flows, adding only complexity to the analysis. The methodology described here, however, can easily be generalized to include these effects (5).

Analytical implementation of the law of conservation to isolated and coordinated intersections by taking time, space, and compressibility into account has been developed recently (5-7). The advantage of these analytical results is that they visually depict the effects of downstream disturbances on upstream flow. Thus, they provide a good insight on the formation and dissipation of queues and congestion in time and space; they also demonstrate (7) that platoon dispersion and

compression is inherent in this modelling, i.e., it does not have to be induced externally. The dynamics of platoons and their behavior can in addition be studied analytically (7) for better understanding of arterial flow behavior. A disadvantage of the analytical solution lies in the oversimplifications needed in the derivation. These include simple initial flow conditions, arrival and departure patterns, absence of sinks or sources, and uncomplicated flow-concentration relationships. Most important, complexities frequently encountered in real situations such as turning lanes and side streets cannot be treated analytically. As in similar problems of compressible flow, these difficulties can be resolved by developing numerical solutions for the state equations.

The methodology proposed here is computational rather than analytical. This eliminates the disadvantages mentioned above by allowing inclusion of complexities one is likely to encounter in practice (turning lanes, sinks and sources, and spillbacks) and employment of realistic arrival and departure patterns,  $u-k$  models as well as empirical considerations. Numerical computation of  $k$ ,  $u$ , and  $q$  proceeds by discretizing the roadway under consideration into small increments  $\Delta x$  (in the order of 30 to 150 ft) and updating the values of these traffic flow variables on each node of the discretized network at consecutive time increments  $\Delta t$  (in the order of 1 sec or so).

Space discretization of a simple signalized traffic link without side streets is presented in Figure 1 in which the dashed segments represent dummy links necessary in the modelling of subsequent sections. It should be stressed that this discretization is only made for computational purposes, i.e., it is not physical. Referring to the solid segments, density on any node  $j$  except the boundary ones (i.e., 1 and  $j$ ) at the next time step  $n + 1$  is computed from density in the immediately adjacent links (both upstream and downstream  $j - 1$  and  $j + 1$ , respectively) at the current time step  $n$  according to the relationship

$$k_j^{n+1} = \frac{1}{3} (k_j^n + k_{j+1}^n + k_{j-1}^n) - \frac{\Delta t}{2\Delta x} (q_{j+1}^n - q_{j-1}^n) \quad (1)$$

in which

$$\begin{aligned} k_j^n, q_j^n &= \text{density and flow rate, respectively, on node } j \text{ at} \\ &\quad t = t_0 + n\Delta t; \\ t_0 &= \text{initial time; and} \end{aligned}$$

$\Delta t, \Delta x$  = time and space increments, respectively, such that  $\Delta x/\Delta t >$  free flow speed.

Once density is determined, speed at  $t + \Delta t$  (i.e., at  $n + 1$ ) is obtained from the equilibrium speed density relationship  $u_e(k)$ , i.e.,

$$u_j^{n+1} = u_e(k_j^{n+1}) \quad (2)$$

For instance, if Greenshields' (8) linear model is adopted, then

$$u_j^{n+1} = u_f \left( 1 - \frac{k_j^{n+1}}{k_{\text{jam}}} \right) \quad (2a)$$

where  $u_f$  is the free flow speed and  $k_{\text{jam}}$  the jam density. It should be stressed that Equation 2 is applicable for any speed density model including discontinuous ones; if an analytical expression is not available, then  $u$  can easily be obtained numerically from the  $u - k$  curve. Finally, flow at  $t + \Delta t$  is obtained from the fundamental relationship

$$q_j^{n+1} = k_j^{n+1} u_j^{n+1} \quad (3)$$

in which the values of  $k$  and  $u$  are first obtained from Equations 1 and 2. As later sections suggest, the measures of effectiveness can be derived from  $k, u$ , and  $q$ .

The above solution requires definition of the initial state of the system, i.e., the values of  $k, u$ , and  $q$  at  $t = t_0$  as well as boundary conditions, i.e.,  $k$  and  $q$  at  $j = 1$  and  $j = J$  (upstream end of the link and stopline, respectively). However, this is essential for analyzing flow regardless of the modelling and solution method, i.e., arrivals and departures at the boundaries and initial flows must always be specified. For practical implementation of Equations 1, 2, and 3 one only needs to specify arrival and departure flow rates; density at  $j = 1$  and  $j = J$  is obtained from the equilibrium  $q-k$  model. It should be noted that Equation 1 does not take into account generation or dissipation of cars; this detail is discussed later. Finally, the discretization of Figure 1 and numerical solution of this section assume all space increments  $\Delta x$  are equal. The case of variable space discretization is presented later; however, regardless of the discretization scheme the relationship  $\Delta x/\Delta t > u_f$  must be maintained at all times for convergence.

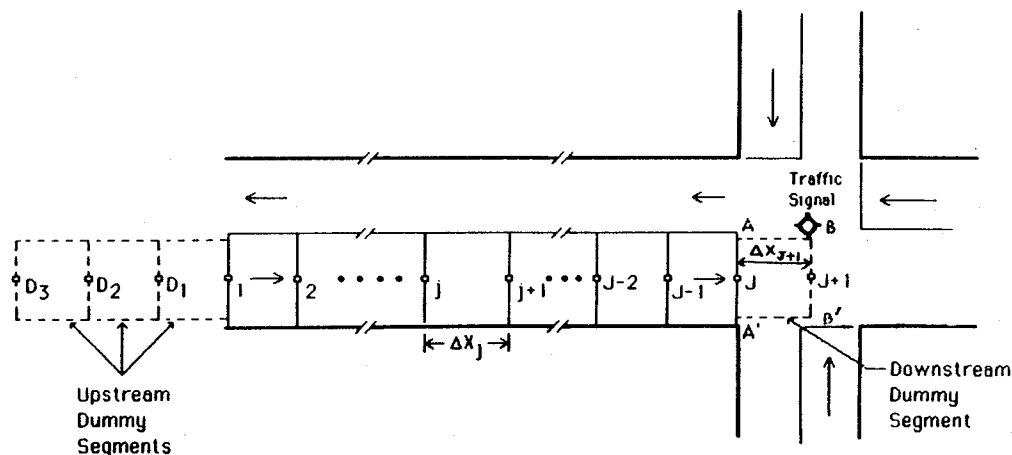


FIGURE 1 Space discretization of a simple link.

### ISOLATED INTERSECTIONS

The basic methodology of the previous section (Equations 1–3) is directly applicable to simple isolated signalized intersections without turning lanes or sinks and sources (Figure 1). Proper definition of the initial and boundary conditions is crucial for accuracy and emulating the particular situation at hand. Initial conditions depend on the state of the signal indication on the approach under consideration, and unless an actual measurement is available, a simplified assumption can be made. For instance, one could initially assume that the approach is empty at  $t = 0$  (i.e.,  $k_j^0 = 0$  at all nodes) and following an initialization interval use the flow conditions at the end of this interval as initial conditions. Another simple alternative is to assume that at  $t = 0$  the signal turns green on the approach under consideration on which an initial queue exists; if flow within the queue is uniform, then on the nodes within the queue  $q_j^0 = 0$  and  $k_j^0 = k_{jam}$  while behind the queue uniform arrival flow conditions, i.e.,  $q_j^0 = q_j^0$  and  $k_j^0 = k_j^0$ , can be hypothesized. In the testing presented in later sections initial conditions are variable.

Upstream and downstream boundary conditions in case of real time applications of the proposed methodology can be obtained on-line from loop detectors, i.e., by measuring actual arrivals and departure flow rates on nodes 1 and  $J$ , respectively. In simulation, these rates can be obtained from the probability distribution one is willing to accept. In some applications, however, it is desirable to reduce the computations related to the generation of stochastic boundary conditions. This is most easily accomplished at the downstream boundary where the variability of departures is less pronounced for the majority of the cycle due to the presence of the traffic signal. For instance, the simplest modelling alternative is the one frequently used in practice, according to which flow at  $j = J$  during the effective green time equals saturation flow as long as there is a queue (i.e.,  $q_j^{n+1} = q_{max}$  = saturation flow) and to arrival flow when the queue dissipates (i.e.,  $q_j^{n+1} = q_j^n$ ). During the effective red interval flow equals zero (right turns on red can also be considered as will be seen later). Only one state variable (flow, density, or speed) need be specified at each boundary; the remaining are obtained from the  $q$ - $k$  or  $u$ - $k$  model and the fundamental relationship  $q = ku$ .

The above and other simple common options for defining boundary conditions were tested extensively using a data base generated by the most recent version of NETSIM (9) (1986 version). Based on this experimentation, effective simplified modelling of the boundary conditions was developed and is described next.

With respect to downstream boundary conditions (stop line), during the green interval, one dummy segment of variable length is introduced immediately downstream of the stop line. This segment is shown with dashed lines in Figure 1. The introduction of this dummy segment smooths out the discharge pattern at the stop line during green, so the actual boundary during green is not the stop line but the line  $B'B$  (dummy node  $J + 1$ ) in Figure 1. At the start of green, free flow conditions are assumed at the dummy node, i.e.,  $k_{j+1}^0 = 0$  and  $u_{j+1}^0 = u_f$  while at the stop line  $k_j^0 = k_{jam}$  and  $u_j^0 = 0$ . The length of the dummy segment  $\Delta X_{j+1}$  at the start of green equals  $\Delta x$ , where  $\Delta x$  is the common discretization length used for the entire link. At every time step during green, the downstream boundary  $J + 1$  is determined from

the dynamics of flow propagation, i.e., the dummy link length is computed from the speed at which the traffic is moving at the stop line during the previous time step; thus,

$$\Delta x_{j+1}^{n+1} = u_j^n * \Delta t \quad (4)$$

The flow at the pseudo boundary  $J + 1$  (dummy node) equals the arrival flow, i.e.,  $q_{j+1}^{n+1} = q_j^n$  and  $k_{j+1}^{n+1} = k_j^n$ , while in all the remaining nodes, including the stop line, flow is determined from Equations 1–3. Here the variable dummy link requires generalization of Equation 1 for calculating flow at the stop line. This equation, including generation terms (not needed for the stop line), becomes

$$k_j^{n+1} = \frac{k_{j+1}^n \Delta x_j + k_{j-1}^n \Delta x_{j+1}}{\Delta x_j + \Delta x_{j+1}} - \frac{\Delta t (q_{j+1}^n - q_{j-1}^n)}{\Delta x_j + \Delta x_{j+1}} + \frac{\Delta t (g_{j+1}^n \Delta x_j + g_{j-1}^n \Delta x_{j+1})}{\Delta x_j + \Delta x_{j+1}} \quad (5)$$

where  $g_j^n$  represents the generation rate in segment  $j$  at the  $n$ th time step. When there is no generation from side streets, then  $g_j^n$  is zero and the last term of Equation 5 vanishes.

The above modelling implies no congestion downstream of the intersection; this event can be taken into account as in congested coordinated intersections described later. During amber time, the dummy segment is not needed, i.e., the stop line becomes the downstream boundary. Flow at this boundary reduces rapidly with time; it was found experimentally that linear reduction of flow at the stop line during amber is sufficient. Thus during yellow, flow at  $j = J$  is

$$q_j(t) = q_j(t_g) * \left[ 1 - \frac{t - t_g}{t_y - t_g} \right] \quad t_g < t \leq t_y \quad (6)$$

where

$$\begin{aligned} t_g &= \text{end of green,} \\ t_y &= \text{end of yellow, and} \\ q_j(t) &= \text{flow at the stop line at } t. \end{aligned}$$

From Equation 6, density during yellow is computed from

$$k_j(t) = k_e [q_j(t)] \quad t_g < t \leq t_y \quad (7)$$

where  $k_e(q)$  is the equilibrium  $k$ - $q$  relationship. This relationship, however, yields in general two values for  $k$  corresponding to a single  $q$  value; care should be exercised to select the one corresponding to the congested flow because at the start of yellow congestion begins to emanate from the stop line. Finally, during red, zero flow conditions prevail at the stop line, i.e.,  $k_j^{n+1} = k_{jam}$  and  $q_j^{n+1} = 0$ . If right turns on red are allowed, then  $q_j^{n+1} \neq 0$ , and this rate can either be specified or determined as in the side street flow case described in the following sections.

As mentioned earlier, at isolated approaches (or external network links) flows at the upstream boundary are more difficult to define due to the absence of signals. If no actual measurements are available (possibly on-line) for defining the upstream arrival flow pattern ( $q_1^{n+1}$ ) in some detail, then at every time step flow must be determined from a statistical distribution when high accuracy is required. The literature on selecting the appropriate distribution is quite extensive and one can select the most appropriate based on personal experience. In the testing performed here the guidelines of Gerlough et al. (10) were followed, with very satisfactory results, at least when compared with NETSIM (9). Once  $q_1^{n+1}$  is spec-

ified,  $k_1^{n+1}$  and  $u_1^{n+1}$  are determined from the equilibrium relationships employed, i.e., for Greenshields' model (8),

$$k_1^{n+1} = 0.5 [k_{jam} - (k_{jam}^2 - 4 k_{jam} q_1^{n+1}/u_f)^{1/2}] \quad (8)$$

$$u_1^{n+1} = u_f (1 - k_1^{n+1}/k_{jam}) \quad (9)$$

The above definition of the upstream boundary conditions assumes that downstream congestion at the previous time step  $n$  does not reach the upstream boundary, i.e., it assumes uncongested upstream flow conditions. When congestion reaches the upstream boundary, the link should be extended artificially by adding dummy segments  $D_1, D_2, D_3, \dots, D_i$  as shown with dashed lines in Figure 1. Extension of the upstream boundary at the next time step  $n + 1$  can be determined by checking density on the node immediately downstream (i.e., node 2 initially) at  $n$ ; if this density falls in the congested region of the  $u-k$  or  $q-k$  curve, then one dummy segment should be added upstream at  $n + 1$ . The threshold value for determining congestion is  $k_m =$  density at capacity; if  $k_{D_{i-1}} \geq k_m$ , a new dummy link is added. Boundary conditions at the first upstream dummy node are determined as before, i.e., from Equations 8 and 9. In some instances an upper limit may exist beyond which no dummy segments can be added; this could be the case when an upstream intersection is reached. In such cases the approach in question can no longer be considered isolated, i.e., the two intersections are coupled and should be treated as a system following the guidelines of later sections.

**GENERALIZATIONS AND EXTENSIONS**

The basic modelling and analysis methodology presented to this point needs further generalizations and extensions for

practical implementation to real situations. These include treatment of sinks and sources, turning lanes, spillback effects, multiple lanes, shared lanes, and coordinated intersections. Due to space limitations the methodology for treating these problems is only briefly presented in this section. A detailed presentation can be found elsewhere (11).

**Sources and Sinks**

When sources and sinks exist, the coupling effects of the main and side streets must be considered. This is accomplished by taking the state equations (conservation) of all system components and solving them simultaneously.

In Figure 2, segment  $S$  represents the case of a side street source. The state equations of the two components (main and side street) have the general form of Equation 5, but they differ in the generation term; this term is zero on the entire side street and all segments of the main street except the merging segment (node  $M$ ) where  $g \neq 0$ . Since the main street has priority, it must be realized that the stop line of the side street is really an internal boundary. The boundary flow condition on node  $J_s$  is restricted by the conflicting flow  $q_M^n$  at the merging node  $M$ ; to be sure,  $q_M^n$  determines the merging capacity at  $J_s$ . This capacity under uncongested main flow conditions was derived experimentally in the form of a curve showing the merging capacity for different main flow conditions. Similar curves can also be found in Tanner's article (12), obtained from the *Highway Capacity Manual* (13), or derived experimentally. For congested conditions, the maximum merging flow remains constant until nearly jam density, where it decreases gradually to zero. The boundary conditions at node  $J_s$  depend on the state of the queue on the side street. Generally, flow at  $J_s$  equals merging capacity, if a queue exists

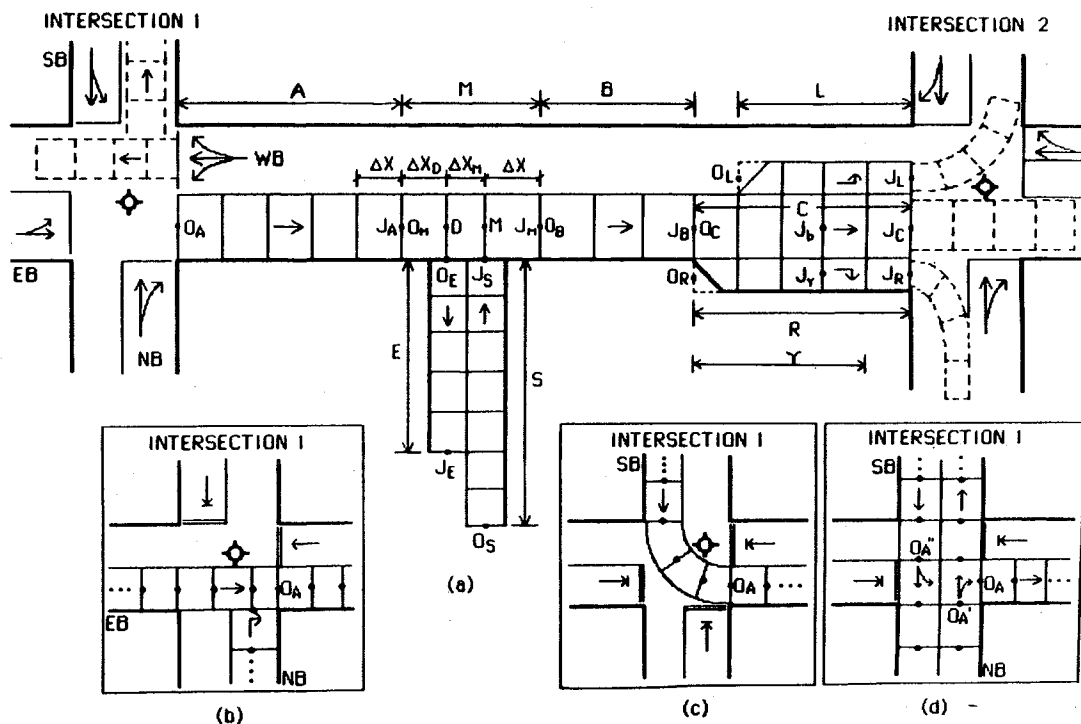


FIGURE 2 Space discretization of a complex arterial link.

on the side street, and equals arrival flow otherwise. Knowing the boundary conditions, Equations 1-3 are applied to determine flow conditions at all nodes of the side street while the generation rate needed for using Equation 5 on the main street is computed from  $q_{Js}$ .

In sinks, as in sources, the problem involves determining the flow at the internal boundary  $O_E$  (Figure 2). Furthermore, when the side street becomes congested a spillback may occur on the main street; this requires additional modelling. The former problem is resolved by adopting the usual assumption that flow at  $O_E$ ,  $q_{OE}^{n+1}$ , is a known fraction  $p = p(t)$  of the main flow at  $D$ ; however,  $q_{OE}^{n+1}$  is restricted by the capacity flow rate  $q_{max}$  at node  $O_E$ . Generally, the capacity flow at  $O_E$  is lower than the capacity of the remaining nodes of the side street due to possible pedestrian movements and the slower speed of vehicles while turning right. This capacity can be derived from the turning speed, which is the speed at capacity at node  $O_E$ . The maximum flow at  $O_E$  at every time step is restricted by the maximum speed that can be achieved at  $O_E$  by the turning flow. The speed at node  $O_E + 1$  influences this speed; therefore, it can be assumed that the free flow speed at  $O_E$  is time varying and equals to the speed at  $O_E + 1$ . The corresponding capacity flow at  $O_E$  can be estimated by continuously adjusting the  $u-k$  relationship according to this free flow speed. However, due to the earlier maximum turning speed restriction (about 13 ft per sec),  $q_{max}^{n+1}$  cannot exceed this capacity value. Thus, whenever the capacity resulting from the speed at  $O_E + 1$  exceeds this value, the latter is imposed. Following determination of  $q_{max}^{n+1}$  at  $O_E$ , not only the  $u-k$  but also the  $q-k$  relationship at this node is adjusted accordingly. When congestion is encountered on the side street, the diverging node  $D$  becomes an internal boundary in addition to the boundary  $O_E$ . Therefore, the flow at  $D$ ,  $q_D^{n+1}$ , is restricted by the through-portion flow  $(1 - p)q_D^n$ , which is allowed to proceed on node  $D$ . If congestion exists on the side street, the flow at  $D$  will decrease, and density will be increasing accordingly because the dissipation rate will decrease. As a result, congestion will begin to build up behind node  $D$  and a spillback will occur.

The simultaneous presence of a source and a sink (Figure 2) necessitates consideration of all three flows, namely the through, merging (from node  $J_s$  to  $M$ ), and diverging (from node  $D$  to  $O_E$ ), as well as their interaction. As Figure 2 illustrates, for increased generality, the network under consideration is equally discretized in space except at nodes  $D$  and  $M$  where  $\Delta x \neq \Delta x_D \neq \Delta x_M$ . This necessitates a slight modification in the calculations presented to this point for the interrupted flow region. The changes required are

At  $j = D$  (diverging area):

$$k_D^{n+1} = \frac{k_M^n \Delta x_D + k_{D-1}^n \Delta x_M}{\Delta x_D + \Delta x_M} - \frac{\Delta t(q_M^n - q_{D-1}^n)}{\Delta x_D + \Delta x_M} - g_D^n \Delta t \quad (10)$$

where the dissipation term is  $g_D^n = q_{OE}^n / \Delta x_D$ .

At  $j = M$  (merging area):

$$k_M^{n+1} = \frac{k_{M+1}^n \Delta x_M + k_D^n \Delta x}{\Delta x_M + \Delta x} - \frac{\Delta t(q_{M+1}^n - q_D^n)}{\Delta x_M + \Delta x} + g_M^n \Delta t \quad (11)$$

where the generation term is  $g_M^n = q_{Js}^n / \Delta x_M$ .

At  $j = D - 1$  (the node immediately upstream of  $D$ ):

$$k_{D-1}^{n+1} = \frac{k_D^n \Delta x + k_{D-2}^n \Delta x_D}{\Delta x + \Delta x_D} - \frac{\Delta t(q_D^n - q_{D-2}^n)}{\Delta x + \Delta x_D} \quad (12)$$

At  $j = M + 1$  (the node immediately downstream of  $M$ ):

$$k_{M+1}^{n+1} = \frac{k_{M+2}^n \Delta x_M + k_M^n \Delta x}{\Delta x_M + \Delta x} - \frac{\Delta t(q_{M+2}^n - q_M^n)}{\Delta x_M + \Delta x} \quad (13)$$

At all the remaining internal nodes, the densities are obtained from Equation 1. Finally, all the corresponding speeds and flows for the above cases are obtained in the usual manner, i.e., from Equations 2 and 3.

All the needed boundary conditions, i.e., the arrival patterns at nodes  $O$  and  $O_s$  and the departure patterns at nodes  $J$  and  $J_s$ , are determined or specified as before. Further, spillback effects at node  $D$  are treated as sinks, while flow at the internal boundaries  $J_s$  and  $O_E$  is determined dynamically as described previously. Finally, it should be repeated that to keep the numerical solutions within reasonable bounds, it is necessary to maintain the relationship  $\Delta x / \Delta t > u_f$ ; this restriction also applies to  $\Delta x_D$  and  $\Delta x_M$ .

### Turning Lanes

Turning lanes frequently encountered in practice present additional modelling difficulties. In this section, treatment of exclusive lanes is summarized. This involves modelling of diverging dynamics and appropriate definition of boundary conditions. Because traffic in turning lanes is diverted from the main stream, it is logical to treat the lanes as simple sinks. However, due to the length of these lanes and possible conflicts at their downstream end, they should be treated differently. As before, the state equations of the main flow and turning lanes must be solved simultaneously and their interactions (exchange of flow and momenta) considered. The conservation equation as described here takes into account the exchange of flow. Exchange of momenta must be considered only with the high order continuum models and is incorporated in the momentum equation as described in earlier publications (4, 14).

When turning lanes exist, exchange of flow between the turning and through lanes is included in the generation rates  $g(x, t)$  of the conservation equation. In the previous section a similar procedure was introduced for traffic sinks. However, the effective diverging area, and hence the computation of dissipation rate, was limited to only one segment  $\Delta x$ .

Figure 2 presents the space discretization of an intersection approach with a right turning lane  $R$ . In general, diversion of flow occurs in section  $r \leq R$  of the main stream and the turning lane. Dynamic description of flow in the main lane and the turning lane begins by considering the three streams involved, i.e., the total flow  $q$  in the through lanes, its diverging component  $q_d$ , and the turning lane flow  $q_r$ . Simultaneous solution of the state equations governing these streams requires knowledge of the generation and dissipation rate  $g(x, t)$ , which is a function of both  $q$  and  $q_r$ . From the physics of the problem the generation rate of any segment  $j$  in section  $r$  at  $t = n\Delta t$ , i.e.,  $g_j^n$ , should be a function of the diverging flow component  $q_{j-1,d}^{n-1}$  of segment  $j - 1$  at  $t = (n - 1)\Delta t$ . The following

relationship can be derived for dynamic determination of the generation rate at turning lanes (11):

$$g_j^n = \frac{q_{j-1,d}^{n-1} S_j^{n-1}}{\Delta x \sum_{i=j}^{J_r} S_i^{n-1}}$$

$$= \frac{q_{j-1,d}^{n-1} (k_{jam} - k_{j,r}^{n-1})}{\Delta x \sum_{i=j}^{J_r} (k_{jam} - k_{i,r}^{n-1})} \quad (14)$$

where

$S_j^{n-1}$  = storage space available in segment  $j$  in  $r$  at  $t = (n - 1)\Delta t$ ;

$k_{j,r}^{n-1}$  = density of node  $j$  in  $r$  at  $t = (n - 1)\Delta t$ ; and

$q_{j-1,d}^{n-1}$  = diverging flow component.

From the above expression and Equation 5, a computational algorithm for dynamic calculation of flow on the through and turning lanes has been developed and described in detail (11). This takes into account both the interactions between the main and turning flows and the proper definition of boundary conditions at the stop line. For the through flow this is accomplished as in "Isolated Intersections." For the right turning flow, however, node  $J_R$  (Figure 2) should be treated in a manner similar to that of node  $O_E$  (described earlier) as the cross street flow affects right turning capacity.

As mentioned earlier, the modelling presented to this point also applies to left turning lanes. The only difference lies in defining the boundary conditions at the stop line of the turning lane, during the green interval. If traffic flow on this lane moves during an exclusive phase, then flow at the stop line is defined as in the through case described in "Isolated Intersections." Naturally, due to the lower turning speed, saturation flow and therefore capacity of the turning lane is lower than the through lanes; estimation of this capacity can easily be obtained from the *Highway Capacity Manual* (13). If flow on the exclusive lane moves concurrently with opposing flow, then left turning saturation flow is lower and at each time increment can be determined from the opposing flow. Estimation of left turning saturation flow with opposing traffic can also be obtained from the *Highway Capacity Manual* (13, Figure 10-3) or related literature (12, 15). Thus, at each time step, if a queue exists, flow at the stop line of the turning lane will equal the opposed left turning saturation flow. Following queue dissipation, arrival flow at the immediately upstream node must be compared to saturation flow at that instant, and the lowest of the two will prevail.

### Spillback Effects

When the flow rate on node  $J_R$  drops during green or the turning volumes are high, spillback from the turning lanes to the main stream is possible. In this case congestion propagates upstream of the stop line (node  $J_R$ ), and as density in the right lane increases, the generation rate decreases. Therefore, density in the through lanes will increase as flow on the right lane becomes congested. As an extreme example, assume that due to congestion in the cross street, flow at the stop line of the turning lane is very low, resulting in compact queueing ( $k \approx k_{jam}$ ) in this lane. As a result, the generation rate will be very low ( $g \approx 0$ ), and the total flow at node  $J_B$  will prop-

agate to the downstream node  $J_B$  almost unchanged, provided flow at  $J_B$  is not disrupted by downstream congestion. This implies that at congested turning flows, the spillback mechanism provided by the main turning lane algorithm is weak and should be improved. For this reason, an improved spillback algorithm was developed (11). Briefly, the algorithm ensures that at each time step the total through volume at  $J_B$  and the remaining through volume in Section  $C$  does not exceed the through volume passing node  $O_c$ .

### Multilane and Shared Lanes Treatment

Treatment of flow on a lane-specific basis may not be needed, especially if turning lanes have already been taken into account, for if we exclude turning movements (considered earlier), little additional lane changing should be expected at the intersection. Thus, aggregate analysis of all through lanes should suffice in most practical situations. Lane-by-lane analysis can also be performed, if desired, using the general methodology presented earlier; such treatment could be justified in segments between intersections.

Analysis of multilane dynamics requires consideration of the state equations of all lanes; in the simplest case this implies one conservation equation per lane. The exchange of flow between neighboring lanes can be included in the generation term of the conservation equation. This exchange can be obtained from the observation that it is related to the density differences among lanes and other location-specific parameters. A complete modelling of multilane dynamics with its numerical implementation is presented in an earlier publication (14).

When no exclusive turning lanes exist, optional or shared lanes serving more than one movement are often employed. This situation is depicted on the westbound approach of the first intersection in Figure 2. Modelling during the red and yellow phases does not present any difficulty as the stop line represents a boundary on which flow is defined as in "Isolated Intersections." During green, this line is no longer a boundary and is either extended by a dummy segment (isolated case) or replaced by the upstream node of the downstream link. In either case the flow in the segment immediately after the stop line includes two dissipation terms, one for each movement. The dissipation term corresponding to the right turning movement is defined as in sinks ("Sources and Sinks"). For the left turns, the dissipation term at each instant is determined from the lowest of two values: (a) the left turning demand, which is a fraction of the total flow at the stop line, and (b) the left turning saturation flow, which is time varying when moving concurrently with opposing traffic and is determined as in left turning lanes.

When congestion builds up at the cross street, the right or left turning generation terms will reduce. To ensure spillback effects are considered, flow at the node downstream of the stop line should not be allowed to exceed the through demand, i.e., this node may become an internal boundary during green. In this case the spillback mechanism described earlier applies.

### COORDINATED INTERSECTIONS

The modelling presented to this point can easily be extended to coordinated intersections. This requires modifications to



the boundary conditions at the upstream end of the coordinated links. Referring to Figure 2, it can easily be realized that the upstream boundary of a signalized link is no longer an external boundary. For instance, consider node  $O_A$ . During the EB green phase at intersection 1,  $O_A$  is treated as any ordinary node, i.e., flow on  $O_A$  is obtained from Equations 1–3 by adding one or more segments connecting  $O_A$  to the stop line upstream as shown in Figure 2b. If exclusive lanes exist on the upstream approach, flow on the nodes falling within the intersection is also computed from Equations 1–3; otherwise, if shared lanes exist, the dissipation terms of the turning movements must be considered (“Multilane and Shared Lanes Treatment”). If right turns on red are allowed from the NB approach, they can be treated as described in “Sources and Sinks” (i.e., as in sources); in such a case the generation term should be added to node  $O_A - 1$ .

Following the EB green phase at intersection 1, several phasing possibilities exist with varying effects on  $O_A$ . Considering the possibility of an exclusive left turning phase from the SB approach, Figure 2c applies, i.e., flow is again computed by connecting  $O_A$  to the stop line of the left turning movements and employing Equations 1–3. During the NB-SB green phase the right turning flows from the NB approach can be treated as in “Turning Lanes” (sinks); in such case node  $O_A$  is an internal boundary analogous to  $O_E$  of section  $E$  in Figure 2a. Stated otherwise, the sink treatment applies in this case where the main flow is in the NB direction. The only complexity is introduced when left turns from the SB flow are also allowed simultaneously with the right turns and through movement NB. This situation requires further modelling. Referring to Figure 2d, one sees the total flow at  $O_A$  consists of the right turning component of the NB stream and the left turning component of the SB stream. In this case flow at  $O_A'$  and the right turns have priority and are treated as before, i.e., nodes  $O_A'$  and  $O_A$  are treated as nodes  $D$  and  $O_E$  in Figure 2a, respectively. Flow at  $O_A''$  is restricted by the total flow at  $O_A'$  and is determined as in the optional left turning case. Thus, at each time step total flow at  $O_A$  will be the sum of the turning flows at  $O_A'$  and  $O_A''$ . Because capacity at  $O_A$  is restricted by the flow at  $O_A + 1$  (see “Sources and Sinks” for sinks), capacity at  $O_A$  restricts flow at  $O_A''$  first and then at  $O_A'$ . When these restrictions are in force, nodes  $O_A''$  and  $O_A'$  become internal boundaries to allow propagation of spillbacks.

## TEST RESULTS

Testing of all the modelling details presented here requires substantial data collection at many locations over a reasonably long time. Due to the limited resources available, extensive model testing could not be performed. As an alternative to field data collection, microscopic simulation was employed to generate the data base necessary to judge the realism and estimate the expected accuracy of the proposed methodology. More specifically, the most recent version of the NETSIM (9) program was employed for this purpose. This program is reasonably well established and documented and has been extensively tested over the many years of its development. In addition to the comparisons against the simulated data, limited testing with field data was also performed and is presented in an earlier publication (16). In these earlier tests, nonlinear

equilibrium  $u-k$  models were also employed including discontinuous ones, and it was concluded that although proper  $u-k$  selection can improve model performance, the linear alternative leads to acceptable accuracy levels. Prior to the comparisons with the data base, model performance was initially determined by inspection of the results, including visual checking of  $q$ ,  $k$ , and  $u$  plots versus time and distance. This was very effective for screening many modelling alternatives that led to unreasonable results.

The measures of effectiveness employed to determine the overall model accuracy and the method of their computation are as follows:

Total Travel (in vehicle-miles):

$$TT = \sum_{j=1}^J \sum_{n=1}^N q_j^n \Delta t \Delta x$$

where  $J$  is the total number of nodes (excluding the dummy segment) and  $N$  is the simulation time/ $\Delta t$ .

Total Travel Time under Interrupted Conditions (in vehicle-minutes):

$$TTTi = \sum_{n=1}^N \sum_{j=1}^J k_j^n \Delta x \Delta t$$

Total Travel Time under Uninterrupted Flow Conditions (in vehicle-minutes):

$$TTTu = \sum_{n=1}^N \sum_{j=1}^J k_{j,u}^n \Delta x \Delta t$$

where  $k_{j,u}^n$  is the density that would have been realized at node  $j$  under ideal conditions, i.e., if the signal were not present. This is obtained by assuming continuous green at the stop line of the approach being considered.

Delay (in vehicle-minutes):

$$\text{Delay} = TTTi - TTTu$$

Average Speed (in mph):

$$u = TT/TTTi$$

Total Arriving Volume (in vehicles):

$$A = \sum_{n=1}^N q_1^n \Delta t$$

Total Departing Volume (in vehicles):

$$D = \sum_{n=1}^N q_J^n \Delta t$$

The deviations of the above measures of effectiveness from the NETSIM estimates were computed and the percentage difference (PD) between them determined. In the results presented next, a positive PD indicates that the modelling proposed here overestimates the corresponding MOE compared to NETSIM; conversely, a negative PD indicates underestimation. Finally, it should be noted that the space increment necessary for obtaining reasonable accuracy was found to be in the range of 30 to 150 ft.

Space limitations do not allow presentation of all test results. For this reason, only the comparisons from two representative situations are presented. The first situation represents a straight

through link without turning movements under light, medium, and heavy flow conditions. The second one represents a coordinated link without turning movements, under varying offsets. In the second situation, for easy intuitive inspection, it is assumed no turning movements enter the link from the cross street at the upstream intersection.

#### Isolated Case

In this case, all through lanes are treated in an aggregate fashion; thus, the results presented and the demand pattern are per lane estimates. The link length is 2,600 ft. The arrival flow changes from 630 vehicles per hour/lane to 900, 1,170, and 360 vehicles per hour/lane every 15 min. These changes replicate a realistic peak hour demand pattern. Assuming a two-phase operation, a cycle of 60 sec, and a yellow interval of 3 sec, the green time of the approach under consideration is varied three times from 39 sec to 27 sec, generating light, medium, and heavy congestion. The average degree of saturation  $X$  corresponding to these timing plans and demand pattern is 0.66, 0.85, and 0.94, respectively. Initial conditions

in each case were determined from the NETSIM (9) program following an initialization interval.

In earlier experimentation (4) the optimal step size was  $\Delta x = 50$  ft and  $\Delta t = 1$  sec; this step size was adopted here. Further, although nonlinear equilibrium  $u-k$  models can result in higher accuracy (4), the linear model was employed here to demonstrate the robustness of the proposed treatment. The model parameters, i.e., the free flow speed  $u_f$  and jam density  $k_{jam}$ , were assumed to be 34 mph and 212 vehicles per mile/lane, respectively.

Table 1 presents the results obtained from the comparisons between the model and NETSIM estimates following 1 hr of analysis. The right half of the table (columns 6–10) presents results obtained by generating the arrivals according to the stochastic process described in "Isolated Intersections" using the 15-min averages previously mentioned. The left half (columns 1–5) of the table corresponds to the NETSIM-generated arrivals that were averaged every 3 min and used for obtaining the model results. In both cases the NETSIM estimates are the same because only the boundary conditions were changed for testing the modelling proposed earlier. As expected, the

TABLE 1 COMPARISONS OF MODEL- AND NETSIM-GENERATED RESULTS

MOE	CASE	NETSIM generated arrivals					Model generated arrivals					
		X=0.66 (1)	X=0.85			X=0.94 (5)	X=0.66 (6)	X=0.85			X=0.94 (10)	
			Cap=1800 Uf=34 (2)	Cap=1800 Uf=32 (3)	Cap=1636 Uf=34 (4)			Cap=1800 Uf=34 (7)	Cap=1800 Uf=32 (8)	Cap=1636 Uf=34 (9)		
TT (veh-mi.)	NETSIM	377.44	377.25	377.39	343.19	376.98	377.44	377.25	377.39	343.19	376.98	
	Model	399.30	401.89	401.46	372.19	383.76	400.87	393.86	416.19	366.78	417.96	
	PD (%) #	(5.8)	(6.5)	(6.4)	(8.4)	(1.8)	(6.2)	(4.4)	(10.3)	(6.9)	(10.9)	
TTTi (veh-min)	NETSIM	895.61	1726.22	1715.51	1543.74	2715.47	895.61	1726.22	1715.51	1543.74	2715.47	
	Model	1026.23	1822.75	1784.84	1620.41	2883.83	1034.98	1787.05	1920.89	1695.85	3126.95	
	PD (%)	(14.6)	(5.6)	(4.0)	(5.0)	(6.2)	(15.6)	(3.5)	(12.0)	(9.8)	(15.2)	
TTTu (veh-min)	Model	806.84	806.78	848.00	746.64	760.91	811.33	791.31	882.15	740.11	939.94	
	Delay (veh-min)	NETSIM	223.70	1063.03	1009.47	954.84	2073.89	223.70	1063.03	1009.47	954.84	2073.89
	Model	219.39	1015.97	936.85	873.77	2122.92	223.66	995.73	1038.74	955.73	2187.01	
PD (%)	(-1.9)	(-4.4)	(-7.2)	(-8.5)	(2.4)	(0.0)	(-6.3)	(2.9)	(0.1)	(5.4)		
Average Speed (mph)	NETSIM	25.30	13.14	13.23	13.39	8.38	25.30	13.14	13.23	13.39	8.38	
	Model	23.35	13.23	13.50	13.78	7.98	23.24	13.22	13.00	12.98	8.02	
	PD (%)	(-7.7)	(0.7)	(2.0)	(2.9)	(-4.8)	(-0.1)	(0.6)	(-1.7)	(-3.1)	(-4.3)	
Total Arrivals (veh)	NETSIM	763.50	763.10	763.20	694.30	763.00	763.50	763.10	763.10	694.30	763.20	
	Model	763.70	763.80	753.80	703.60	746.70	767.00	746.20	782.20	699.50	894.60	
	PD (%)	(0.0)	(0.0)	(-1.2)	(1.3)	(-2.1)	(0.4)	(-2.2)	(2.5)	(0.7)	(17.2)	
Total Departures (veh)	NETSIM	769.50	769.40	770.20	700.30	763.40	769.50	769.40	770.20	700.30	763.40	
	Model	843.00	780.50	782.60	718.90	735.10	844.30	773.40	803.40	710.10	754.90	
	PD (%)	(9.6)	(1.4)	(1.6)	(2.7)	(-3.7)	(9.7)	(0.5)	(4.3)	(1.4)	(-1.1)	

\* Results above correspond to one hour simulation.

# PD (%) = percentage of difference from data = (Model-NETSIM)/NETSIM \* 100%

results of the left half of the table are closer to the NETSIM data, although both are equally satisfactory.

Columns 2–4 and 7–9 corresponding to  $X = 0.85$  include two additional values of  $u_r$  and capacity affecting the equilibrium  $u-k$  relationship to demonstrate the model sensitivity to this relationship. As the table suggests, although model performance deteriorates, the results are still acceptable. To be sure, all the estimates obtained by the model are satisfactory as they deviate mostly well below 10 percent from the NETSIM data. When the degree of saturation increases, model performance improves; this suggests that the proposed modelling is more appropriate for congested flows.

Before concluding, it should be noted that with respect to uninterrupted total travel time  $TTTu$ , the differences between the model results and NETSIM were more pronounced. This is because of the difference in the method of calculating  $TTTu$ . According to NETSIM,  $TTTu$  or “ideal travel time” is estimated according to a “target speed” for the link; however, no further information on this speed is included in the program documentation, and nothing is mentioned about the existence of the signal and how it is taken into account in the calculations.

### Signalized Links

During the testing of oversaturated coordinated links, it was quickly realized that for the reasons mentioned later in this section, NETSIM in its present form is not suitable for generating a reliable data base, i.e., it does not treat oversaturated signalized links effectively. Therefore, the effectiveness of the model can only be judged by intuitive inspection of the results. Two test cases are presented here: one demonstrating a spillback to the upstream link and another without spillback. As in the isolated case, all through lanes are treated in an aggregate fashion. As such, the results presented here pertain to per lane estimates. A two-phase operation of a one-way signalized link is assumed with a base cycle length of 150 sec and a yellow interval of 3 sec. The layout is as in the EB direction of Figure 2a; however, no side streets and turning lanes are taken into account in these test cases. The arrival flow begins at 864 vph/lane (capacity is 1,800 vph/lane) for 15 min and drops to 576 vph/lane for another 15 min. The green times are 90 sec and 60 sec at the first and second intersections, respectively. The average degree of saturation corresponding to this time plan and demand pattern is 0.80 for the upstream link and 1.20 for the downstream link. A two-cycle initialization period, with a flow of 720 vehicles per hour/lane, was used to define initial conditions. A free flow speed of 40 mph and a jam density of 180 vehicles per mile/lane were assumed. The step size for both cases is  $\Delta x = 60$  ft and  $\Delta t = 1$  sec. As noted previously, no turning movements are assumed in this example; this implies no input to the upstream end of the signalized link during red, for easy intuitive inspection of the formation and dissipation of congestion.

In the first case, the signalized link length is assumed to be 2,000 ft; this represents the case in which no spillback occurs. Spillback results when the queue from the downstream link extends into the upstream link. In the second case, the signalized link length is 1,500 ft; this represents spillback of congestion during the test period. In both cases, the offsets are increased in steps from zero with a step size approximately one quarter of a cycle. The test runs are for a period of half

an hour. The results indicated that in the second case, the queue from the downstream signal backs up into the upstream link in the fourth cycle. The spillback recedes after five cycles (i.e., in the ninth cycle) due to the lower arrivals into the system after the first 15-min. interval.

The test runs also indicated that the offsets do not significantly affect the model results (MOEs) such as total travel, total travel time, and average speed, while total arrivals and departures are the same. When an attempt was made to verify this result by NETSIM, it was realized that this program presently cannot effectively treat oversaturated signalized links. This is because the NETSIM results demonstrated great variability with change in offsets. This latter result does not agree with experience and intuition according to which at high saturation levels, such as those presented, offsets should not produce such large differences, i.e., they should not significantly affect the MOEs. Part of the problem lies with the NETSIM program, which presently does not deal with severe congestion. For instance, when the upstream end of the link is reached, NETSIM begins to store cars vertically ignoring space and spillback effects. Improvements to the NETSIM program are being made to take spillback effects and severe congestion into account.

A visual depiction of the model's performance is seen in Figure 3, which shows the density plots versus space and time for both cases. The plot includes the cycles at which congestion is maximum. In the first case congestion barely reaches the upstream intersection, while in the second case it spreads to the upstream link. The blank spaces in Figure 3 are generated because, as explained earlier, no input occurs during the red phase to the signalized link to allow some dissipation of congestion. In this manner the trajectory of the queues becomes clearer and their evolution easier to track. Finally, compression and dispersion characteristics are manifested.

### CONCLUDING REMARKS

In this paper an attempt was made to treat traffic flow at congested intersections and arterials in a unified and consistent macroscopic fashion. This was done without ignoring important components and characteristics of urban streets and making the oversimplifications usually encountered in practical and theoretical models. Major advantages of the approach taken here are the explicit inclusion of both time and space and compressibility characteristics. Unlike empirical models, these characteristics are not induced artificially, but are inherent in the modelling.

Another consideration concerning the usefulness of the modelling presented here is that it can be implemented in microcomputers. This is because of the simplicity of the numerical schemes employed and the macroscopic nature of the models, which allow efficient computations and minimal memory requirements. The test results were in fact obtained by implementing the proposed methodology in IBM-PC-based software. Similar modelling was employed in an earlier interactive microcomputer-based simulation program for freeways (17).

The testing presented here and testing against actual field data presented in an earlier publication (16) are limited. Although this earlier testing included more complex equilibrium models (nonlinear and discontinuous ones), more testing is needed for model refinement and verification. This could

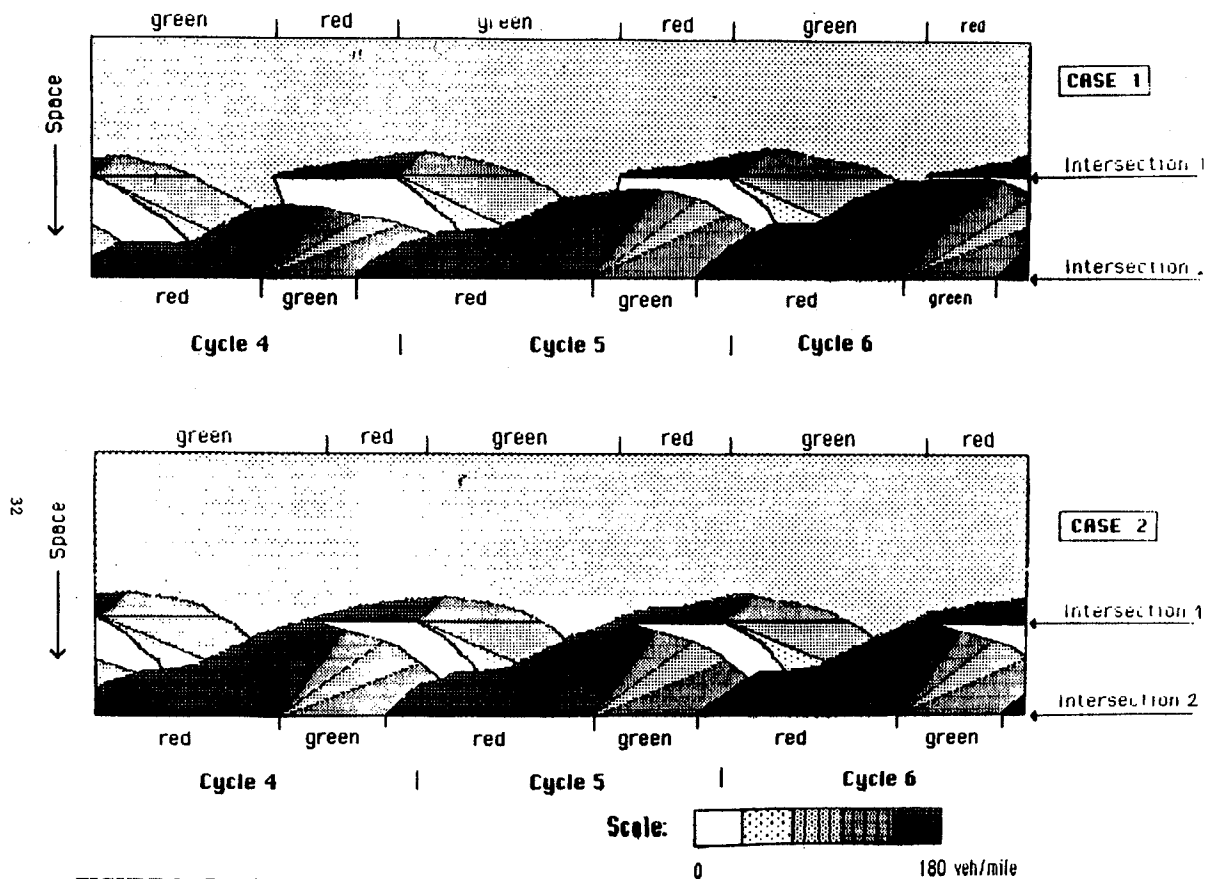


FIGURE 3 Density plots demonstrating propagation of congestion.

lead to simplifications that are probably necessary for more widespread practical implementation of the methodology presented here. Although the computational requirements of the proposed modelling are substantially lower than those of microscopic models, they are probably higher than empirical models. A sacrifice in computational effort, however, should be expected for more detail, accuracy, and realism.

#### ACKNOWLEDGMENTS

Financial support for this research was provided by the National Science Foundation and by the Exxon Petroleum Violations Escrow Fund as administered through the U.S. Department of Energy.

#### REFERENCES

1. M. H. Lighthill and G. B. Whitham. On Kinematic Waves-II. A Theory of Traffic Flow on Long Crowded Roads. *Proceedings, Royal Society (London)*, A229, No. 1178, 1955, pp. 317-345.
2. H. J. Payne. Models of Freeway Traffic and Control. *Proc., Mathematics of Public Systems, Simulation Council*, Vol. 1, No. 1, 1971, pp. 51-61.
3. W. F. Phillips. A New Continuum Traffic Model Obtained from Kinetic Theory. *IEEE Transactions on Automatic Control*, Vol. AC-23, No. 2, 1978, pp. 1032-1036.
4. J. Lin. *Modelling and Analysis of Traffic Flow in Complex Transportation Networks*. Ph.D. dissertation. Department of Civil and Mineral Engineering, University of Minnesota, Minneapolis, 1985.
5. G. Stephanopoulos and P. G. Michalopoulos. Modelling and Analysis of Traffic Queue Dynamics at Signalized Intersections. *Transportation Research*, Vol. 13A, 1979, pp. 295-307.
6. P. G. Michalopoulos, G. Stephanopoulos, and V. B. Pisharody. Modelling of Traffic Flow at Signalized Links. *Transportation Science*, Vol. 14, No. 1, 1980, pp. 9-41.
7. P. G. Michalopoulos and V. B. Pisharody. Platoon Dynamics on Signal Controlled Arterials. *Transportation Science*, Vol. 14, No. 4, 1980, pp. 365-396.
8. B. D. Greenshields. A Study of Traffic Capacity. *HRB Proc.*, Vol. 14, 1934, pp. 448-477.
9. *Traffic Network Analysis with NETSIM—A User Guide*. Implementation Package FHWA-IP-80-3. FHWA, U.S. Department of Transportation, 1980.
10. D. L. Gerlough and M. J. Huber. *Special Report 165—Traffic Flow Theory: A Monograph*. TRB, National Research Council, Washington, D.C., 1975.
11. P. G. Michalopoulos and L. Wung. *Treatment of Interrupted Traffic Flow in Complex Congested Arterials*. Department of Civil and Mineral Engineering, University of Minnesota, Minneapolis, 1986.
12. J. C. Tanner. A Theoretical Analysis of Delays at an Uncontrolled Intersection. *Biometrika* 49, 1962, pp. 163-170.
13. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
14. P. G. Michalopoulos, D. E. Beskos, and Y. Yamauchi. Multilane Traffic Flow Dynamics: Some Macroscopic Considerations. *Transportation Research*, 18B, No. 4/5, 1984, pp. 377-395.
15. P. G. Michalopoulos, J. O'Connor, and S. M. Novoa. Estimation of Left-Turn Saturation Flows. In *Transportation Research Record 667*, TRB, National Research Council, Washington, D.C., 1977, pp. 35-41.
16. P. G. Michalopoulos, D. E. Beskos, and J. Lin. Analysis of Interrupted Traffic Flow by Finite Difference Methods. *Transportation Research*, 18B, No. 4/5, 1984, pp. 409-421.
17. P. G. Michalopoulos and J. Lin. A Freeway Simulation Program for Microcomputers. *Proc., 1st National Conference on Microcomputers in Urban Transportation*, San Diego, Calif., 1985, pp. 330-341.

# Traffic Distribution Fitting: A Systematic Methodology

RUSSELL THOMPSON AND WILLIAM YOUNG

**Traffic distributions are an integral part of many traffic analyses. The correct determination of these distributions is therefore essential. This paper presents a methodology for the determination of the appropriate distribution. The methodology incorporates the data investigation, model selection, parameter estimation, and goodness-of-fit stages in distributions fitting. Many traffic analysts pay scant attention to the first two steps. The methodology is formulated within a microcomputer program, TRANSTAT, that has been developed to determine appropriate traffic distributions. TRANSTAT incorporates many recent developments in exploratory data analyses, expert systems, and outlier analyses. The paper describes the program and illustrates its application to the determination of parking duration curves.**

Traffic engineering data analysis often requires the investigation of statistical distributions (1). For instance, distributions are needed for traffic simulation, checking of distributional assumptions in statistical analysis, and the determination of outliers (non-typical observations, inconsistent with the other values in the sample). This paper emphasizes a methodological approach for determining appropriate distributions.

The probability distribution fitting process can be thought of as a set of interconnected steps. The steps in the process are:

- Data collection,
- Data investigation,
- Model selection,
- Parameter estimation,
- Goodness-of-fit determination, and
- Application.

It should be noted that curve fitting is an iterative process. It is a search of possible alternatives for an appropriate explanation of the data. The backward link between goodness-of-fit testing and data investigation should continue until a "representative" model has been found. The data investigation and model selection stages of this process are often given scant attention, with analysts moving straight from data collection to parameter estimation. This paper emphasizes the need to consider these steps fully.

The overlooking of the data investigation and model selection stage often results from the time needed to prepare graphs and plots of the data. The formalization of the methodology, therefore, owes much to recent developments in microcom-

puters and computer graphics. These developments have created the rapid interactive environment that allows the data investigation and model selection stages to become an important part of the distribution fitting methodology. The discussion of the methodology is therefore couched in terms of a microcomputer package (TRANSTAT). TRANSTAT has been developed to fit common univariate distributions to traffic data. The need for such a package was identified largely from the inadequacy of the existing commercial statistical packages (2). Most packages did not include any facilities to fit distributions to data. Some microcomputer statistical packages do allow probability densities to be fitted (*e.g.*, STATGRAPHICS, SPSS/PC+). However, the treatment is usually very limited. TRANSTAT contains unique modules relating to outlier analysis, model guidance and diagnostic plots. Also, critical regions as well as test statistics are presented for all tests. These features render TRANSTAT superior compared with conventional analysis packages. TRANSTAT provides a comprehensive and user-friendly package to aid the investigation and modeling of statistical distributions.

This paper describes TRANSTAT, emphasizing the need to take a good look at the data before attempting model fitting, then illustrates its application to the determination of the distribution of parking duration times.

## TRANSTAT OVERVIEW

TRANSTAT is written in the Microsoft's QuickBASIC computer-programming language and is designed to run on IBM PC-XT/AT microcomputers. Data input is via an ASCII file, and individual data values must be separated by at least one space. There is a data limitation of 2000 observations. The file input data facility allows data to be directly interfaced with automatic data collection devices enabling the transfer of data without manual intervention. This can speed up analysis considerably and allow observations to be measured with a high degree of accuracy. TRANSTAT is available from the Department of Civil Engineering, Monash University, Australia, for \$100 (Aust.).

A conscious effort has been made to make TRANSTAT as efficient as possible. Calculation speed has been given high priority. For instance, when sorting data, the Quicksort routine has been used (3). This is accepted as being the fastest general sorting procedure available for computers.

Color graphics have been used extensively throughout TRANSTAT to enhance the modeling process. The package allows either the Enhanced Graphics Adapter (EGA) or the standard Color Graphics Adapter (CGA) to be used. EGA

```

          TRANSTAT

FILENAME: SPEED.DAT

          MAIN MENU

1...  EXPLORATORY DATA ANALYSIS
2...  SUMMARY STATISTICS
3...  OUTLIER ANALYSIS
4...  MODEL GUIDANCE
5...  DISTRIBUTION FITTING
6...  EXIT PROGRAM

ENTER OPTION?

```

FIGURE 1 TRANSTAT's main menu.

graphics offer a higher resolution output with many more colors.

Interactive menus are used throughout the package in an attempt to make distribution modeling simple. The main menu (Figure 1), allows any one of five modules to be chosen. After selecting an option, another menu or prompt will be presented requesting information. The structure of the main menu is deliberately designed to encourage the distribution-fitting methodology introduced above, and hence good modeling practice, by the ordering of the options.

The following sections of the paper describe the options offered in the main menu. They discuss the application of exploratory data analysis, summary statistics, outlier analysis, model guidance, and distribution fitting (the Chi-squared test and the Kolmogorov-Smirnov test). The first two steps provide a strong indication of the type of distribution. The distribution fitting stage provides statistical measures of the degree of fit between the observed and the expected distributions.

### Exploratory Data Analysis

Exploratory Data Analysis (EDA) techniques (2) are becoming very popular. These techniques seek to guide the analyst into the appropriate analysis procedure through the visual examination of data. Numerous plots of data can be used to illuminate its structure and subtleties. Tukey (4) states that EDA is "numerical detective work" and notes that its strength lies in forcing "us to notice what we never expected to see."

EDA should precede statistical inference and stochastic model-building. Its role is to enhance the analyst's knowledge of the data before any inference is attempted. It is recommended that several plots of the data be produced before any models are fitted. The EDA module appears first on TRANSTAT'S main menu to encourage its early use in the modeling process.

Although there are many standard types of EDA plots, EDA itself is not a set of well defined procedures. It consists of any plot or data analysis technique which attempts to illus-

trate the structure or subtleties of the data. The plots included in TRANSTAT are commonly used for investigating one dimensional data. The box, quantile, and jitter plots are available in TRANSTAT and provide a valuable tool for model identification. To illustrate the variety of plots the histogram, jitter, and box plots will be discussed.

The frequency histogram (Figure 2) provides an effective method of displaying frequency data. However, since it summarizes the data, it can produce a deceiving picture and care should be taken in its interpretation. This distortion can be caused by the length and number of the class intervals. TRANSTAT offers the user the option of specifying the class size and the start of the first interval to alleviate this problem. It is good practice to vary the class size in order to gain a varied picture of the data's distribution. In practice, about twenty classes over the data's range usually provide a reasonable picture of the distribution.

Chambers et al. (5) introduced an alternative to the frequency histogram, a one-dimensional scatter plot. This plot is called a jitter plot. Dots are used to represent observations. The dots are placed in the appropriate position on the major axes. This often results in loss of data due to point overlap. The overlap is reduced by spreading the points randomly on the vertical axis. The latter plot can be used to complement a frequency histogram by providing an undistorted picture of the data's local density. The data's range, density and symmetry can be easily seen from this plot. An example of the combined frequency and jitter plots for different TRANSTAT distributions is presented in Figure 2.

Another useful EDA plot is the box plot. Initially introduced by Tukey (4), this plot summarizes the data by plotting the upper and lower quartiles as well as the median. These quartiles are represented by the top and bottom horizontal lines of the rectangle with the median being portrayed by the line within the rectangle. Generally, vertical lines extend from the box to the minimum and maximum values. However, if a very large or small value falls outside the upper quartile plus or minus 1.5 times the inter-quartile range, it is highlighted by being plotted individually. The vertical lines then extend to the next highest or lowest value falling inside this range. This provides a simple method of identifying possible outliers, singling them out for further examination. This plot conveniently summarizes many distributional properties, including symmetry, spread, and range. Figure 3 presents box plots for the distributions available in TRANSTAT. Distribution types can be rejected on the basis of symmetry, range, etc., narrowing the possible model types to be considered.

Box plots can also be used to study numerous "batches" of data. Multiple box plots can also be used to partition sets of data into homogeneous classes. Figure 4 provides an example of the application of this type of plot to the presentation of parking duration data over time. Here, the skewness of the data suggests that the exponential or Erlang may be appropriate.

### Summary Statistics

Summary statistics calculated from the sample are important in determining the type of distribution. Most distributions have a constrained range and other properties based on sample statistics.

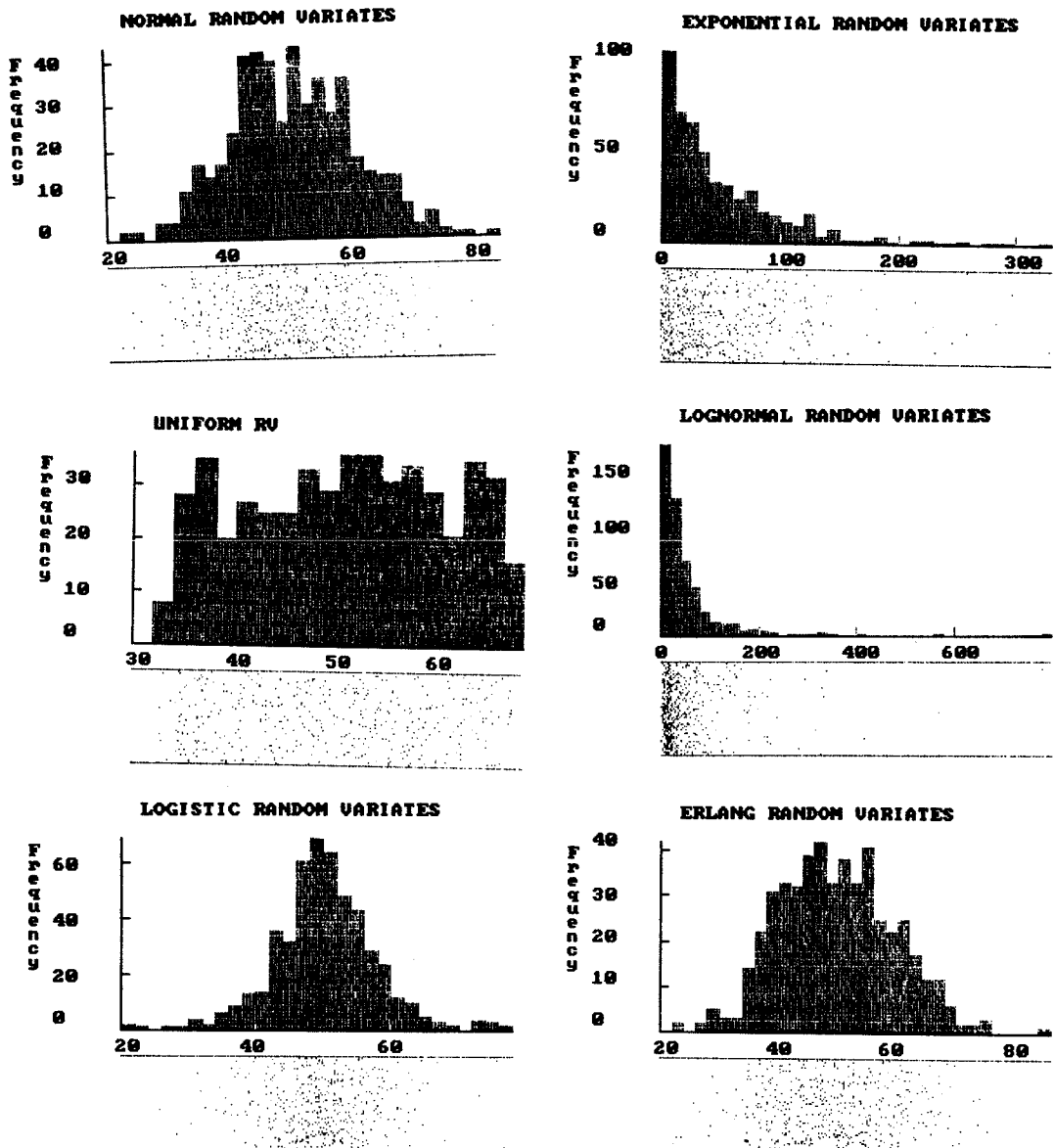


FIGURE 2 Histograms of various distributions.

Certain statistics measure the central tendency of data. The mean, median, and mode are included in TRANSTAT. The mean can often be distorted when outliers are present. It is called a non-robust estimator as a result. A possible remedy, the trimmed mean, for making the mean less sensitive to outliers is present in TRANSTAT. The trimmed mean is where a small proportion of data from each end of the sample is trimmed, and the remaining observations are used to calculate the average (6). The median is less sensitive to outliers and, for nonsymmetric distributions, provides a more robust measure of location. The mode presents a poor measure of location. Measures of dispersion indicate the spread or variability of the data and include the range, standard deviation, and coefficient of variation. The standard deviation provides a suitable statistic to judge the reliability of the mean as a measure of location. The standard deviation is a measure of absolute dispersion, and its units are the same as the data's. The coefficient of variation is a measure of relative dispersion. It

allows populations to be compared. The peakness of the distribution is called the kurtosis. The symmetry of a distribution is measured by the coefficient of skewness. A zero value indicates a completely symmetric distribution. A distribution can be skewed to the right (positive) or left (negative). Figure 5 illustrates the summary statistics output produced from TRANSTAT. The use of four decimal places enables the analyst to get an indication of the need for rounding; smaller numbers of decimal places can be used.

#### Outlier Analysis

An important aspect of the distribution fitting methodology is the determination of outliers. Outliers can be due to coding or input errors. Further, with the increasing prevalence of electronic equipment in data collection, protocol errors and instrument errors are becoming common. Outliers can also

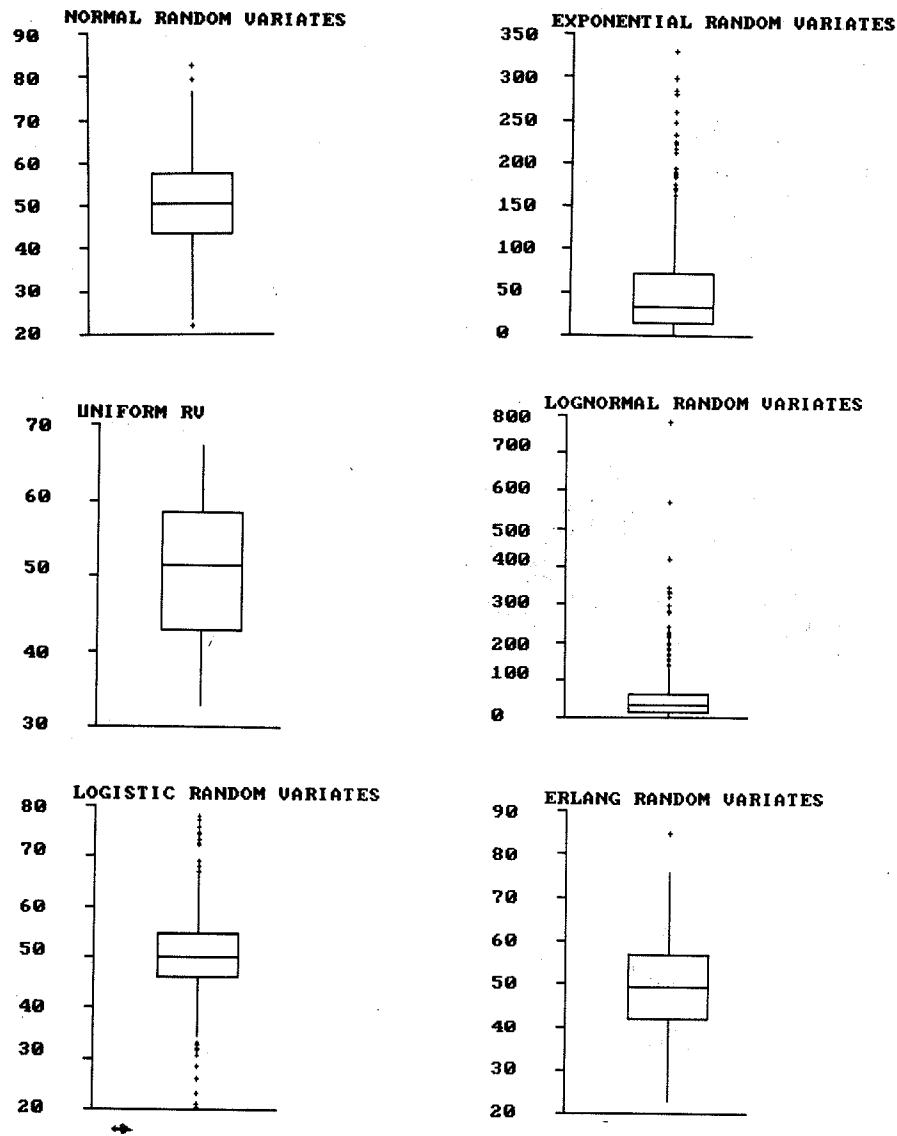


FIGURE 3 Box plots from various distributions.

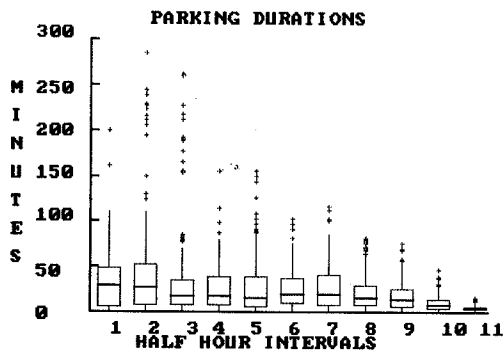


FIGURE 4 Multiple box plots of half-hourly parking duration data.

be due to measurement or copying errors, both of which can easily occur, if the data set is large. Since these values can significantly affect sample estimates, they should be identified and investigated for their validity.

The criteria used to detect potential outliers in the sample

data are those of Tukey (4) and Barnett and Lewis (7). The first method is based on the test used for highlighting observations outside the fences in the box plot. Any value exceeding the distance 1.5 times the interquartile range from the median is selected. This method is distribution-free, which is most favorable since at this stage in the modeling process no distributional assumptions can be inferred. It provides a convenient and simple rule of thumb for identifying possible suspect observations. However, since it assumes symmetry, it is only relevant for samples which are approximately symmetric.

Barnett and Lewis (7), in their comprehensive summary of outlier tests, give little attention or hope of any non-parametric methods being used to detect outliers. This is colorfully expressed by their statement, "To deliberately abandon the model, by seeking non-parametric (or distribution free) methods in some broad aim of robustness, smacks of throwing out the bath water before the baby has even been immersed." Therefore, it is not surprising that nearly all Barnett and Lewis's (7) "discordancy" tests are based on knowledge of the underlying distribution. TRANSTAT also uses a "discordancy" test to identify any possible outliers for skewed



TRANSTAT	
DURATIONS INTERVAL 2	
NO. OF OBSERVATIONS	= 90
MINIMUM OBSERVATION	= 1.0000
MAXIMUM OBSERVATION	= 285
SAMPLE MEAN	= 52.8778
SAMPLE MEDIAN	= 28.0000
STANDARD DEVIATION	= 71.1630
SKEWNESS COEFF.	= 1.7924
KURTOSIS COEFF.	= 4.9763
COEFF. OF VARIATION	= 1.3458
MODE	3.0000

FIGURE 5 Summary statistics produced by TRANSTAT.

distributions. It is based on the assumption that the sample comes from an exponential distribution. This can be justified by the fact that many known distributions have the exponential distribution as their limit, and many natural phenomena are distributed this way. Furthermore, while being careful to assume not all distributions are exponential, if data is significantly skewed this test is useful and performs quite well.

The system provides information relating to the observations which may be outliers based on the above criteria. It gives the user the option of including these values in subsequent analysis. This relieves TRANSTAT of any decision regarding the inclusion of these values in the analysis but indicates that, if these values are to remain, their influence will be significant (Figure 6). This information should encourage the user to investigate these observations for their validity. Individual values can be deleted from the analysis using this option if an outlier is confirmed.

### Model Guidance

A small "expert system" has been included in TRANSTAT to help users identify potential representative models. This can eliminate many unlikely models from detailed examination. An "expert system" is described by Marksjo (8) as being "a piece of computer software giving the illusion of being a human expert within a restricted field of competence." There has been considerable interest in expert software applications in statistics, with emphasis being given to systems where consultant (expert) and client (user) interactions are imitated (5,9). This new type of software can suggest actions which will achieve the goals of the analysis and provide explanations of output and of recommendations.

TRANSTAT provides a list of suggested distributions which should receive closer attention. The selection criteria are based on inferences obtained from distributional properties (10). These potentially should provide a reasonable fit, but formal and informal tests should be used to confirm or reject this preliminary advice. Numerous tests are performed based on the data's symmetry, kurtosis, and range to provide this initial selection of distributions. Confidence limits are generated for the skewness and kurtosis coefficients as well as for the mean and standard deviations. These are then compared with the properties of the theoretical models and recommendations given as to their similarities. The user can accept or reject the system's advice concerning suggestions. The system provides a list of most suitable distributions and ranks these by their apparent suitability. Reasons, in the form of explanations, are also given for the rejected and suggested models. This will help educate analysts. The advice provided must be interpreted as preliminary guidance, and the suggested models should be subjected to the appropriate statistical tests before any conclusions can be made. A sample of the model guidance produced by TRANSTAT is provided in Figure 7.

TRANSTAT		
PRELIMINARY SCANNING OF THE DATA HAS IDENTIFIED 1 UNREPRESENTATIVE OBSERVATION(S) WHICH MAY POSSIBLY BE OUTLIERS		
OBSERVATION(S) :		
58		
THE EFFECT OF THESE OBSERVATION(S) ON SAMPLE STATISTICS CAN BE SEEN BY THE FOLLOWING COMPARISONS:		
STATISTICS	WITH	WITHOUT UNREPRESENTATIVE VALUES
MEAN	30.1548	30.0840
STDVN	7.1155	6.9840
MIN	14.0000	14.0000
MAX	58.0000	50.0000
DO YOU WISH TO DELETE THE ABOVE OBSERVATIONS (Y/N)?		

FIGURE 6 Outlier detection from TRANSTAT.

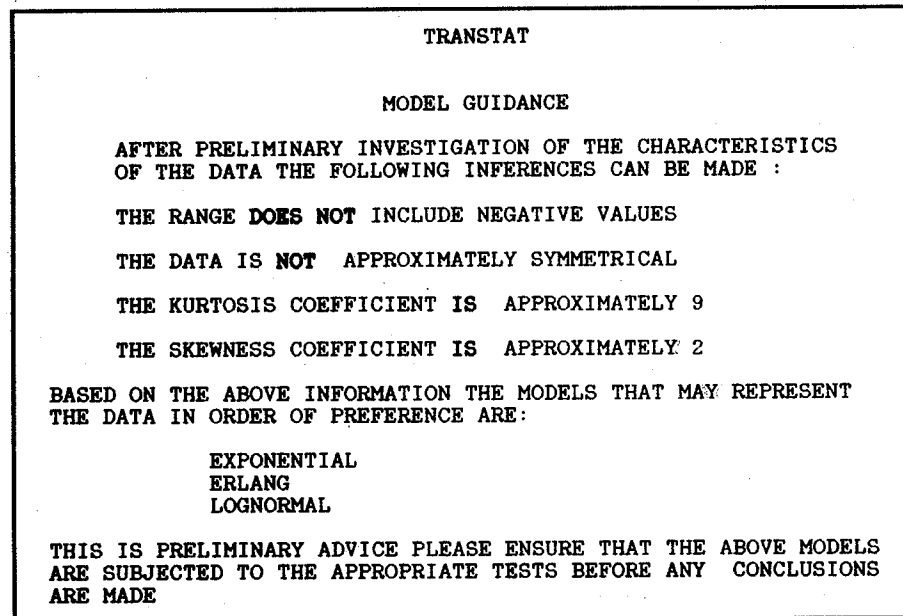


FIGURE 7 Model guidance output from TRANSTAT.

### Chi-Square Test

The previous discussion illustrated the general application of a number of techniques to the investigation of the data and the determination of the distributions likely to fit the data. These are important steps in the distribution-fitting methodology, and it is only after this investigation that data fitting should be attempted.

The statistical fit of the data to the theoretical distribution is tested in two ways. The first is a chi-square test. To test an hypothesis that enumerative data falls into  $k$  classes with probabilities  $p_1, p_2, \dots, p_k$  the chi-square test compares the actual counts  $o_1, o_2, \dots, o_k$  with the number expected in each class of  $k$  classes,  $np_1, np_2, \dots, np_k$ .

The comparison is made using the test statistic,

$$\chi^2 = \sum (O_i - E_i)^2 / E_i \quad \text{all } i$$

where

$O_i$  = observed frequency in class  $i$

$E_i$  = model expected frequency in class  $i$ .

The null hypothesis is that the two distributions are the same; and if this is true, the test statistic has a chi-square distribution for large samples ( $n > 40$ ).

Large value of  $\chi^2$  indicates that it is unlikely the hypothesized probabilities  $p_1, p_2, \dots, p_k$ , are correct and the hypothesized distribution is incorrect. The number of degrees of freedom is the number of classes,  $k$  less 1, for each linear restriction placed on the cell counts (11).

This test was originally designed for categorical data but can be applied to continuous data models. It involves subjectively choosing class lengths and the beginning of the first cell. Different class sizes can be used, which can affect the significance of the test (12). The subjective choice of class intervals may be a limitation with this approach. An additional problem with this test is that low cell expectancies can seriously affect the test statistic. Even with these serious prob-

lems, the chi-square test remains the most popular test used in distribution fitting. Its popularity seems to be due to its ease in calculating the test statistic.

The size of the classes can be specified in TRANSTAT. There seems to be little indication of what is an optimal class size; in practice, many should be tried. Cochran (13) reports that expected frequencies should be equal. This produces unequal class lengths. Mann and Wald (14) suggest the following optimum number of classes ( $k$ ), based on a sample size ( $n$ ):

$n$	200	400	600	800	1000	1500	2000
$k$	31	41	48	54	59	70	78

The chi-square statistic is usually only calculated for classes where the expected frequencies are greater than 5. However, there seems to be little agreement among statisticians as to the minimum expected number for each class. Cochran (13) suggests that, "since the discrepancy between observed and a postulated distribution is often most apparent at the tails, the sensitivity of the chi-square test is likely to be decreased by an overdose of pooling at the tails." In this light, he considers that the inflexible use of a minimum expectation of 5 may be harmful. A sample of the output produced from this module of TRANSTAT is given in Figure 8.

### Kolmogorov-Smirnov Test

The second test of goodness-of-fit is the Kolmogorov-Smirnov test. It is for continuous data and can be used for samples of any size (15). The test statistic is based on the measurement of the maximum absolute vertical difference between the two cumulative distributions (Figure 9).

The test compares  $F(x)$ , the hypothesized (cumulative) distribution function, with  $S(x)$ , the empirical distribution function. If the data follow the hypothesized distribution, then  $|F(x) - S(x)|$  should be small.

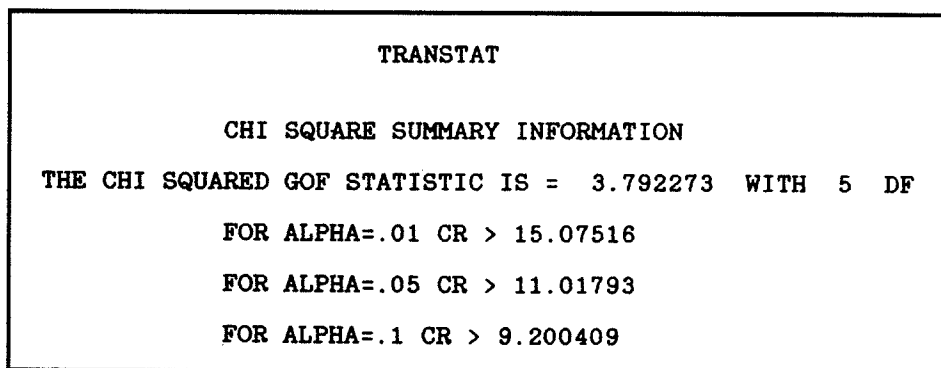


FIGURE 8 Chi-square output produced from TRANSTAT.

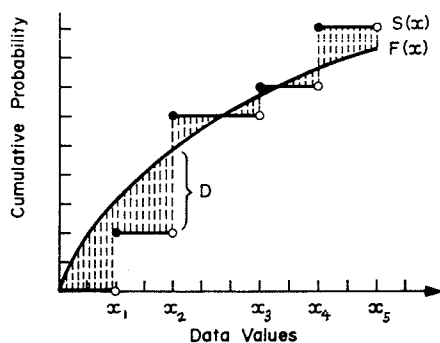


FIGURE 9 Calculation of the Kolmogorov-Smirnov test statistic.

The Kolmogorov-Smirnov test statistic for samples of size  $n$  is  $D = \text{Supremum } |F(x) - S(x)|$  over all  $x$ .

The advantages of this test are that it uses all the data points, it involves no subjective grouping, and it can be used on small data sets. It therefore eliminates many of the weaknesses of the chi-square test. If the parameters of the hypothesized distribution are previously known, the critical regions are distribution-free and, for  $n > 30$  are equal to  $1.36/n$  at the 5% significance level.

However, if the parameters of the hypothesized distribution are estimated from the sample data, the common tabulated values are not valid; and if used, they result in a conservative test. Recognizing this major weakness, Lilliefors (16, 17) developed new tables of critical values for when the parameters are estimated from the sample. Adjusted critical regions were derived using Monte-Carlo techniques for the Exponential and Normal distributions. The Monte-Carlo critical values presented are approximately two-thirds that of the common tables. Monte-Carlo techniques have also been used to generate adjusted critical values for the Log-Normal distribution (2). Over 2000 distributions were generated for odd sample sizes up to 30, and linear interpolation was used to produce the even sample numbers. The results seem consistent with those of Lilliefors (16).

Durbin (18) showed that critical regions can only be generated exactly if the estimated parameters of the distribution are measures of location and spread. This indicates the limitation of Lilliefors's (16) method to the Normal, Log Normal, and Exponential distributions. However, Durbin described a half sample device where the parameters are estimated using

a random sample of half the data. This procedure is asymptotically equivalent to the original test, and hence the common critical regions can be used. The only drawback with this method is that large samples (100 or more data points) are required. TRANSTAT allows the user to specify whether or not to use the half-sampling technique. If the half sample technique is chosen, critical regions are displayed for the 1%, 5%, and 10% levels of significance for all models.

Another use of  $D$  is in comparative studies. Since  $D$  is a measure of the model's goodness-of-fit, it can be used as a comparative statistic for evaluating various models. For example, a result of model fit may produce a  $D$  equal to 0.11, which can be interpreted as meaning that the two cumulative distributions differ by at most 11% and should be preferred to a model with a  $D$  value of 0.16.

The computation of the Kolmogorov-Smirnov test statistic appears to be quite lengthy and complex, but close examination indicates both of these problems can be reduced. D'Agostino and Nothier (19) state that this maximum distance does not necessarily occur at an observation point, and therefore only using these points in the calculation of  $D$  generally results in a conservative test. However, using mathematical analysis it can be shown that the maximum difference will occur either at an observation point of the empirical distribution or a very small distance before one of these points (10). Figure 10 presents some possible comparisons between the theoretical cumulative distribution ( $F(x)$ ) and the observed one ( $S(x)$ ). It can be seen that the maximum  $D$  value will be found at, or just before, an observation. Using this result, the computational effort is reduced to arrive at maximum: twice the number of different observation points. This is significantly more efficient than comparing these two functions at numerous small increments. The accuracy of the test is dependent upon how close to the data points the  $D$  value chosen is. This in practice should be in the order a one tenth of the measured significance.

Numerous plots can aid the interpretation of  $D$ . First, the plot of both the model and the empirical data illustrates the differences between them (Figure 11a). A second plot shows the difference between the theoretical model and the observed data points over the data's range (Figure 11b). More specifically, Figure 11b presents the  $D$  value in terms of the independent variable and provides the 5% confidence intervals. The plot provides an immediate view of the quality of the fit on the distribution to the data.

**PARKING DURATION ANALYSIS USING TRANSTAT**

To illustrate the distribution-fitting methodology, TRANSTAT is applied to the investigation of parking durations. A parking study was performed on a Saturday morning in July 1986 at the Mountain Gate Shopping Center lower carpark, Ferntree Gully Road, Ferntree Gully, Australia. An input/output technique was used to collect the data. Observers recorded the time in minutes and the registration number of vehicles entering or leaving the facility. This carpark has three entrances and exits, a good number for such a survey. The carpark has approximately 300 stalls with a variety of shops adjacent to it, including a supermarket, newsagents, and a restaurant. The data collected was summarized into half-hourly intervals of durations by arrival times. This enabled the temporal variation in parking duration curves to be investigated. The first period was for the half-hour from 8:00 a.m. to 8:30 a.m., with the last period being from 1:00 p.m. to 1:30 p.m. Short-term parkers mainly consisted of short shopping trips (e.g., out-of-hours banking at the Commonwealth Handy Bank Automatic Teller and staff being dropped off). High durations were mainly from the staff of the shops and from the supermarket patrons.

After the collection of the data, the first step in the distribution fitting methodology is to begin to build knowledge about the data. This can start by looking at the summary statistics. The summary statistics (Table 1) of the duration data by half-hour intervals provides valuable insight into the data's structure. The durations are presented visually in the multiple box plot in Figure 4. It can be seen that there is a general decrease in the average duration as the study progresses, and that there are many large durations in the first few periods. The large durations were vehicles of the staff of the businesses in the shopping center. They were therefore deleted from analysis as they were considered to be outliers. The criterion used for identifying long-term parkers was durations longer than 180 minutes. The mean duration time of vehicles, corrected for outliers, decreased steadily by the interval of arrival (Figure 12). A simple linear regression equation was fitted to this data with the following relationship established:

$$\text{Mean Duration} = 46.6601 - 3.4481 * (\text{Interval no.})$$

The fit was reasonable with an *R*-squared value of 78.44%. This result is very useful, as the model type that is chosen to represent durations by intervals has the mean as a parameter. The estimation of this parameter can be made with associated confidence intervals using this relationship.

The standard deviation of the durations also decreased in proportion to the mean with a constant linear relationship evident. A linear regression model was fitted providing a very good relationship (*R* squared of 93.74%). The following relationship was established using Least Squares Regression:

$$\text{Mean} = -9.8457 + 1.4784 * (\text{Standard deviation})$$

Data for all the intervals were positively skewed ranging from 1.2741 to 2.5641. These high values of skewness indicate that the normal, logistic, and rectangular distributions would not be suitable for modeling this data. The kurtosis coefficients were between 3.9 and 9.00, with those around 9 indicating the suit-

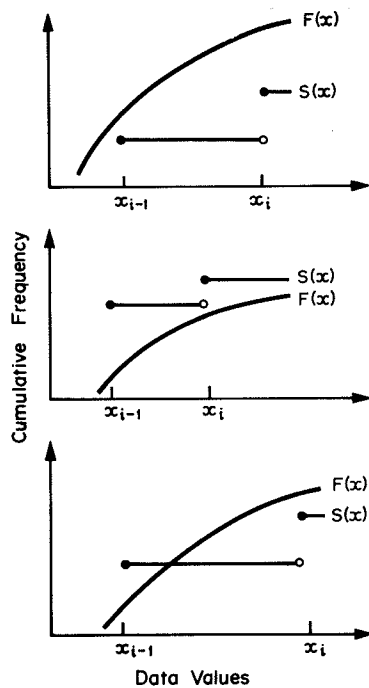
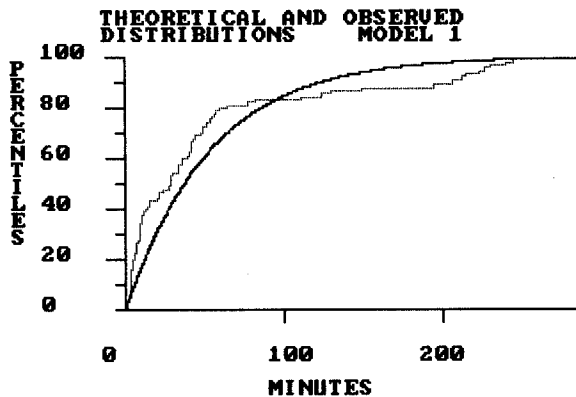
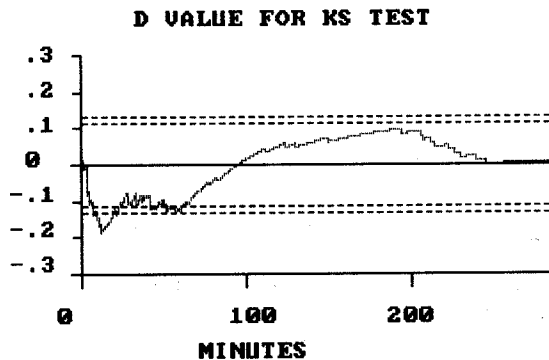


FIGURE 10 Calculation of *D*.



(a)



(b)

FIGURE 11 Kolmogorov-Smirnov analysis plots produced from TRANSTAT: (a) plot of empirical and theoretical cumulative distributions and (b) difference between observed and theoretical distributions.

TABLE 1 SUMMARY STATISTICS FOR PARKING DURATION DATA BY INTERVAL

INTERVAL NO.	N	MAX	MIN	MEAN	MEDIAN	STANDARD DEV.	SKEW. COEFF.	KURTOS. COEFF.	COEFF VAR.
1	48	200	1	35.19	30	40.00	2.21	8.82	1.14
1/A	47	160	1	31.68	29	32.13	1.73	6.98	1.01
2	90	285	1	52.88	28	71.16	1.79	4.98	1.35
2/A	79	149	1	28.92	20	30.71	1.84	6.74	1.06
3	153	262	1	37.05	17	53.94	2.56	9.00	1.46
3/A	143	165	1	24.90	16	27.92	2.67	11.94	1.12
4	160	155	1	25.14	17	24.01	1.94	8.58	0.96
5	181	155	1	26.35	15	30.14	2.21	8.56	1.14
6	219	102	1	25.31	19	21.52	1.27	4.31	0.85
7	193	116	1	27.35	20	24.17	1.28	4.41	0.88
8	171	81	1	21.98	16	19.01	1.29	3.94	0.87
9	104	75	1	18.08	14	16.42	1.44	4.76	0.91
10	63	47	1	11.48	8	10.28	1.47	4.74	0.90
11	38	15	1	4.90	4	3.40	1.54	4.96	0.70
ALL	1420	285	1	26.96	16	34.19	3.46	18.98	1.27
ALL/A	1398	165	1	23.96	15	24.42	2.10	9.19	1.02

NOTE: /A = data with extreme low values removed.

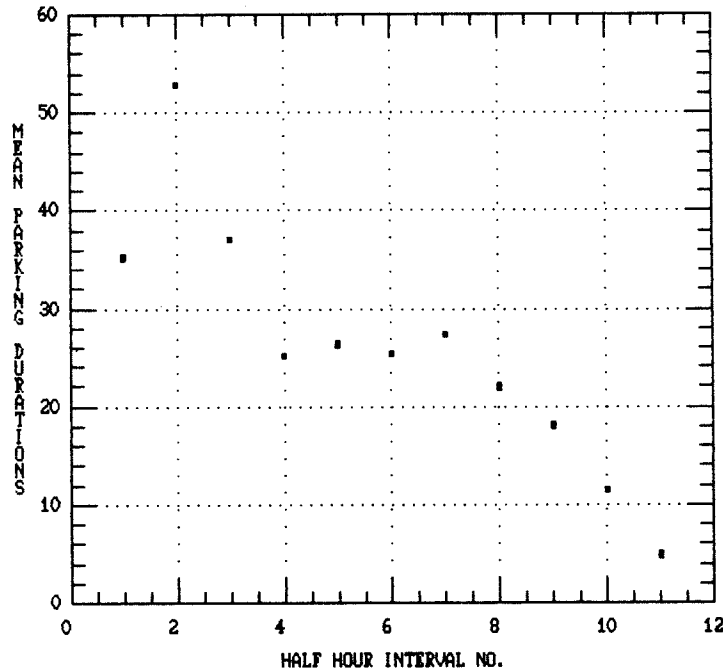


FIGURE 12 Plot of mean duration by arrival interval.

ability of the exponential distribution. A histogram and jitter plot of the data for half-hour interval number 5 is presented in Figure 13. The data for the whole morning can be conveniently summarized in a multiple box plot (Figure 4). The positive skewness of the data by intervals is evident from the small distance from the median and the minimum value compared to the maximum value. The steady decay of duration times by interval is obvious from the reducing medians and maximum values. The high number of outliers, particularly in the early intervals, highlights the long-term (staff) parkers. The general decrease in frequency with increasing duration is consistent with an exponential or Erlang distribution.

From the exploratory data investigation, it was considered that only the Exponential, Erlang, and Log Normal models could possibly represent the data to an acceptable level.

Observations over 180 minutes were deleted from the analysis, as their properties were of little interest. The Kolmogorov-Smirnov test was used for analyzing the goodness-of-fit of the distributions and the half-sample parameters were used when the sample size was over 150. This enabled critical regions to be established for these data sets. For the intervals where there was a small number of observations, the Monte-Carlo generated values were used.

The results of each of the models performance are presented in Thompson (10). Overall, only one data set (half-hour interval no. 11) failed to be satisfactorily represented by any of the models. To summarize the model fit, a ranking procedure was adopted. A score of 3 points was given for a model fit with no significant difference between the model and the data at the 10% level. A score of 2 was assigned to

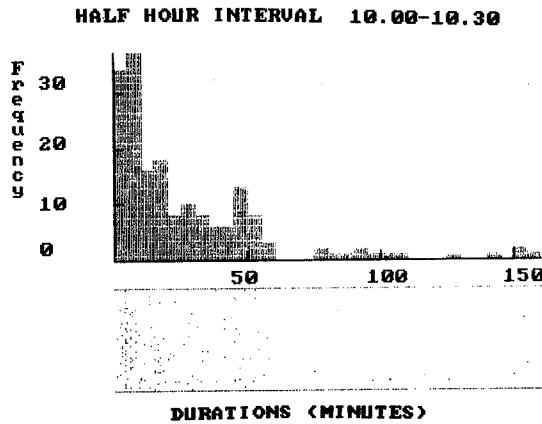


FIGURE 13 Histogram of parking durations half-hour interval no. 5.

a fit where a significant difference was detected at the 10% level. A score of 1 was assigned to a fit where a significant difference was detected at the 5% level. A zero score was assigned if a significant difference was detected at the 1% level. Using these rankings, the results are summarized in Table 2.

The best overall model was the Exponential, with the Log-normal distribution also performing well. The Erlang model rarely fitted the data well. Figure 14 illustrates the fit of both the Exponential and the Log Normal models to the duration data observed in half-hour interval number 8. An alternative ranking procedure based on the smallest *D* value for each model produced similar results. They are presented in Table 3.

The closeness in performance of the Exponential and Log Normal models is highlighted in Tables 2 and 3. Overall, the

Exponential distribution appears to accurately represent the parking durations of shoppers by half-hourly intervals. The associated parameters are a function of the mean of the data. This provides a convenient method for the generation of different distributions with varying means, since the mean/half-hourly interval relationship allows for the temporal variation to be included.

The more common method of looking at the distribution of parking durations is to group the data for the entire study period. The distribution of parking durations for all the data is therefore presented in the box plot in Figure 15. It illustrates the high degree of skewness and large number of outliers that are present.

The summary statistics of the durations for all the data (Table 1) exhibit many of the same characteristics as the individual half-hour interval data: high skewness and kurtosis coefficients. This general trend is further highlighted in the frequency and jitter plots presented in Figure 16. A very high coefficient of kurtosis of 18.7883 was calculated.

The distribution of the parking duration for the entire study period was also investigated. It showed no distribution could be accepted at the 10% significance level. However, after excluding the outliers from the data, the Log Normal model just failed to be accepted at the 10% level, with a *D* value of 0.0477. Poorer fits were found when every distribution was applied to the data sets with the long-term parkers included.

CONCLUSIONS

This paper has outlined a methodology for traffic distribution fitting. The methodology incorporates the steps of data investigation, model selection, parameter estimation, and goodness-of-fit testing. The methodology has been formulated in

TABLE 2 RANKING OF SIGNIFICANCE OF ERLANG, EXPONENTIAL, AND LOG NORMAL DISTRIBUTIONS TO DATA

Model	Half-Hour Interval Number											Rank Sum
	1	2	3	4	5	6	7	8	9	10	11	
Exponential	2	1	3	3	2	0	3	3	3	3	0	23
Erlang	3	0	3	3	0	0	0	0	0	0	0	9
Log normal	0	0	3	2	3	3	2	3	1	3	0	20

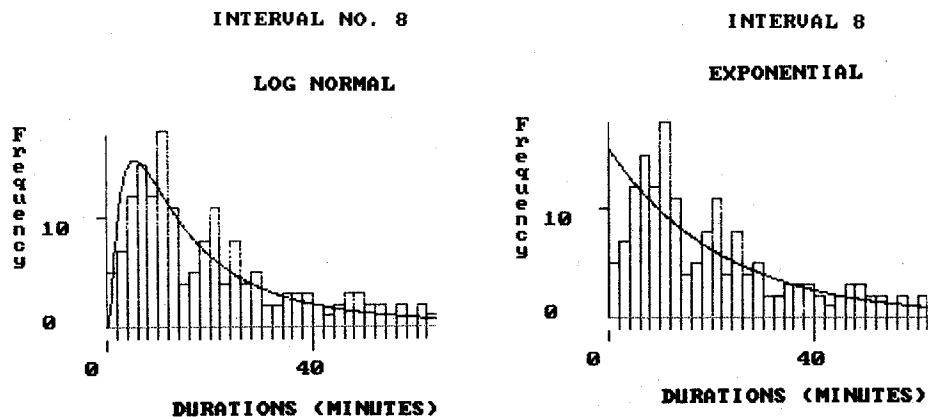


FIGURE 14 Exponential and Log Normal models of interval 8.

TABLE 3 RANKINGS OF DISTRIBUTION FITS BASED ON D

Model	Half-Hour Interval Number											Rank Sum
	1	2	3	4	5	6	7	8	9	10	11	
Exponential	2	2	3	2	2	3	3	2	3	2	1	25
Erlang	3	1	2	3	1	1	1	1	1	1	2	17
Log normal	1	3	1	1	3	2	2	3	2	3	3	24

NOTE: 3 = smallest D, 2 = 2nd smallest D, 1 = largest D.

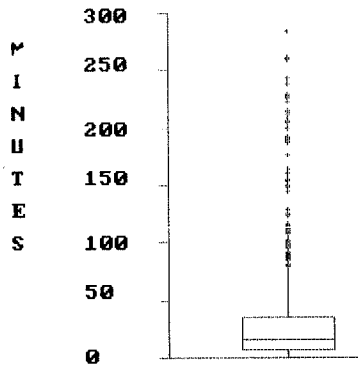


FIGURE 15 Box plot of durations for all intervals.

preparing input to simulation models, and performing probability and analysis.

The distribution fitting methodology was applied to the investigation of the distribution of parking times. The distribution of parking times was shown to be an exponential distribution. However, more importantly, the study demonstrated that the distribution of parking times varied throughout the day. The mean parking time was found to decrease by approximately 3.5 minutes/half hour, from a maximum of 46 minutes in the first half-hour to a minimum of 9 minutes. Other applications of TRANSTAT look at the speed of vehicles in parking lots and headway distributions (10).

REFERENCES

1. M. A. P. Taylor and W. Young. *Traffic Analysis: New Technology and New Solutions*. Hargreen Publishing Co., Melbourne, Australia, 1987.
2. R. G. Thompson and W. Young. New Developments in Distribution Fitting. Presented at the International Symposium on Simulation and Mathematical Modeling, Melbourne, Australia, 1987.
3. S. Dvorats and A. Musett. *Basic in Action*. Butterworths Publishing Co., Sydney, Australia, 1984.
4. J. W. Tukey. *Exploratory Data Analysis*. Addison Wesley Publishing Co., Boston, Mass., 1977.
5. J. M. Chambers, W. S. Cleveland, B. Kliener, and P. A. Tukey. *Graphical Methods for Data Analysis*. Duxbury Press, Boston, Mass., 1983.
6. D. F. Andrews, P. J. Ickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. *Robust Estimates of Location*. Princeton University Press, Princeton, N.J., 1972.
7. V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley & Sons, New York, N.Y., 1978.
8. B. S. Marksjo. Expert Systems for Local Government Planners. In *Microcomputers for Local Government Planning and Management* (P. W. Newton and M. A. P. Taylor, eds.), Hargreen Publishing Co., Melbourne, Australia, 1985.
9. G. J. Hahn. More Intelligent Statistical SOFTWARE and Statistical Expert Systems: Future Directions. *The American Statistician*, 39(1), pp. 1-16.
10. R. Thompson. *TRANSTAT: A Statistical Package for Traffic Planners*. Master of Engineering Science thesis. Monash University, Australia, 1987.
11. Mendenhall and Scheaffer. *Mathematical Statistics with Applications*. Duxbury Press, North Scituate, Mass., 1973.
12. E. M. Gumbel. On the Reliability of the Classical Chi-Square Test. *Annals of Mathematical Statistics*, 14, 1943, pp. 253-263.
13. W. G. Cochran. The Chi Square Test of Goodness of Fit. *Annals of Mathematical Statistics*, 23, 1952, pp. 315-345.
14. H. B. Mann and A. Wald. On the Choice of the Number of Class Intervals in the Application of the Chi Square Test. *Annals of Mathematical Statistics*, 13, 1942, pp. 306-317.
15. J. R. Benjamin and C. A. Cornell. *Probability, Statistics and Decision for Civil Engineers*. McGraw Hill, New York, N.Y., 1970.
16. H. W. Lilliefors. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *American Statistical Association Journal*, 1967, pp. 399-402.

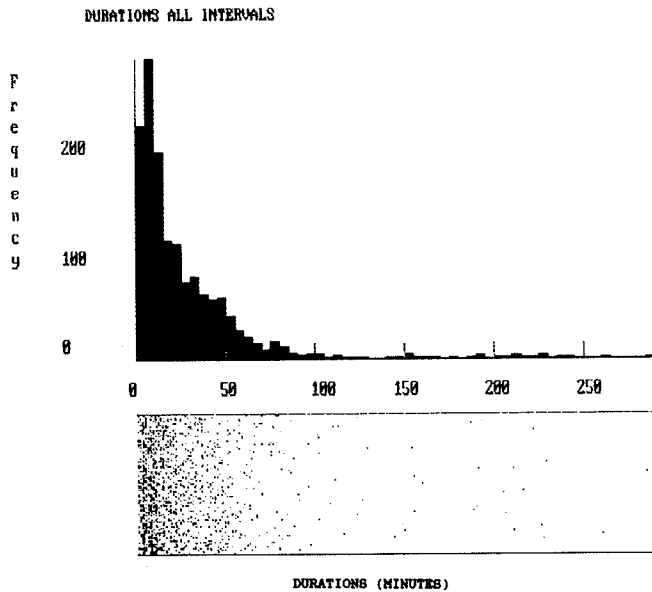


FIGURE 16 Histogram of all durations.

terms of a statistical package (TRANSTAT). The package has incorporated recent developments in exploratory data analysis, outlier determination, and expert systems. It is a comprehensive probability distribution-fitting package which enables probability density functions to be easily fitted to sample data. Full statistical details are provided enabling subsequent probability analysis. The package offers the traffic analyst an interactive, fast, and accurate package to aid in distributional analysis. TRANSTAT has been used for checking distributional assumptions in classical statistical analysis,

17. H. W. Lilliefors. On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *American Statistical Journal*, 1969, pp. 387-389.
18. J. Durbin. *Distribution Theory for Tests Based on the Sample Distribution Function*. Society of Industrial and Applied Mathematics, Philadelphia, Penn., 1973.
19. R. B. D'Agostino and G. E. Noethier. On the Evaluation of the Kolmogorov Statistic. *The American Statistician*, 27(2), 1973, pp. 81-82.

---

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



# CARSIM: Car-Following Model for Simulation of Traffic in Normal and Stop-and-Go Conditions

R. F. BENEKOHAL AND JOSEPH TREITERER

A CAR-following SIMulation model, CARSIM, with more realistic features to simulate not only normal traffic flow but also stop-and-go conditions on freeways, has been developed. The features of CARSIM are: (1) marginally safe spacings are provided for all vehicles, (2) start-up delays of vehicles are taken into account, (3) reaction times of drivers are randomly generated, (4) shorter reaction times are assigned at higher densities, and (5) dual behavior of traffic in congested and non-congested conditions is taken into consideration in developing the car-following logic of this model. The validation of CARSIM has been performed at microscopic and macroscopic levels. At the microscopic level, the speed change patterns and trajectories from CARSIM were compared with those from field data; whereas at the macroscopic level, average speed, density, and volume computed in CARSIM were compared with the values from real world traffic conditions. The regression analysis of simulation results versus field data yielded  $R^2$  values of 0.98 and higher, indicating that the results from CARSIM are very close to the values obtained from field data. One example of the application of CARSIM to study traffic-wave propagation is presented.

Since the beginning of traffic simulation in the mid-1950s, more realistic features have been added to newly developed models. In the more recent models, the emphasis has shifted from machine-processing efficiency to human efficiency in using the models (2, 3). A comprehensive review of car-following studies is provided (1), as are a comparison of car following models (4, 5, 9) and a review of traffic simulation models (6, 7, 8). The feasibility of developing an integrated simulation system to improve the human and computational efficiency has been investigated by Federal Highway Administration since 1975. As a result, an integrated traffic simulation model, called TRAF, has been developed and can be used for the evaluation and development of traffic control and traffic management policies (2). The creation of TRAF was not involved with new model development, but with the enhancement of the best existing traffic simulation models.

The most complete and updated microscopic freeway simulation model is INTRAS (9), which was included in phase I of TRAF. INTRAS is a highly complex stochastic model and is capable of simulating a variety of traffic/flow conditions. However, the INTRAS model is not capable of realistically

simulating the behavior of traffic in stop-and-go situations on freeways. This is mainly due to the following assumptions made in developing its car-following algorithm:

- (1) INTRAS uses a constant value of 0.3 seconds to represent the reaction time of drivers (a lag).
- (2) INTRAS does not take into account the start-up delay of the stopped vehicles.
- (3) In INTRAS, the dual behavior of traffic in congested and non-congested conditions has not been taken into consideration.

Therefore, a new car-following model or a substantial improvement in the car-following algorithm of INTRAS is needed, if the model to be used for simulation of stop-and-go conditions on freeways.

The CAR-following SIMulation model, CARSIM, was developed to take into consideration the aforementioned shortcomings of the INTRAS car-following algorithm and to offer additional realistic features and capabilities for simulation of stop-and-go conditions on freeways. In its present form, CARSIM simulates only the car-following behavior of freeway traffic. CARSIM has been validated at microscopic and macroscopic levels using field data. At the microscopic level, the speed change patterns and the trajectories of vehicles obtained from CARSIM were compared with those from field data. At the macroscopic level, however, the average speed, density, and volume computed in CARSIM were compared with the values calculated from the field data. The development and validation of CARSIM as well as its features and capabilities are briefly discussed in the following sections.

## FEATURES OF CARSIM

The following features were included in the development of car-following logic of CARSIM:

1. The vehicles' acceleration and deceleration rates were kept within the reasonable values observed in actual traffic conditions, and yet marginally safe spacings were provided for all vehicles.
2. The delay in response of the following driver to the lead car's deceleration was taken into consideration. The delay is equal to the reaction time of the driver.
3. The start-up delay of a stopped vehicle was taken into consideration. The start-up delay is, on the average, less than 2 seconds.

R. F. Benekohal, Department of Civil Engineering, University of Illinois at Urbana-Champaign, 208 North Romine Street, Urbana, Ill. 61801. J. Treiterer, Civil Engineering Department, Ohio State University, Columbus, Ohio 43210.

4. The dual behavior of traffic in congested and non-congested conditions was taken into consideration. For non-congested flow condition (density less than 60 vpm), the following and lead vehicle both have the same maximum deceleration rate. However, a maximum deceleration rate of 13 ft/sec/sec is used for the following vehicle when density exceeds 60 vpm, while the maximum deceleration rate for the lead car is 16 ft/sec/sec. The use of different maximum deceleration rates for congested flow condition is merely for computational purposes.

5. CARSIM uses varying reaction times for an individual driver and different reaction times for different drivers. The reaction time of a driver in congested flow conditions is less than his reaction time in free flow conditions. A driver's reaction time is randomly selected from a category of twelve different reaction times.

6. CARSIM has the capability to simulate stop-and-go conditions, a feature that is extremely important in studying the effects of kinematic disturbances on traffic.

## DESCRIPTION OF CARSIM

The car-following and the acceleration algorithms are the two most critical routines in CARSIM. A detailed discussion of these algorithms will be given in the following sections. A brief discussion of the inter-arrival time of vehicles, vehicle generation, reaction time of drivers, and speed distribution will be also presented. For programming of CARSIM, SIMSCRIPT II.5, a simulation language, is used (10).

### Car-Following Algorithm

The car-following algorithm is basically a vehicle-advancing mechanism that facilitates the movement of vehicles from one point to another along the road. In conjunction with the acceleration routine, it determines the proper acceleration or deceleration rate a vehicle should maintain in a given time interval. Once the acceleration or deceleration rate is determined, it is used to compute the speed and the location of the vehicle at the end of that time interval. In CARSIM, the following vehicle will be advanced to a position that provides it enough space headway to decelerate to a safe speed or a complete stop when the lead car reduces its speed.

The very important parameter in advancing a vehicle throughout the system is finding the proper acceleration or deceleration rate ( $AXL$ ).  $AXL$  is determined in the acceleration routine. Once this rate is known, the speed and the location of the vehicle, at the end of updating time interval, is calculated using the following equations:

$$V_F = V_F + (AXL)(DT) \quad (1)$$

$$X_F = X_F + V_F(DT) + 0.5(AXL)(DT)^2 \quad (2)$$

where:

$AXL$  = proper acceleration or deceleration rate for that time interval;

$DT$  = scanning time interval (1 second);

$V_F$  = velocity of the following vehicle at the beginning or end of the scanning time interval;

$X_F$  = position of the following vehicle at the beginning or the end of the scanning time interval.

The car-following algorithm checks whether or not the vehicle is the first unit in the system (leader). If the vehicle is the first unit in the system and is traveling at its desired speed or at the speed limit, an acceleration rate of zero is used to update the speed and location of this vehicle. However, for a vehicle traveling slower than the desired speed or the speed limit, the rate ( $AXL$ ) is computed from the acceleration routine. When the first vehicle in the system is traveling faster than the desired speed or speed limit, either an acceleration of zero or a comfortable deceleration rate is used to update the speed and location of this vehicle.

For the following vehicles in the system, the acceleration routine is called to determine the acceleration or deceleration rate. The following section describes how the acceleration or deceleration rates are determined in the acceleration routine.

### Acceleration Algorithm

The acceleration routine determines the acceleration or deceleration rate a following vehicle should have while satisfying all safety and operational constraints. Several acceleration or deceleration rates are computed for different situations, and the most suitable one is selected for each vehicle in every time interval.

The acceleration or deceleration rates are computed for every vehicle in 1-second time intervals for the following situations:

1. The following vehicle is moving but has not reached its speed limit or desired speed:  $A1$ .
2. The following vehicle has reached its speed limit or desired speed:  $A2$ .
3. The following car was stopped and has to start from a standing still position:  $A3$ .
4. The following vehicle's performance is governed by the car-following algorithm while space headway constraint is satisfied:  $A4$ .
5. The following vehicle is advanced according to the car-following algorithm with non-collision constraint:  $A5$ .

In addition to these, comfortable deceleration rates and maximum allowable deceleration rates of the following and the lead vehicles are taken into consideration in order to limit the computed values within a reasonable boundary.

The discussion of the procedures for the computation of the acceleration or deceleration rates are given in the following sections.

### Computation of $A1$

$A1$  is acceleration rate of a moving vehicle or a vehicle ready to move, constrained only by the mechanical ability of the vehicle.  $A1$  is taken from Table 1 for a given vehicle type and speed. This table is constructed based on maximum acceleration-rate information given in the *Transportation and Traffic Engineering Handbook (T&TEH) (11)*.

The normal acceleration and deceleration rates for a pas-

TABLE 1 TYPICAL ACCELERATION RATES (ft/sec/sec) FROM STANDING START TO 15 mph AND 30 mph, AND AT VARIOUS RUNNING SPEEDS THEREAFTER, ON A LEVEL ROAD

vehicle type	speed (mph)					
	0-15	15-30	30-40	40-50	50-60	>60
passenger car	8.80	5.50	5.17	4.17	3.08	2.09
tractor semi- trailer	2.20	1.10	0.88	0.44	0.44	0.44

TABLE 2 NORMAL ACCELERATION AND DECELERATION RATES (ft/sec/sec) FOR PASSENGER CARS WHEN THE DRIVERS ARE NOT INFLUENCED TO REACT RAPIDLY

speed change (mph)		acceleration	deceleration
from	0-15	4.84	7.77
	15-30	4.84	6.74
	30-40	4.84	4.84
	40-50	3.81	4.84
	50-60	2.93	4.84
	60-70	1.91	4.84

senger car are given in Table 2 (II). These values were observed where the drivers were not influenced to react rapidly. For trucks, 75% of the values in Table 2 are used.

#### Computation of $A_2$

$A_2$  is the acceleration or deceleration rate a vehicle needs to reach the desired speed ( $DS \cdot V_F$ ) or the speed limit. A driver will try to reach his desired speed or the speed limit as fast as possible while satisfying all safety and operational constraints.

The desired speed will be equal to the speed limit when a driver does not want to drive above the speed limit. In CARSIM, a courtesy factor which shows what percentages of the drivers will obey the speed limit or the suggested speed is assigned by the user.

#### Computation of $A_3$

$A_3$  is the acceleration rate of a stopped vehicle when it starts moving after a start-up delay. When a platoon of vehicles is subjected to a kinematic disturbance, the speed of the vehicles will decrease gradually and finally they may come to a complete stop. After the lead car moves, the follower will move after a few seconds of delay: "start-up delay." The investigation of Ohio State University trajectory data (12) revealed that the start-up delay for the cars in a platoon is about 1 to 3 seconds.

It is assumed that the drivers with shorter reaction times will wait less than the drivers with longer reaction times. In CARSIM, less than 20% of the drivers have a reaction time of 0.68 seconds in a surprise situation. These drivers will wait 1 second, but the rest of the drivers will wait a maximum of 2 seconds, before moving again. Comparison of the trajectory plots generated by CARSIM for four stop-and-go conditions indicated that the assumed start-up delays are very reasonable. Less satisfactory results were obtained when a 2-second start-up delay was used for all drivers. The propagation speed of the starting wave would be about 14 ft/sec when the assumed start-up delays are used.

A vehicle will not be allowed to move, regardless of how long it has been stopped, as long as the non-collision constraint is not satisfied. For a vehicle starting from a stand-still position, the acceleration rate is 2 ft/sec/sec for passenger cars and 1 ft/sec/sec for trucks for the first second of the movement. Thereafter, the acceleration is determined according to the car-following algorithm.

#### Computation of $A_4$

$A_4$  is the acceleration computed from the following equation (using equality):

$$X_L - (X_F + V_F (DT)) + 0.5 (A_4) (DT)^2 \geq L_L + K \quad (3)$$

where:

- $X_L$  = position of the lead vehicle;
- $L_L$  = length of the lead vehicle;
- $K$  = buffer space between vehicles;  $K = 10$  ft when density is not very high.

The other terms have been defined previously.

$A_4$  is the acceleration rate when a space headway of greater than or equal to the sum of the length of the lead car and a buffer space of  $K$  feet is provided. A buffer space of 10 feet, as suggested by the INTRAS model, provided satisfactory trajectory plots when density was not very high. However, study of the Ohio State University trajectory data (12) indicated that the use of a 10-foot buffer space cannot be justified at high densities. For example, for platoon 123 with fifteen cars, minimum space headway which occurred near jam density was in the range of 18.41 to 30.99 feet, and the average of the minimum distances was 22.44 feet. The minimum spacing would be different depending on the density of traffic: for near-jam density conditions, a 5- to 7-foot buffer space was used.

#### Computation of $A_5$

$A_5$  is the acceleration or deceleration rate of a vehicle when the non-collision constraint is satisfied. The following equation is used to assure that enough spacing is provided:

$$X_L - (X_F + V_F(DT) + 0.5(A_5)(DT)^2) - L_L - K \geq \text{maximum of } \begin{cases} [V_F + (A_5)(DT)](BRT), \text{ or} \\ [V_F + (A_5)(DT)](BRT) \\ + \frac{[V_F + (A_5)(DT)]^2}{2(MX.F)} - \frac{V_L^2}{2(MX.L)} \end{cases} \quad (4)$$

where:

- $BRT$  = brake-reaction time of a driver;
- $V_L$  = velocity of the lead car at the end of time interval;
- $MX.F$  = maximum deceleration rate of the following car;
- $MX.L$  = maximum deceleration rate of the lead car.

It can be seen that the non-collision constraint built in the logic of CARSIM is:

$$[V_F + (A_5)(DT)](BRT) + \frac{[V_F + (A_5)(DT)]^2}{2(MX.F)} - \frac{V_L^2}{2(MX.L)} \geq 0 \quad (5)$$

The value of  $A_5$  is determined such that after moving a following vehicle to its new position there will be enough space headway for this vehicle to react to a decelerating lead vehicle, and stop or reach a safe driving speed. Assuming the non-collision constraint is satisfied, solve equation 4 for  $A_5$  and simplify it to obtain equation 6:

$$(DT)^2(A_5)^2 + B(A_5) + C < = 0 \quad (6)$$

where

$$B = 2(V_F)(DT) + 2(MX.F)(DT)(BRT) + (MX.F)(DT)^2$$

$$C = -2(MX.F)(X_L - X_F - V_F(DT) - L_L - K - V_F)(BRT) + \frac{V_L^2}{2(MX.L)} + V_F^2$$

Solve equation 6 for  $A_5$ .

$$A_5 = [-B + (B^2 - 4(DT)^2(C))^{1/2}] / [2(DT)^2] \quad (7)$$

$A_5$  computed from equation 7 may be for acceleration or deceleration depending on the values of  $B$  and  $C$ .

Since  $A_5$  is computed using the non-collision constraint, equation 4 is evaluated using  $A_5$ . By substituting  $A_5$  in equation 4, determine whether or not the left-hand side of equation 4 is greater than or equal to the maximum of the two expressions on the right-hand side. If this condition is not satisfied, use the first expression on the right-hand side of equation 4 to compute  $A_5$ .

To use CARSIM as a dual-regime model, computations are performed with different  $MX.F$  and  $MX.L$  values. When the density is greater than 60 vpm, an  $MX.F$  of 13 ft/sec/sec, instead of 16 ft/sec/sec, is used in computations of  $A_5$ . When the density is less than 60 vpm, a maximum deceleration rate of 16 ft/sec/sec is used for both the following and the lead vehicles.

#### Proper Acceleration or Deceleration Rate

Once  $A_1$ ,  $A_2$ ,  $A_3$ ,  $A_4$ , and  $A_5$  have been computed, a comfortable deceleration rate ( $AC$ ) is also determined for each speed group from Table 2. The comfortable deceleration rate will be used when a driver slows down just to reach the posted speed limit. Using a comfortable deceleration rate would prevent a sudden decrease of speed which might cause another kinematic disturbance.

To choose the proper acceleration value, the program finds the minimum of  $A_1$ ,  $A_2$ ,  $A_3$ ,  $A_4$ , and  $A_5$  and uses this positive number as the acceleration rate. The proper deceleration value is either comfortable deceleration ( $AC$ ), or  $A_2$ , or  $A_5$ . It is always less than the maximum deceleration rate of 16 ft/sec/sec. If  $A_2 < AC$  and  $AC < \text{minimum of } (A_4, A_5)$ , then  $AC$  is used. If  $AC < A_2$  and  $A_2 < \text{minimum of } (A_4, A_5)$ , then  $A_2$  should be used; otherwise,  $A_5$  is used.

After determining the proper acceleration or deceleration rate, the speed and the position of a vehicle are computed using equations 1 and 2. This vehicle is advanced to a new position and the rest of the vehicles are moved in a similar fashion. This process is repeated in 1-second time intervals for all vehicles in the system. For each driver-vehicle unit in CARSIM, a set of attributes is assigned before it is allowed to enter the model. These attributes are discussed in the following section.

#### Driver-Vehicle Characteristics

Vehicles are generated one at a time, and a set of attributes is assigned to each vehicle upon generation. The driver-vehicle characteristics such as type and length of vehicle, emergency deceleration, complying index, location, identification number, desired speed, brake reaction time, and arrival time

are assigned to every unit. The traffic mix is a user-specified input variable.

In developing the model, several decisions are made with regard to the inter-arrival time, brake reaction time, and speed distribution of the vehicles. These topics will be discussed in the following sections.

For light traffic, where there is less vehicle interaction, the inter-arrival time constitutes an exponential function. As traffic volume increases, the interaction between vehicles becomes more frequent. Considering the behavior of drivers in the car-following situations, one may expect the headway distribution to have an exponential tail. Therefore, the headways can be generated from functions such as log normal, truncated normal, or shifted exponential. For this model, the time headway between successive vehicles is generated from a negative shifted exponential distribution.

Another characteristic assigned to each vehicle is the desired speed. One important factor influencing the shape of the speed distribution plot is the density of traffic. In high density traffic the variation in speed is not as much as in free-flow traffic conditions, thus more uniformity is expected. Duncan (14), Breiman et al. (13), and Pahl (15) suggested normal distribution, while Ashworth (16) proposed an Erlang distribution instead of a normal distribution for speed. In CARSIM, the desired speed of vehicles is generated from a truncated normal distribution with a mean of 55 mph and a standard deviation of 5 mph.

A full description of the other attributes used in CARSIM is described elsewhere (1). In the following section, the discussion about brake reaction time is presented.

### Brake Reaction Time

The response of a driver jointly varies with that driver's stimulus and sensitivity. Thus, the brake-reaction time not only varies among the drivers but also changes for a given driver under different conditions. Johansson and Rumer (17) reported that the median brake-reaction time was 0.66 seconds and the range was between 0.3 to 2.0 seconds in alerted situations. A recent study by Olson et al. (18) used two groups of drivers: a younger group, age 40 or younger; and an older group, 60 or more years old.

For surprise conditions, the 5th and 95th percentile range was 0.85 to 1.6 seconds for young drivers and very close to these values for older drivers. For young drivers, the 5th and 95th percentiles for alerted conditions were 0.57 and 1.37. For older drivers, the values were a little longer.

The reaction times used in CARSIM were obtained from a cumulative distribution plot based on Johansson and Rumer's data. The reaction time varies from 0.4 to 1.5 seconds for alerted (congested) conditions, in increments of 0.1 seconds. There are twelve different reaction times, and one of them is assigned randomly to each driver. The probability of assigning each one of the twelve values is not the same. The percentage of drivers having reaction times of less than or equal to the specified value are given in Table 3 for alerted and surprise situations. The reaction times used in CARSIM are very close to the young drivers' reaction times in Olson's study.

The use of varying reaction times is more realistic than using a constant value for all densities and more reasonable than using a constant value for all drivers. When a driver is in a platoon or in unexpected traffic congestion on the freeway, he would be driving with more attention to the situation than when he sees only a few cars on the freeway. Due to this fact, CARSIM uses shorter reaction times when density is greater than 60 vpm.

However, adjustments are made for trucks such that no truck with a reaction time of less than 1 second is allowed to enter the system.

### DATA BASE

The data used for validation of CARSIM are the Ohio State University trajectories data collected by Treiterer (12) using aerial photogrammetric techniques. The photographs were taken in 1-second time intervals. The location of a vehicle is estimated to be accurate within  $\pm 0.50$  feet and the speed is determined with an error of  $\pm 1.0$  mph. The data was collected from I-70 in Columbus, Ohio, a 3.5-mile-long section with three on- and three off-ramps. The data provides spacings, headways, longitudinal positions, and speeds for all vehicles.

From the data, the following four platoons which experi-

TABLE 3 REACTION TIME OF DRIVERS (IN SECONDS) IN LIMITED ANTICIPATION (ALERTED) AND SURPRISE CONDITIONS

% of drivers	alerted situation	surprised situation
100	1.50	2.03
98	1.40	1.89
96	1.30	1.76
94	1.20	1.62
90	1.10	1.49
88	1.00	1.35
81	0.90	1.22
72	0.80	1.08
64	0.70	0.95
48	0.60	0.81
20	0.50	0.68
4	0.40	0.54

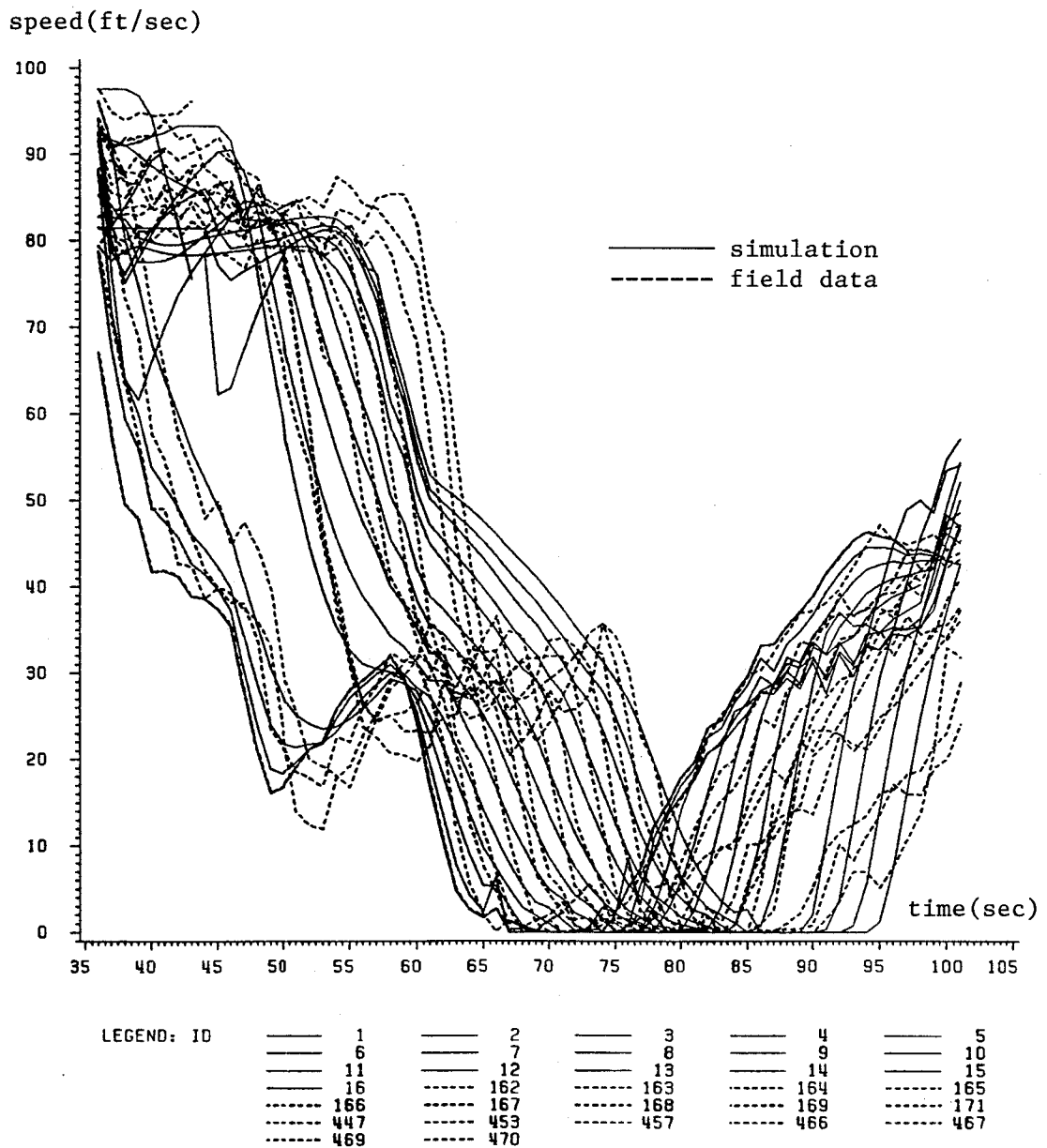
enced a stop-and-go condition are used in the validation of CARSIM:

- (1) Platoon 123: A group of fifteen vehicles which has no vehicles entering or leaving the platoon.
- (2) Platoon 126: The platoon started with fifteen vehicles, and after several seconds two cars left the platoon and one car joined to it. This group of fourteen vehicles went through a kinematic disturbance and lost another car before recovering and one more car after recovering from the disturbance.
- (3) Platoon 127: This platoon reached a very high density. It started with ten vehicles and remained a ten-vehicle platoon.
- (4) Platoon L123X: A group of five vehicles following each other for 202 seconds. The vehicles in this platoon are also in Platoon 123.

**VALIDATION OF CARSIM**

Validation of CARSIM was performed at microscopic and macroscopic levels. At the microscopic level, the location and the speed of each vehicle from the field data were compared with those computed from CARSIM; whereas at the macroscopic level, the average speed, density, and volume from the field data were compared with the values calculated by CARSIM. The discussion of different aspects of the procedures used for validation of this simulation model are provided elsewhere (19-26).

Four different platoons of vehicles covering a wide range of traffic operations were used for the comparison of field data with CARSIM's results. Here, only the results for platoon 126 are presented. (The other platoons' results were very similar.)



**FIGURE 1** Speed change patterns generated by CARSIM and field data for a stop-and-go condition (platoon 126).

Five independent replications of CARSIM were made for each real-world condition. In the replications, the characteristics of vehicles and drivers were generated randomly and assigned to them. However, the speed and location of the first vehicles in CARSIM and the real-world platoon were the same.

**Validation at Microscopic Level**

The average speed and location of each vehicle were computed at time intervals of 1 second and were compared to the values obtained from the field data for that vehicle. For platoon 126, the plot of the average speeds computed by CARSIM versus the values obtained from the field data are shown in Figure 1. As can be seen, the speed change patterns generated by CARSIM were very close to those from the field data. It should be noted that the last three vehicles of the real-world platoon were not closely following the other cars in the platoon, while the last cars in the simulated platoon were keeping up with the lead cars.

For the trajectory comparison, the average location of each vehicle computed from the replications was compared with the location determined from the field data, as shown in Figure 2. This figure shows that all of the vehicles in platoon 126 were forced to stop for awhile and then move. CARSIM is

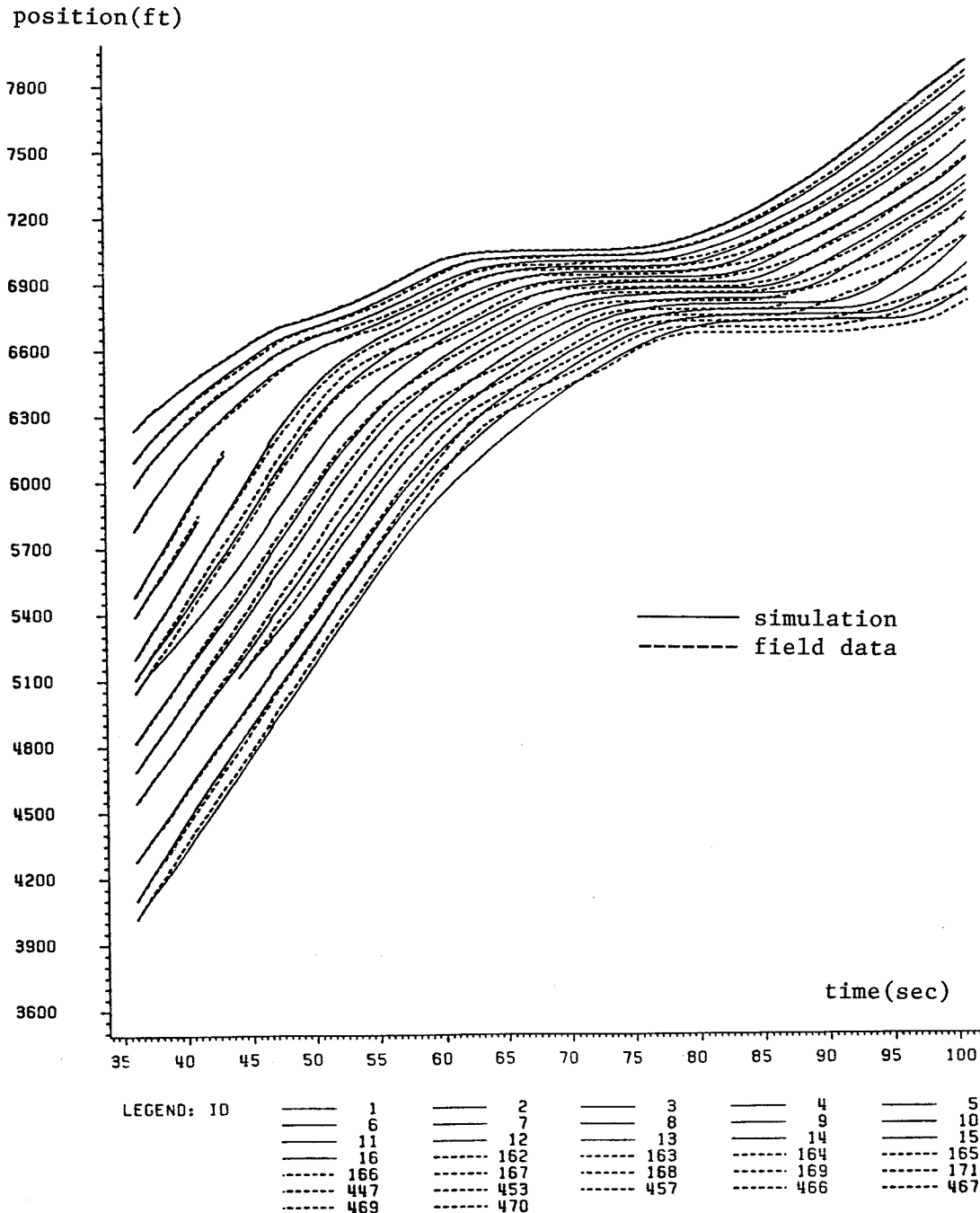
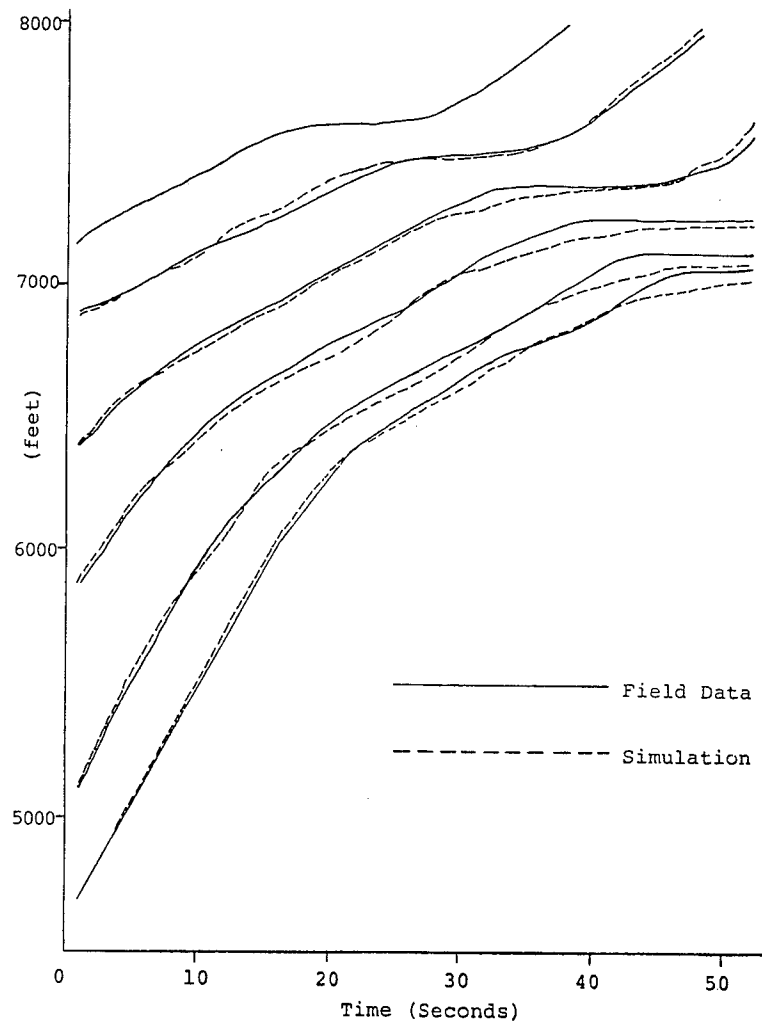


FIGURE 2 Trajectories generated by CARSIM and field data for a stop-and-go condition (platoon 126).



**FIGURE 3** INTRAS trajectory validation graph, showing every fifth car in the platoon (9).

capable of simulating such stop-and-go operations in a realistic manner. No statistical tests were run; only visual comparisons were made at the microscopic level.

When the traffic flow is not in a steady state condition, it is quite challenging to generate trajectory plots from simulation models that would replicate the actual trajectory plots. It is not clear how the existing car-following models would perform when subjected to this challenge. For example, the INTRAS model (9) used only one phase of going through a kinematic disturbance, namely the deceleration phase, and did not use the second phase (Figure 3). For comparison of the trajectories generated by INTRAS and CARSIM, see Figures 2 and 3. The results from CARSIM show the deceleration and the acceleration phase (stop-and-go), but the results from INTRAS show only the deceleration phase.

#### Validation at Macroscopic Level

At the macroscopic level, the overall performance of a platoon of vehicles, rather than the performance of an individual vehicle, was evaluated.

For each platoon, the following comparisons were made between CARSIM's results and field data:

1. Comparison of flow parameters over time,
2. Comparison of fundamental relations of traffic flow, and
3. Regression of similar results versus the field data.

#### Comparison of Flow Parameters

The flow parameters used for comparisons are: speed, density, and volume. These parameters are computed from CARSIM and the field data at 1-second time intervals for every platoon. For platoon 126, the plot of the average speeds from CARSIM versus the speeds from the field data is shown in Figure 4. As can be seen, the speed computed by CARSIM was very close to the actual speed of the platoon. It is important to note that platoon 126, from its initial speed of about 80 fps, reached a speed of near zero in less than 1 minute; and that CARSIM simulated such a rapid speed reduction. The simulation curve shows less fluctuation than the curve for the field data, as expected.



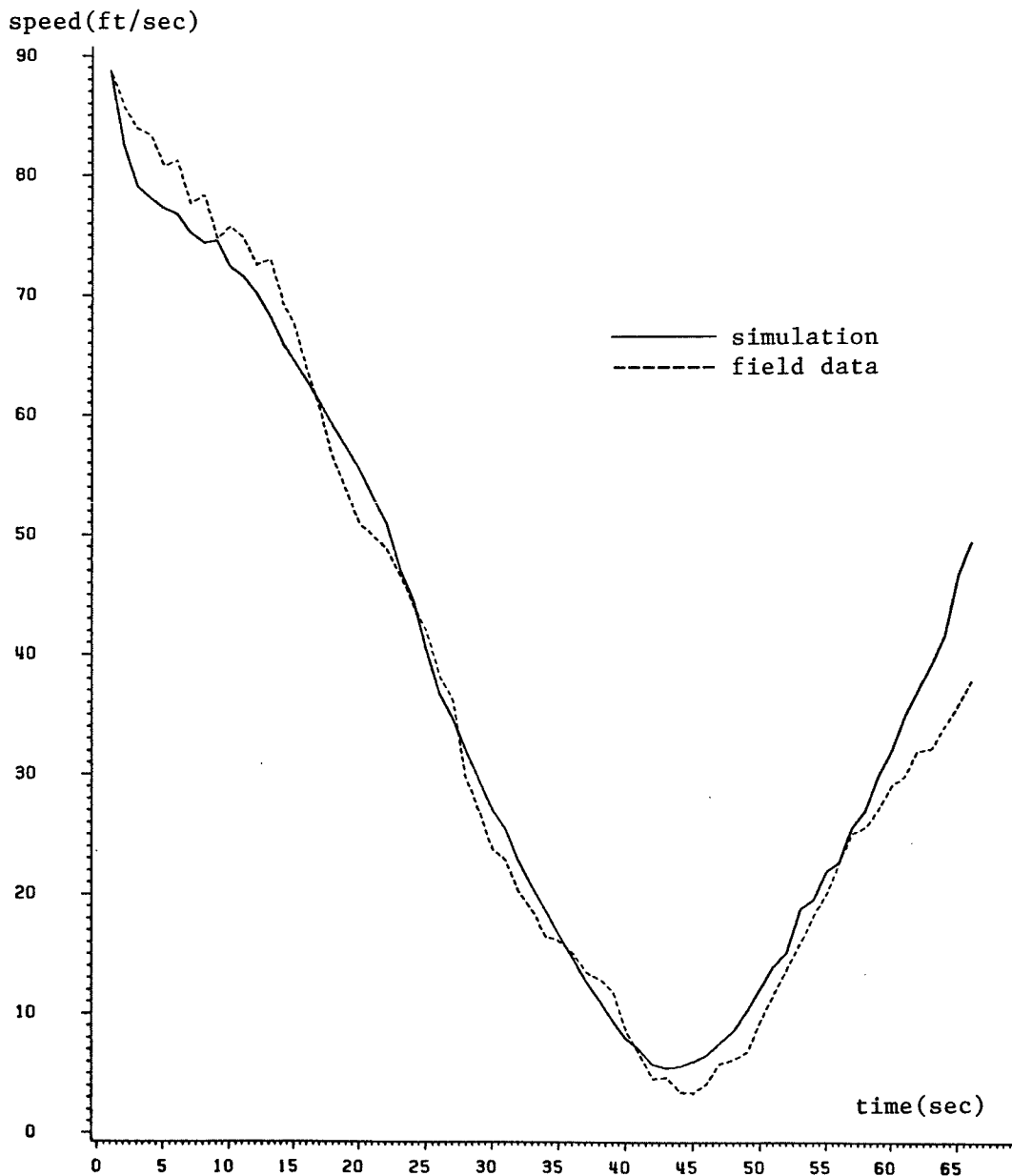


FIGURE 4 A platoon speeds computed from CARSIM and field for a stop-and-go condition (platoon 126).

The plot of densities computed from CARSIM and the field data are shown in Figure 5. This graph shows very similar density fluctuation curves for the simulated platoon and for the actual platoon. It should be noted that the time in which the simulated platoon reached the jam density was very close to that of the actual platoon. One should be very careful in using the density of platoon when comparing simulation results with field data, because the density of a platoon is computed from the position difference of the first and the last car in the platoon and is very dependent on the spacing between these vehicles.

#### Comparison of Speed-Density Plots

The relationship between speed, density, and volume computed from the field data and CARSIM was compared for all

platoons (1). For platoon 126, the speed-density curves from CARSIM and field data are shown in Figure 6. A nonlinear relationship and a loop between the values obtained before and after the disturbance is produced by the field data and by CARSIM. The loop was more distinct when the results from each replication of CARSIM were plotted against the field data. The loop is an indication of the dual behavior of traffic before and after a disturbance.

In addition to the graphical presentation of the results, regression analysis of the simulation results versus the field data was also carried out. The results will be discussed in the following section.

#### Regression of Simulation Results vs. Field Data

Regression of the speed, density, and volume computed from CARSIM versus the values obtained from the field data were

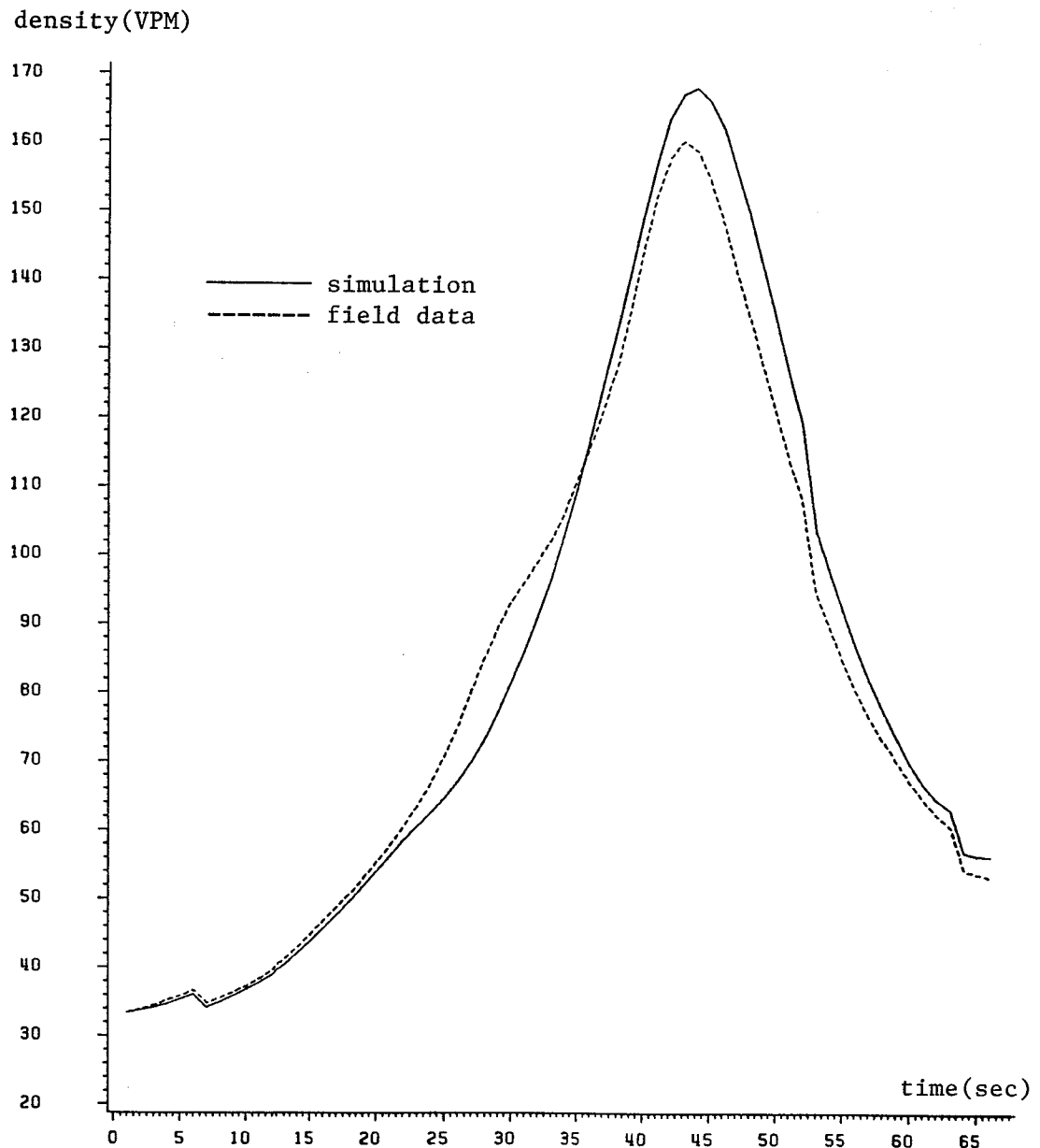


FIGURE 5 A platoon densities computed from CARSIM and field data in a stop-and-go condition (platoon 126).

carried out for all platoons. The average values of speed, density, and volume were computed from the replications in 1-second time intervals for all platoons and used for comparison. Table 4 gives the summary of the regression analysis. It can be seen that the slopes of the regression lines ( $b_1$ ) are very close to 1, and the y-intercepts ( $b_0$ ) are almost zero. This combination of slope and y intercept indicates that the plot of CARSIM's results versus the field data are scattered around a line going through the origin with a 45-degree angle. This means CARSIM's results are very close to the value from the field data.

The regression analysis indicates that there is a strong agreement between the simulation and the real-world results. The  $R$ -squared values for the regression of speed and density from CARSIM versus the values from the field data are 0.98 or higher. Such high  $R$ -squared values and low variability on

the slopes of the regression lines indicate that the results obtained from CARSIM are very close to the values computed from the field data.

The graphical comparison and the regression analysis of the results from CARSIM and the field data indicate that CARSIM realistically reproduces normal and stop-and-go traffic flow conditions on freeways. From this comparison, the result seems reliable. For further validation, more field data from stop-and-go conditions on freeways is needed. In the following section, one application of CARSIM is demonstrated.

#### APPLICATION

This example is to demonstrate how CARSIM handles a road blockage and how the traffic waves propagate and dissipate.

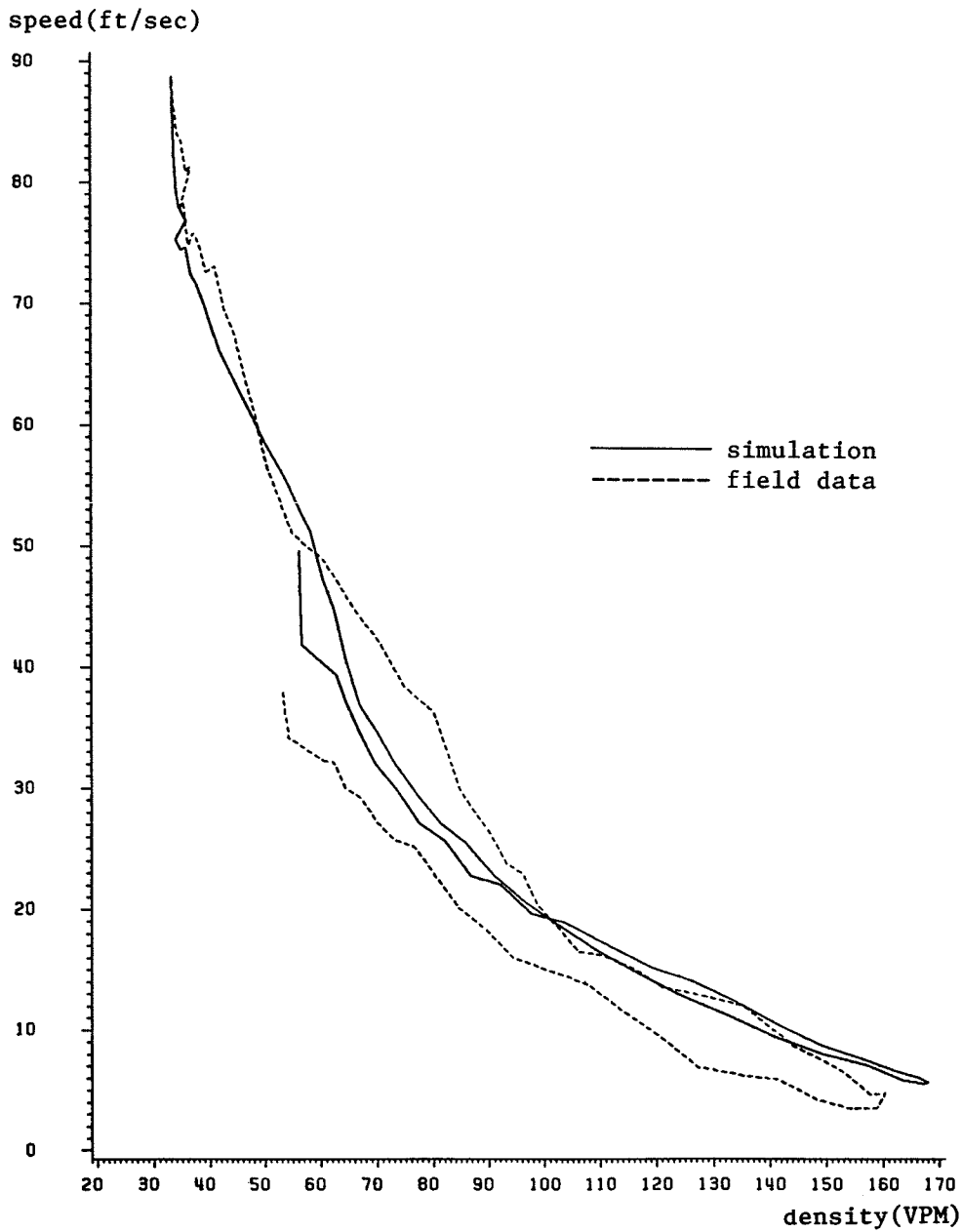


FIGURE 6 A speed-density relationship computed from CARSIM and field data for a stop-and-go condition (platoon 126).

TABLE 4 RESULTS OF REGRESSION OF CARSIM'S VALUES VS. FIELD DATA

type of data	platoon number	b0	b1	R**2	s(b0)	s(b1)
	123	1.70554	0.94705	0.98383	0.49177	0.01220
Speed	126	3.26491	0.93782	0.98621	0.63354	0.01386
	127	-1.14070	1.02664	0.98708	0.52923	0.01105
	123	1.43827	0.98432	0.98922	0.88620	0.01033
Density	126	-5.12594	1.08229	0.98183	1.65879	0.01841
	127	2.46060	0.93949	0.97765	1.22708	0.01336

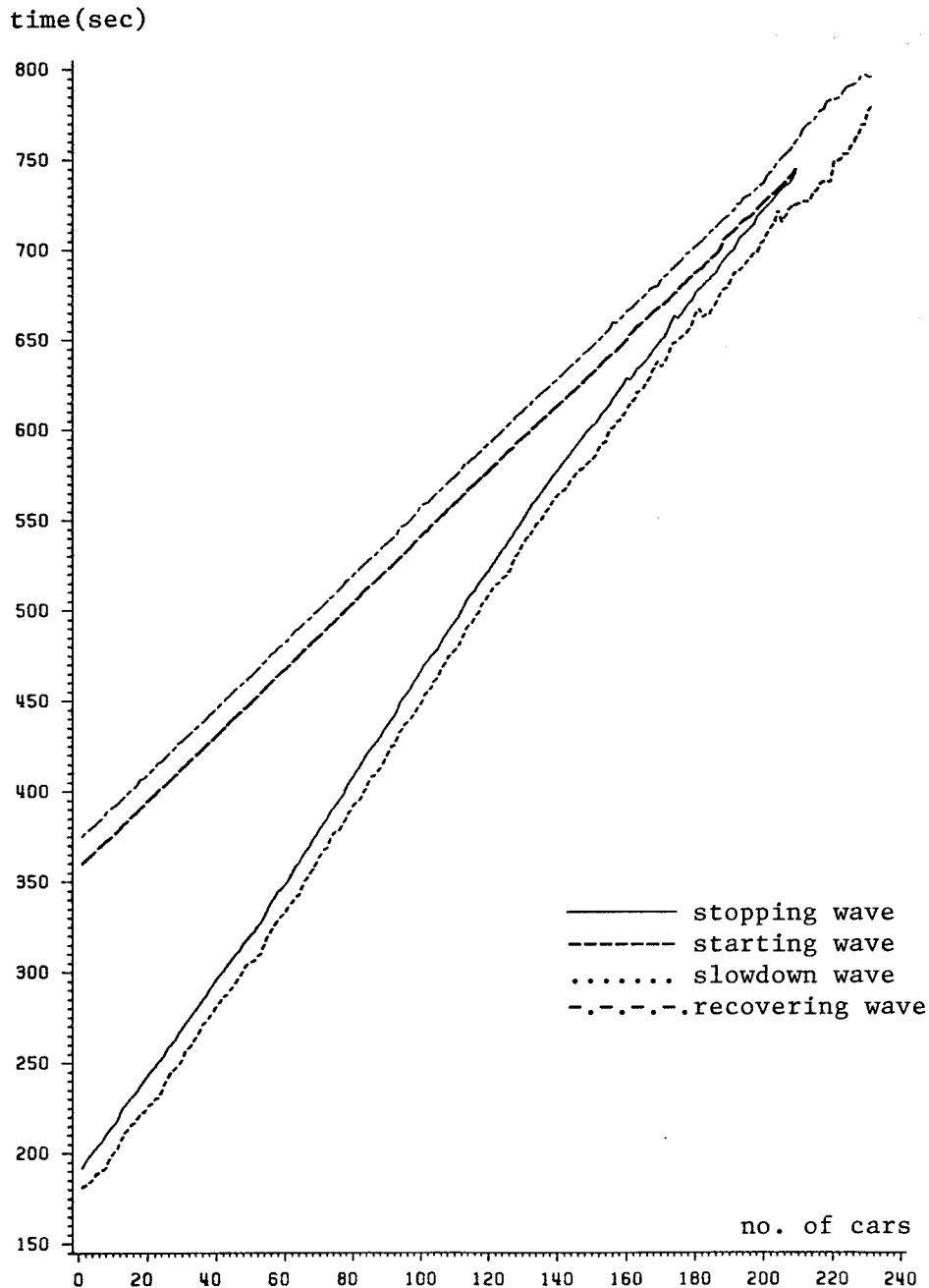


FIGURE 7 Simulation of effect of a 3 minute road blockage on propagation and dissipation of traffic waves (volume is 1,200 VPH).

Due to an incident, a roadway is blocked for 3 minutes and the arriving vehicles form a very long queue. The closed section is a single-lane road with no exit, such as a long bridge or a construction zone. The vehicles are allowed to accelerate to a desired speed of 55 mph when the incident is removed. While the cars in the front of the queue accelerate to reach the desired speed, the arriving cars join the rear of the queue. Through this process, the location of the bottleneck moves to upstream traffic. The traffic volume is 1200 vph and the desired speed is 55 mph (80.67 fps). The incident happens 180 seconds after the lead car enters the system and forces the cars to decelerate to a complete stop. There are enough cars in the system (about 240 cars) that the effect of 3 minutes'

road blockage does not reach the last vehicle in the system. After 3 minutes, the lead car is allowed to move and accelerate to the desired speed. The time a vehicle slows down, stops, starts moving again, and reaches the desired speed is determined by CARSIM for all vehicles affected by the road blockage.

Four different kinematic traffic waves are identified as: (1) slow-down wave, (2) stopping wave, (3) starting wave, and (4) recovering wave. The waves are evidenced by the effect imposed on the following car due to the road blockage. Figure 7 shows the time each wave is reached by a given vehicle.

At the point a starting wave reaches a stopping wave, the queue of stopped cars is eliminated. Similarly, at the point

the recovering wave reaches the last affected vehicle, the effect of road blockage is completely eliminated. The relationships between these waves can easily be used to study characteristics of traffic congestions. As it can be seen in Figure 7, the slope of the starting wave is less than the slope of stopping wave. This indicates that more vehicles will move, rather than stop, in a certain period of time. When the slope of the moving wave is equal to or greater than that of the slope of the stopping wave, there will always be a queue of stopped vehicles. Application of CARSIM for traffic wave studies is discussed elsewhere (1).

## CONCLUSIONS

A car-following model, CARSIM, is developed to simulate not only normal but also stop-and-go traffic flow conditions on freeways. More realistic features to reflect the behavior of traffic in stop-and-go conditions are included in the car-following algorithm of this model. For validation of CARSIM, graphical and statistical techniques were used and the results from CARSIM were compared to the field data. The comparison of the trajectory and the speed-change plots yielded satisfactory results. The regression analysis of speeds and densities computed from CARSIM versus those from the field data resulted in R-squared values of 0.98 or higher, indicating a strong agreement between the simulation results and real-world traffic data.

## REFERENCES

1. R. F. Benekohal. *Development and Validation of a Car Following Model for Simulation of Traffic Flow and Traffic Wave Studies at Bottlenecks*. Ph.D. Dissertation. The Ohio State University, Columbus, Ohio, 1986.
2. E. B. Lieberman. *Integrated Traffic Simulation Model Phase 1*, Vol. 1, Executive Summary. Report FHWA-RD-80-086. FHWA, U.S. Department of Transportation, Sept. 1980.
3. G. Radelat. Simulation Developments in Progress. In *Special Report 194*, TRB, National Research Council, Washington, D.C., 1981.
4. D. N. Goodwin. *Examination of Nine Freeway Simulation Models*. System Development Corporation TM-4638/014/01, Santa Monica, Calif., July 1972.
5. Y. S. Hsu and P. K. Munjal. Freeway Digital Simulation Models. In *Transportation Research Record 508*, TRB, National Research Council, Washington, D.C., 1974, pp. 29-41.
6. A. D. May. Models for Freeway Corridor Analysis. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, pp. 23-32.
7. D. R. P. Gibson. Available Computer Models for Traffic Operations Analysis. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, pp. 12-22.
8. A. S. Byrne, et al. *Handbook of Computer Models for Traffic Operations Analysis*. Report FHWA-TS-82-213. FHWA, U.S. Department of Transportation, 1982.
9. D. Wicks, et al. *Development and Testing of INTRAS, A Microscopic Freeway Simulation Model*, Vols. I-IV. Final Reports. FHWA, U.S. Department of Transportation, 1980.
10. E. C. Russell. *Building Simulation Models With SIMSCRIPT II.5*. C.A.C.I., Inc., Los Angeles, Calif., 1982.
11. *Transportation and Traffic Engineering Handbook*, 2nd ed. Institute of Transportation and Traffic Engineering, Washington, D.C., 1982.
12. J. Treiterer. *Investigation of Traffic Dynamics by Aerial Photogrammetry Techniques*. Transportation Research Center, Department of Civil Engineering, Ohio State University, Final Report EES 278, Feb. 1975.
13. L. Breiman, R. Lawrence, D. Goodwin, and B. Bailey. The Statistical Properties of Freeway Traffic. *Transportation Research*, Vol. 11, 1977, pp. 221-228.
14. N. C. Duncan. *A Method of Estimating the Distribution of Speeds of Cars on a Motorway*. Transport and Road Research Laboratory, Report 598, 1973.
15. J. Pahl. The Effect of Discrete Time Measurement on Speed Data. In *Highway Research Record 349*, HRB, National Research Council, Washington, D.C., 1971.
16. R. Ashworth. An Alternative Procedure for Estimating Speed Distribution Parameters on Motorways and Similar Roads. *Transportation Research*, Vol. 10, 1976, pp. 25-29.
17. G. Johansson and K. Rumer. Drivers' Brake Reaction Times. *Human Factors*, Vol. 13, No. 1, Feb. 1971, p. 23.
18. D. L. Olson, et al. *NCHRP Report 270: Parameters Affecting Stopping Sight Distance*. TRB, National Research Council, Washington, D.C., 1984.
19. R. S. Sargent. Verification and Validation of Simulation Models. In *Progress in Modeling and Simulation* (F. E. Cellier, ed.), Academic Press, 1982.
20. G. S. Fishman. Problems in the Statistical Analysis of Simulation Experiments: The Comparison of Means and the Length of Sample Records. *Communications of the ACM*, Vol. 10, No. 2, 1967, pp. 94-99.
21. J. F. Torres, A. Halati, and A. Gafarian. *Statistical Guidelines for Simulation Experiments*. Executive Summary, Prepared for FHWA by JFT Associates, Culver City, Calif., 1983.
22. J. P. C. Kleijnen. Design and Analysis of Simulations: Practical Statistical Techniques. *SCI Simulation*, Vol. 28, No. 3, March 1977.
23. J. P. C. Kleijnen. Experimentation with Models: Statistical Design and Analysis Techniques. In *Progress in Modeling and Simulation* (F.E. Cellier, ed.), Academic Press, 1982.
24. J. R. Wilson and A. B. Pritsker. A Survey of Research on the Simulation Startup Problem. *Simulation*, Vol. 31, No. 2, Aug. 1978.
25. J. R. Wilson and A. B. Pritsker. Evaluation of Startup Policies in Simulation Experiments. *Simulation*, Vol. 31, No. 3, Sept. 1978.
26. J. P. C. Kleijnen. Antithetic Variates, Common Random Numbers and Optimal Computer Time Allocation in Simulation. *Management Science*, Vol. 21, No. 10, 1975.

---

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

# Work Zone Analysis Model for the Signalized Arterial

C. THOMAS JOSEPH, ESSAM RADWAN, AND NAGUI M. ROUPHAIL

The purpose of this paper is to illustrate new theoretical concepts used to represent traffic flow in work zones. The paper presents the development and applications of a microcomputer program—Work Zone Analysis Tool for the Arterial (WZATA)—for the analysis and evaluation of a system consisting of a lane closure between two signalized intersections. The program consists of two parts: a semi-simulation model to represent and analyze flow between the intersections, and a macroscopic model to represent traffic characteristics at the downstream intersection. New techniques were developed to represent percentage merges and vehicle merge characteristics before the lane closure. The user was provided direct control over the merging of every vehicle. A modified version of the continuum-flow theory was utilized to represent and analyze traffic flow at the downstream intersection. Flow was considered to be composed of two parts: the platoon flow and the non-platoon flow. The techniques of analysis used also considered acceleration, deceleration, and start/stop losses. The program is written in Microsoft BASIC for the IBM-PC/XT/AT and is structured to facilitate easy modification of data as well as analysis of various hypothetical situations. Attempts are being made to include graphical display facilities in the model.

Work zone management constitutes the most common management and maintenance task of any transportation system. Many streets require regular repair or upgrading. Construction and maintenance activities performed often require the closure of at least one lane to traffic. The decrease in the number of lanes usually manifests itself in an increase in traffic congestion and consequent reduction of the level of service. An effort should therefore be made to schedule work zone activities and modify signal characteristics to reduce delay and congestion to the minimum.

A considerable number of analytical and computer techniques and models are available to maintenance personnel to aid in decision making and scheduling of work zone lane closures on the arterial. The drawback to almost all these techniques is that, in the final analysis, an arterial is divided into several segments and each is analyzed individually—with no relation to the other. For example, in a typical analysis of a work zone in an arterial, the arterial is divided into three sections:

1. The intersection;
2. The strip between the intersections, but not including the lane closure; and
3. The work zone, including the lane closure.

C. Joseph and E. Radwan, Center for Advanced Research in Transportation, Arizona State University, Tempe, Ariz. 85287. N. Roupail, University of Illinois, Chicago, Ill. 60680.

Each section was analyzed individually for the different scenarios, and the final decision was based on a logical combination of the results of all three analyses. To date, while several comprehensive computer programs (such as QUEWZ(1) and FREECON(2)) have been developed to analyze work zones on freeways, no model has been developed to analyze the arterial system with the work zone as an overall comprehensive unit.

This paper culminates the development of a macroscopic semi-simulation model to analyze the work zone in an arterial system as an overall comprehensive unit.

## MODEL DEVELOPMENT

A review of the literature on work zones revealed that a completely macroscopic model would be too approximate to represent conditions of traffic flow in an arterial with a work zone, since the phenomena of merging and diverging before and after lane closure is complex and involves several inter-related parameters. It was therefore decided to develop the model in two parts: a semi-simulation model to represent flow between the intersections and through the lane closure zone, and a macroscopic model to represent traffic characteristics at the downstream intersection. Illustration of a few fundamental theoretical aspects of the developed model follow. A detailed explanation of all aspects of the model can be found elsewhere(3).

## SEMI-SIMULATION MODEL

A technique utilizing a concept of "real" and "imaginary" vehicles was used to simplify the simulation model to a considerable extent.

### Real Vehicles

These vehicles represent actual vehicles in the open lane that follow each other according to the car-following rules. All parameters like delay, travel time, and bandwidth are estimated for these vehicles alone.

### Imaginary Vehicles

These vehicles, on the other hand, are not actual vehicles but pseudo vehicles introduced to simulate merging and diverging

behavior before and after the lane closure. They could be considered as vehicles originally traveling in the lane about to be closed. Outside the proximity of the lane closure, they have the same characteristics as their real counterparts in the open lane. Within the proximity of the lane closure, however, their characteristics are manipulated to represent the additional headway required by the real vehicles to accommodate the merging of vehicles from the closed lane. Delay and other parameters of interest are not estimated for these vehicles.

The basic assumptions made in the semi-simulation model are as follows:

1. Initial headway of vehicles is assumed to be directly proportional to initial velocity. All vehicles beginning from the upstream intersection have the same initial velocity and therefore the same headway.
2. Vehicles are distributed equally and have similar characteristics in all lanes.
3. Passenger cars constitute 100% of the traffic.
4. When vehicles merge from the closed lane, they merge at equal intervals in such a way that the traffic volume and characteristics upstream from the lane closure are similar in both lanes at all times.
5. All vehicles merging from the closed lanes distribute themselves equally to all open lanes.
6. Only vehicles in the platoon are assumed to experience any sort of delay in travel between the intersections. Non-platoon vehicles are assumed to be able to travel at their desirable speed and hence experience no travel delay.

All the above assumptions were made in an attempt to reduce the complexity of the simulation model without introducing major errors in the estimation of traffic flow characteristics and calculated parameters. Assumption 1 is realistic in an arterial setting that is part of a network, since signals tend to group vehicles together and attribute similar characteristics to them. Assumption 3 is also appropriate in an arterial, since it is well known that traffic in an arterial, in general, constitutes a very low percentage of trucks. Assumptions 4 and 5 are directly related to the default values assumed by the model during execution. Options are provided to alter conditions to analyze situations when some of the assumptions are not valid.

#### CAR-FOLLOWING RULES USED IN SEMI-SIMULATION MODEL

The basic philosophy adopted for the car-following rules is to provide the following driver enough gap between vehicles so that the status of his vehicle can be brought to that of his leader, without ever reducing the gap between the vehicles to less than a minimum.

Two conditions, those of steady and varying flow, were considered in the development of this model. During steady flow (no acceleration or deceleration of vehicles), vehicles travel behind each other at a distance equal to the product of the reaction time times the velocity plus a minimum headway (headway at zero velocity) between vehicles.

Mathematically:

$$HW = V \cdot C + HW_{\min}$$

where:

- $HW$  = the distance headway between vehicles,
- $V$  = velocity of the vehicle of interest,
- $C$  = car-following constant, and
- $HW_{\min}$  = minimum acceptable headway between vehicles

Under varying flow conditions (acceleration/deceleration of vehicles), different cases are considered based on the relationship between the velocities of the leading and the following vehicles. A more detailed explanation of the car-following rules is provided elsewhere (3).

#### MODEL REPRESENTATION OF TRAFFIC CHARACTERISTICS IN CLOSED AND OPEN LANES

To make the simulation of vehicles less complex, the arterial system (including the work zone) was simplified in a two-stage process.

In stage 1, all arterial segments, with both the total number of lanes and the number of lanes closed having a common divisor, are divided until further "irreducible". For example, a 4-2 configuration would be reduced to a 2-1 configuration.

In stage 2, vehicles from the open and closed lanes are combined in one lane and configured to represent a symmetrical and uniform merge situation (default) of vehicles in a single lane at the lane closure.

The two-stage process is illustrated in Table 1. As mentioned earlier, the concept of "real" and "imaginary" vehicles is used to represent vehicles that were initially (before the work zone) in the open or closed lane, respectively. After the default configuration is estimated by the model, the user is provided an option to modify it if necessary.

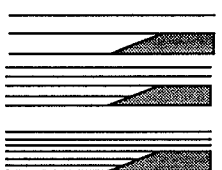
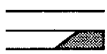
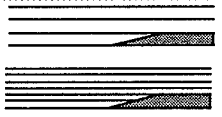

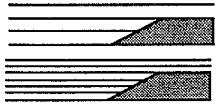
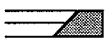


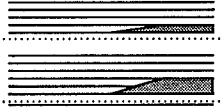

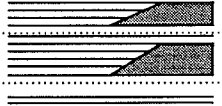
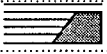
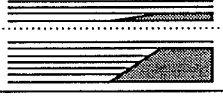

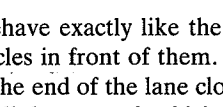
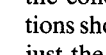
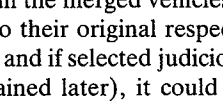
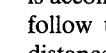
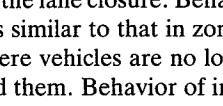
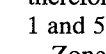
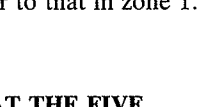
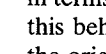
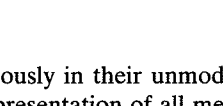
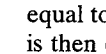
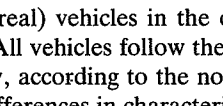
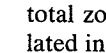
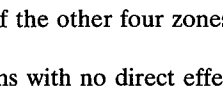
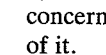


The entire length of roadway between the intersections is divided into five zones, each representing sites of different flow characteristics.

Zone 1 represents the length of roadway before the work zone where there is no modification of flow characteristics directly due to the existence of a lane closure ahead. Characteristics of the imaginary vehicles are exactly the same as those of their counterpart real vehicles in the open lane.

Zone 2 represents that part of the roadway where the driver is directly influenced by the work zone ahead, but not yet alongside, the actual lane closure. A logical beginning of this zone could be either the distance at which the warning sign is visible, the distance at which the lane closure itself is visible, or the distance at which the drivers are expected to react to the effect of the work zone ahead. Characteristics of the imaginary vehicles in this zone begin to differ from their counterpart real vehicles in the open lane. This is to slowly introduce additional headway between the real vehicles to accommodate prospective mergers.

Zone 3 is the zone actually constituting the lane closure alone. The configuration referred to in the previous section and displayed in column 3 (model default) of Table 1 represents actual vehicle positions of merged (imaginary) and real vehicles in this zone, as depicted by this model. Imaginary vehicles begin to follow real vehicles in the open lane according to the normal car-following rules. This is the only zone

TABLE 1 ILLUSTRATION OF THE TWO STAGE SIMPLIFICATION PROCESS OF A ROAD SEGMENT WITH THE WORK ZONE

WORK ZONE CONFIGURATION	MODIFIED	MODEL DEFAULT
		RIRIRIRIRIR
		RRRIRIRIRRI
		RIIRIRIRIRI
		RRRIRIRIRRI
		RIIRIRIRIRI
		RRRIRIRIRRI
		RIIRIRIRIRI
		RRRIRIRIRRI
		RIIRIRIRIRI
		RRRIRIRIRRI
		RIIRIRIRIRI
		RRRIRIRIRRI
		RIIRIRIRIRI
		RRRIRIRIRRI
		RIIRIRIRIRI

in which the imaginary vehicles behave exactly like the real vehicles in terms of following vehicles in front of them.

Zone 4 is the zone beginning at the end of the lane closure and extending to a location where all the merged vehicles are assumed to have diverged back into their original respective lanes. This location is user-defined; and if selected judiciously with the "diverge constant" (explained later), it could represent any "diverge" behavior after the lane closure. Behavior of imaginary vehicles in this zone is similar to that in zone 2.

Zone 5 is the "normal" zone where vehicles are no longer influenced by the work zone behind them. Behavior of imaginary vehicles in this zone is similar to that in zone 1.

**TRAFFIC CHARACTERISTICS AT THE FIVE ZONES**

The configurations discussed previously in their unmodified form (column 3, Table 1) are a representation of all merged (imaginary) and already existing (real) vehicles in the open lane of the lane closure (zone 3). All vehicles follow the one in front, whether real or imaginary, according to the normal car-following rules in this zone. Differences in characteristics of traffic flow as existing in each of the other four zones are illustrated below.

Zones 1 and 5 represent locations with no direct effect of

the conditions in the work zone. Hence, no merging operations should take place; and the configuration would represent just the real vehicles following each other, and not an imaginary vehicle that may exist between them. This modification is accomplished by assigning to all the imaginary vehicles that follow the real vehicle in this zone the same velocity and distance as that of the closest real vehicle ahead. A configuration like RIRIRIRIRIRI, as represented in Table 1, would therefore be equivalent to an RRRRRR configuration in zones 1 and 5 (Figure 1a).

Zones 2 and 4 represent zones where the effect of the work zone is being sensed, resulting in a general reaction reflected in terms of the merging or diverging of vehicles. To represent this behavior, a two-step manipulation of the parameters of the original configuration is undertaken, as follows:

Step 1: The imaginary vehicles are first attributed velocities equal to the real vehicle ahead and closest to them. Headway is then calculated using steady flow analysis.

Step 2: The ratio of the distance traveled by the real vehicle involved, from the beginning of the concerned zone to the total zone length, is then estimated. The headway as calculated in step 1 is then multiplied by this term, raised to some power (user-defined). The resulting term is again multiplied by the total number of vehicles between the imaginary vehicle concerned (including itself) and the closest real vehicle ahead of it.



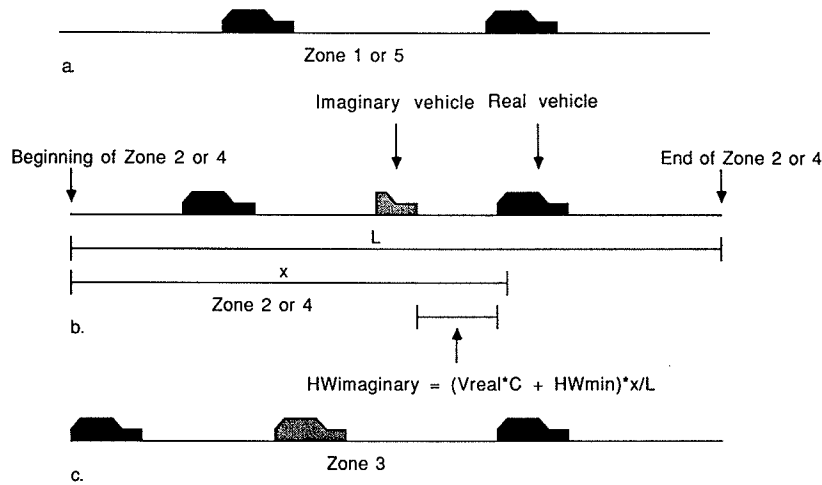


FIGURE 1 Illustration of traffic characteristics in the different zones.

Mathematically, the resulting term for Zone 2 is

$$HW_{imaginary} = (V_{real} * C + HW_{min}) * r * m * n$$

For Zone 4,

$$HW_{imaginary} = (V_{real} * C + HW_{min}) * (1 - r) * n$$

where

- $r$  = the ratio as explained above,
- $n$  = number of vehicles between itself and the real vehicle closest and ahead of it (including itself),
- $m$  and  $l$  = coefficients  $> 0$  and  $< \alpha$  that represent merging and diverging behavior of vehicles.

Physically,  $r$  varies from 0 (when the real vehicle is at the beginning of the concerned zone) to 1 (when at the end of it).

To illustrate the principle, let a real vehicle be at a distance  $x$  from the beginning of the zone 2, which has a total length  $L$ . The ratio  $r$ , using the above definition, is  $x/L$ . The first imaginary vehicle behind the real vehicle will be placed at a distance:

$$HW_{imaginary} = (V_{real} * C + HW_{min}) * (x/L)^{m+1}$$

Assuming  $m = 1$ , (input by the user), the final expression is:

$$HW_{imaginary} = (V_{real} * C + HW_{min}) * x/L \quad (\text{Figure 1b})$$

**REPRESENTATION OF DIFFERENT "MERGE LOCATION CONFIGURATIONS"**

In the above illustration,  $m$  is considered to be equal to 1. The resulting term increases linearly with increase in  $x$ . Therefore, it could be perceived that, for example, at a location midway in zone 2 (between the work zone warning and the lane closure), only "half a vehicle" has merged into the open lane. This is the case for all the vehicles in the closed lane. Another way of perceiving the situation is to assume only half the total number of vehicles have managed to merge completely, while the other half are still in the closed lane at that location. Thus, the merging phenomena can be assumed to be occurring uniformly all along the zone, between the lane closure warning and the lane closure. Different merging characteristics, including early and late merging, can be represented by manipulating the value of  $m$  (Figure 2).

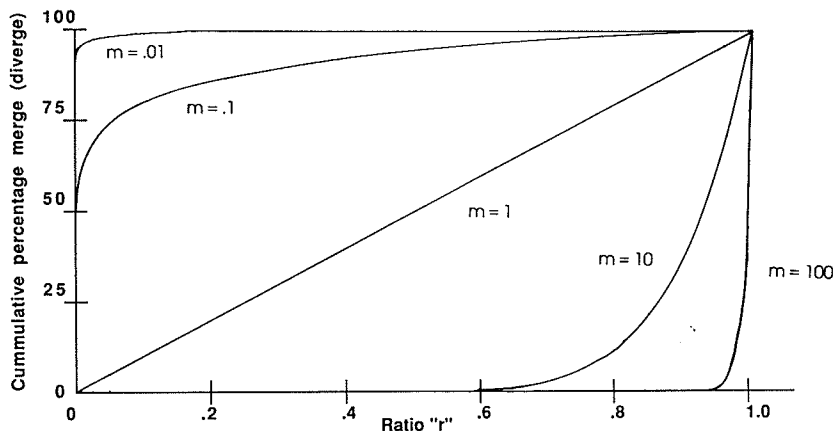


FIGURE 2 Relationship between ratio  $r$ , and the cumulative percentage of vehicles that have merged (or diverged) from the closed (open) lane.

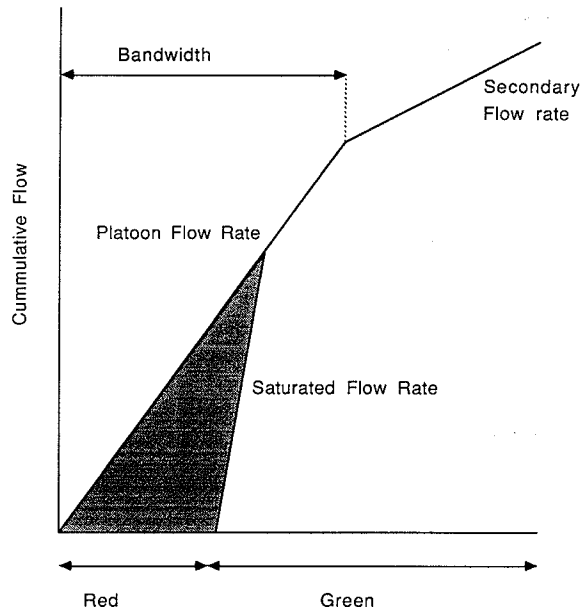


FIGURE 3 Modified continuum model to introduce two flow rates.

**INTERSECTION DELAY MACROSCOPIC MODEL**

The conventional continuum model was modified and a new model was developed to more realistically represent the characteristics of platoon flow. The fundamental concept of this model was borrowed from an idea originally developed by Rouphail (4). Instead of assuming a single average flow rate (as in uniform flow), two different flow rates were considered, one representing the average platoon flow rate and the other representing the secondary flow rate.

**ESTIMATION OF INTERSECTION DELAY AND RELATED PARAMETERS**

The macroscopic model uses a continuum model as depicted in Figure 3. Configurations like the one depicted in the figure can be defined by relationships between three attributes: signal (green time and cycle time), bandwidth, and offsets. It is not possible to establish a general formula to calculate the delay and other related parameters. The model is therefore split into several sub-models, and distinct formulas are derived for each of these sub-models. All parameters obtained from the semi-simulation model either serve as direct input or are used to estimate parameters required for input to the macroscopic model at the intersection. Formulas are also derived to actually introduce losses due to the reaction time and acceleration of vehicles after the start of the green. Detailed illustration of the formulas and method of estimation of parameters can be found elsewhere (2, 3).

**INPUT REQUIREMENTS AND MODEL APPLICATIONS**

The input requirements for execution of the program are divided into five sections. They are listed below with respect to a typical arterial system with a work zone (Figure 4).

1. Traffic Characteristics:
  - a. Initial velocity = 20 mph.
  - b. Desired velocity = 45.
  - c. Desired velocity in the work zone = 45.
  - d. Maximum allowable acceleration = 7 ft/sec<sup>2</sup>. This parameter is only used with respect to the first vehicle, to develop its velocity profile.
  - e. Maximum allowable deceleration = 21. This param-

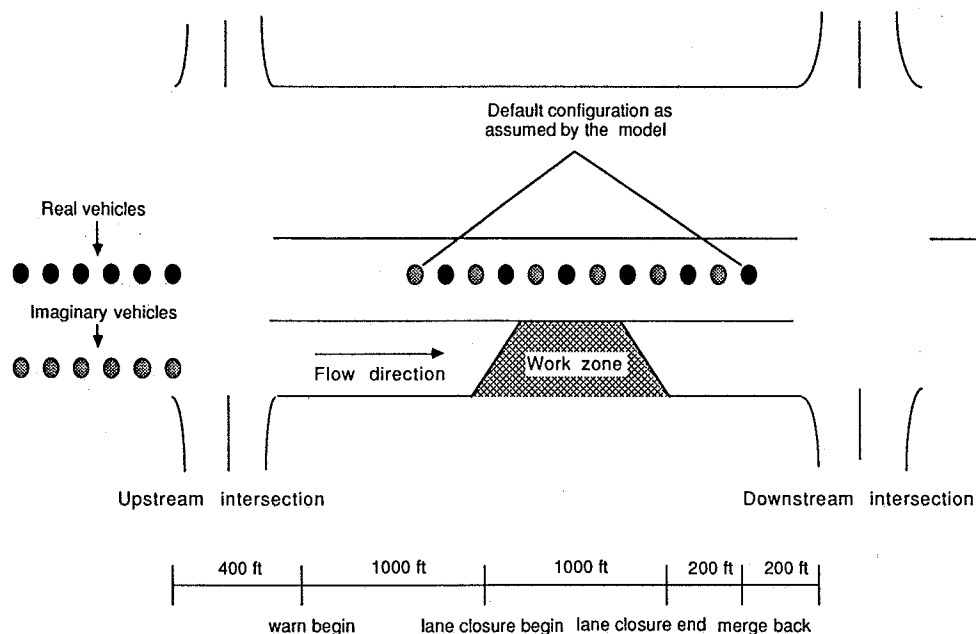


FIGURE 4 Typical arterial with a work zone.

eter is used with respect to all vehicles and is related to basic vehicle characteristics.

- f. Minimum headway = 16 ft. This parameter represents an allowable distance between the front or back end of two average vehicles, when stationary.
2. Intersection/Work Zone Characteristics:
- a. Lane closure warning location = 400 ft. This parameter defines a location with respect to the upstream intersection, from which point the vehicles are directly aware of the lane closure ahead. It should be selected with caution. The program considers this location as the beginning point of zone 2, where all merges originate.
  - b. Lane closure "begin" location = 1400. This parameter represents the location, with respect to the upstream intersection, of the beginning of the lane closure (zone 3).
  - c. Lane closure "end" location = 2400. This parameter represents the location of the end of the lane closure with respect to the upstream intersection (beginning of zone 4).
  - d. Post-lane closure merge-back completion location = 2600. This parameter represents the location, downstream of the lane closure, at the end of which all vehicles that had merged are assumed to have merged back into their respective lanes. This parameter can be manipulated to represent different merging behavior.
  - e. Total number of lanes = 2.
  - f. Number of closed lanes = 1.
  - g. Downstream intersection location = 2800.
  - h. Average grade = 0.
3. Signal Characteristics:
- a. Cycle time = 60 sec.
  - b. Green time = 30.
4. Platoon Characteristics:
- a. Saturation flow rate = 2000 veh/hr/lane (at the downstream intersection).
  - b. Secondary flow rate = 200 (non-platoon flow rate at the downstream intersection).
  - c. Number of vehicles per lane = 6 (in the platoon).

5. Options:

- a. C value outside the work zone = 1.45. This parameter represents a constant for car following [discussed in detail elsewhere (3)].
- b. C value in the work zone = 1.45. This parameter is used instead of the above only in the work zone. Caution should be used when selecting this parameter, since it would differ from the value used outside the work zone only for specific situations [discussed in detail elsewhere (3)].
- c. Constant for merging behavior = 1. This parameter is responsible for the merging behavior before the lane closure. All positive values are acceptable. (See Figure 2.)
- d. Constant for diverging behavior = 1. This parameter represents diverging or merging back of the "imaginary" vehicles into the original lanes, and operates similar to the parameter discussed above.
- e. Time step for simulation = 1 sec.
- f. Listing of simulation process = N.
- g. Optimize offsets = Y
- h. Listing of the optimization process = N.
- i. Direct run = N.

The program has an option of directly estimating intersection delay without going through the simulation process.

The results of a run with the above data are tabulated in column 1 of Table 2.

**SPECIAL APPLICATIONS**

**Analysis of Flow Without Work Zone**

Analysis of flow without the presence of a lane closure between the intersections can be conducted by placing all four parameters that define the work zone before the upstream intersection or after the downstream intersection. Since the upstream intersection is located at the reference point for all other distances (zero distance), attributing negative values to all the parameters that define the work zone would place it before this intersection.

TABLE 2 ILLUSTRATION OF MODEL OUTPUT FROM THE ANALYSIS OF DIFFERENT SCENARIOS AT THE SAME WORK ZONE

Configuration	RIRIRIRIRI	RRRRR	RRRRIIRRII	IIIRRIIRRR
Work zone present	Y	N	Y	Y
Platoon Flow Rate (veh/hr)	1731	2886	1990	1667
Delay Between Intersections (secs)	5.87	1.64	3.33	8.40
Delay at Intersection (sec)	5.64	5.96	5.80	5.50
Total Delay (sec)	11.51	7.59	9.13	13.90
Travel Time of Bandwidth(sec)	44.12	44.12	44.12	44.12
Bandwidth length (sec)	16.15	7.69	12.77	17.91
Optimized Offset Range (sec)	34.8 40.1	26.1 40.1	31.5 40.1	36.1 40.1
Total Length of Queue (cars)	0 *	0 *	0 *	0 *
Longest Time in Queue (sec)	0 *	0 *	0 *	0 *
Queue Dissipation Time (sec)	7.78	7.78	7.78	7.78

\* Only with respect to the platoons (zero because offset considered is optimized)

Four input parameters were modified in the above example to analyze conditions without the lane closure:

1. Lane closure warning location (ft) -2400,
2. Lane closure "begin" location -1400,
3. Lane closure "end" location -400, and
4. Post-lane closure merge completion location -200.

The results obtained after running the program are tabulated in column 2 of Table 2.

#### Analysis of Asymmetrical Flow

The model can be used to analyze a situation with asymmetrical or non-uniform merging, or that of traffic flow that displays distinct differences in characteristics in both lanes. To illustrate the procedure, consider the work zone described in the previous section. (See Figure 4.) For the sake of simplicity, assume that the number of vehicles in both lanes is equal and that only the merge configuration is asymmetrical. The first two cases illustrate flow when the vehicles are merging uniformly in such a way that flow characteristics upstream of the intersection are always symmetrical in both lanes, *i.e.*, if *R* represents vehicles in the open lane and *I* represents vehicles in the closed lane, then the configuration of vehicles in the open lane at the lane closure appears as follows:

*R I R I R I R I R I*

← (traffic direction)

This is the default condition assumed by the model. However, if this is not the case, the user can alter the configuration. Assume that the configuration at the lane closure after the merge is as follows:

*R R R R I I I R R I I I*

(traffic conditions not conducive to uniform merging of closed vehicles)

The results obtained when the program is run for this configuration are listed in column 3 of Table 2.

It is obvious that, for the above configuration, there would be asymmetrical flow characteristics just upstream of the intersection. The vehicles merging from the closed lane would certainly experience a greater delay than the vehicles that were already in the open lane. Since the model is only capable of representing the average delay of all the vehicles in the open lane (*R*), the aforementioned delay is only with respect to the vehicles in the open lane (the "real" vehicles). To estimate the average delay for the vehicles in the closed lane, the model is "tricked" by running it again—this time, with all the designations of the vehicles interchanged (easily done using the mirror command on the original configuration of the same data set). The new configuration is as follows:

*I I I I R R R I I R R R*

The results obtained after the run with this configuration are tabulated in column 4 of Table 2 and represent the average delay of vehicles in the closed lane.

In some cases, it may not be possible for vehicles both to be merged according to the configuration specified and, at the same time, to satisfy the car-following rules. Such a sit-

uation is flagged as a "crash" between vehicles, and the program is terminated. The user should then adjust the configuration to one that is more likely under the circumstances.

#### SUMMARY AND CONCLUSIONS

The model developed was found to be a very flexible tool; and when calibrated accurately, it can be used as an effective tool to analyze almost any situation of a work zone in an arterial. Analysis of work zones with asymmetrical flows in lanes or with no work zone at all can also be conducted using simple manipulative procedures in running the program. Work is at present being conducted to implement graphical facilities into the model. Preliminary validation of the model indicated general agreement with the commonly witnessed traffic behavior in work zones.

#### Model Advantages

Several advantages of the model developed are apparent from the initial runs and validation, as follows:

1. Since vehicle travel is actually simulated between the two intersections, direct effects of various parameters—like distance of the warning to the lane closure, location of the lane closure in between the intersections, length of the lane closure, initial speed of vehicles, flow rate at the lane closure, characteristics of the merging phenomena, and offsets between intersections on delay, among others—can be studied.
2. The user has complete control of vehicle merging. He or she actually inputs the likely configuration of vehicles at the single open lane of the lane closure.
3. Unlike conventional simulation models, where merges are based on the philosophy of gap acceptance, this model actually "forces" gaps between vehicles in the open lane for vehicles in the closed lane to complete their merge procedures. This is based on the generally observed fact that, in arterials, where the speeds are not as high as on freeways, merges commonly occur for two reasons: (a) Vehicles from the closed lane "force" themselves into the open lane, assuming that the following driver would automatically reduce his/her speed and adjust the headway to accommodate the merging vehicle; and (b) drivers in the open lane reduce their speed prior to the actual merge phenomena, thereby providing acceptable gaps for drivers from the closed lane to complete their merge.
4. The user has complete control over the merging locations of the vehicles. By altering a single parameter in the input data, a whole new vehicle merge location configuration can be represented.
5. The macroscopic model is accurate in principle since it is developed along the lines of continuum models, which are based directly on first principles. Additional accuracy is obtained by representing flow at the intersection as a composite of two flows and rather than, as in conventional continuum models, one.
6. The model provides an option to analyze the system for an unlimited set of offsets. Since there is no relation between the offset and the travel time delay between the intersections,

only the macroscopic section of the program is to be executed to test for different offsets. The analysis of different offsets is therefore rapid and simple. The model also provides an option to estimate the offset range wherein the delay at the intersection is the least (optimized offset).

7. Special techniques are used to include the deceleration, acceleration, and start losses in the overall estimation of delay at the intersection. Thus, a major disadvantage of conventional continuum models is removed.

8. The model is completely menu-driven and the user has total control over the manipulation of parameters and the execution of the program at all times. Further, there are the advantages associated with microcomputer programs over the main-frame counterparts.

### Model Shortcomings

The shortcomings associated with the model in general are listed below.

1. All vehicles are initiated at a constant uniform velocity and a constant headway. No initial distributions of either headway or velocities can be analyzed.

2. No merging phenomena are allowed to occur anywhere outside either the lane closure warning distance or the post-lane closure merge distance.

3. The whole phenomenon of merge initiation is user-dependent and not model-generated. Thus, the user should obtain some prior information on the flow characteristics in the system. When more information and data are available in this respect, it could be possible to perceive patterns and form regression equations; these regression equations might then be directly implemented in the model.

4. Delay between the intersection is not estimated for non-platoon flows. All vehicles constituting the "non-platoon" are considered to travel through the work zone without any delay. Delay at the intersection is, however, estimated for both platoon and non-platoon flow.

5. The macroscopic model is only capable of analyzing pre-timed signal control systems.

6. The macroscopic model is only capable of analyzing symmetrical flows at intersections. If flow as estimated by the simulation of vehicles before the intersection is asymmetrical, then an average flow is estimated and considered for further analysis.

7. The macroscopic model is not capable of analyzing over-saturated conditions, because the delays associated with over-saturated conditions are never constant and increase with the cycle length.

### REFERENCES

1. J. L. Memmott and C. L. Dudek. Queue and User Cost Evaluation of Work Zones (QUEWZ). *Transportation Research Record 979*, TRB, National Research Council, 1984, pp. 12-19.
2. N. M. Roupail. *A Model of Traffic Flow at Freeway Construction Lane Closures*. Ph.D. dissertation. Ohio State University, 1981.
3. C. T. Joseph. *Model for the Analysis of Work Zones in Arterials*. M.S. thesis. Arizona State University, 1987.
4. N. M. Roupail. Delay Models for Mixed Platoon and Secondary Flows. *ASCE Journal of Transportation Engineering*, in press.

---

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

# Application of the Image Analysis Technique to Detect Left-Turning Vehicles at Intersections

YEAN-JYE LU, YUEN-HUNG HSU, AND GUAN C. TAN

**This study applies the image analysis technique to detect left-turning vehicles at intersections. The problems existing in the current method of left-turn data collection are discussed. An optical image device system including a CCD video camera, an interface board, an image monitor, and an IBM PC/AT was used. An algorithm with linear time complexity was developed for detecting left-turning movements at intersections. Microcomputer software was derived for the algorithm. This computer system is a real time system. The accuracy of the algorithm is about 80% for detecting the left-turning vehicles whose signal lights are on. This study so far has had only limited success. The difficulties of attaining greater accuracy as well as the possibilities of improving the developed image analysis system are also discussed.**

The goal of a traffic system is to provide rapid, economical, and safe movement of vehicles and pedestrians in a city or in an urban area. Since a traffic system works as a circulation system, its operating efficiency will significantly affect the city's growth and economic development. However, intersections probably represent the most critical elements of the traffic system. Since two or more streets share the space at an intersection, the capacity of the intersection is generally limited. One of the main factors which affects the capacity of such an intersection is the presence of left-turning vehicles. Thus, traffic problems such as traffic congestion, fuel consumption, air pollution, and noise pollution have usually occurred at these intersections, particularly when heavy left-turn traffic was present.

Left-turning vehicles cause fewer problems at an intersection under low-volume conditions. As the traffic volume of an intersection approaches capacity, however, fewer opportunities for left-turning maneuvers exist. Both the left-turning vehicles and the non-turning vehicles which queue frequently behind the turning vehicles suffer long delays before clearing the intersection. Thus, drivers sometimes become impatient and make hazardous maneuvers after experiencing long delays at an intersection.

Before making improvements to the traffic system, traffic engineers must have adequate and accurate traffic data. Traffic data can be divided into three types: traffic volumes, vehicle speeds, and intersection delays. These traffic data are impor-

tant in practice because they may indicate locations where improvements are needed. In addition, these data are used in before-and-after studies to determine the effectiveness of changes in parking prohibitions, signal timing, one-way streets, or turning prohibitions. Furthermore, the left-turn volume during different hours of the day is one of the major criteria for determining these changes. This is because the left-turn volume significantly affects vehicle speeds and intersection delays.

## NEED FOR RESEARCH

Basically, there are two methods for gathering traffic data: manual and automatic. The manual method is labor-intensive. This method requires high labor costs for data collection and data processing, particularly when massive quantities of traffic data are required. In addition, the data collected by an observer is sometimes inaccurate because it is difficult for an observer to keep an eye on a street scene for hours. However, the manual method allows data to be collected on the basis of turning movements and vehicle classifications. On the other hand, the automatic method uses automatic recording devices for gathering field data and further processes these data by using microcomputers. The automatic recording devices include pressure-type road tubes, electrical contact tapes, photo-electric detectors and inductive loop detectors (1). The automatic method is able to continuously record and process traffic data for long periods of time. This method is generally more economical and accurate than the manual method.

Today, there is a real need among traffic engineers, not only for more data, but also for data of a more complex nature. Although the automatic method for gathering traffic data has been widely used in major cities, this method has problems in providing left-turn traffic data. These problems are discussed below.

1. No Turning Movement Data: The automatic recording devices are incapable of determining turning movements except for the intersections equipped with exclusive turning lanes. However, the turning movement data are especially significant in design, channelization, lane marking, signal timing, and the application of control devices.

2. Wheel-Oriented Data: The automatic recording devices can only detect the number of wheels passing over the detection devices. Thus, the automatic method provides the num-

Y. J. Lu, Department of Civil Engineering, Concordia University, 1455 DeMaisonneuve Boulevard West, Montreal, Quebec, H3G 1M8, Canada. Y. H. Hsu, Department of Computer Science, McGill University, Montreal, Canada. G. C. Tan, CAE System Division, Tektronix, Sunnyvale, Calif.

ber of wheels for the traffic volume data, and the number of wheel-seconds for the traffic delay data. This type of traffic data may not be suitable for traffic network systems analysis.

This study has reviewed some published literature related to the topics on traffic data collection and the application of image analysis to traffic engineering. The results of the literature review indicate that in the last decade a variety of systems have been developed specifically for traffic data collection and/or monitoring. Dickinson and Waterfall (2) provided an excellent review of these systems. They covered the general subject of image processing and the analysis of traffic scenes. They concluded that the hardware and software of the video image system have to be improved in order to provide a general-purpose traffic data collection system.

Branston (3) used a time-lapse photography method to collect traffic data on speeds and headways. Garner and Uren (4) used a coordinate reader and an electronic computer to reduce the time of data collection and analysis for the aerial photographic method. Mountain and Graner (5) provided a review of the types of photography and traffic data that may be collected by using the aerial photographic method. In addition, Ashworth (6) and Polus, et al. (7) described a video recording system for measuring traffic data such as speeds, occupancies, and volumes. Wootton and Potter (8) used video cameras to record traffic movements and the recording was linked to a TV monitor and a microcomputer. Ashworth and Kentros (9) used an ultrasonic detector to measure vehicle occupancy.

Since the mid-1970s, the U.S. Department of Transportation has been funding research on image processing applied to freeway surveillance at the Jet Propulsion Laboratory (JPL) in Pasadena. Hilbert, et al. (10) in 1978 described the conceptual design for the wide-area detection system (WADS). The major objective of this system is to track individual vehicles within the scene. In order to track the vehicle, they used the cross-correlation method to produce "best-fit" locations for the vehicle in each frame. Thus, traffic data such as speed, acceleration, and lane changing can be estimated. Schlusmeyer, et al. (11) in 1982 described several practical problems with the cross-correlation method for vehicle tracking. These problems include the high computational cost, the difficulties of tracking a vehicle if its gray values changed, and the low accuracy for distant vehicles. Thus, JPL engineers reverted to a simpler approach using a single lane across the traffic lanes and monitoring it to detect any vehicle entering the detection area close to the camera. Then the leading and trailing edge of the vehicle image are located in two frames and the vehicle velocity is calculated subsequently.

Dickinson and Waterfall (12) described the development of a multi-microprocessor system for preprocessing video images of traffic scenes. Moreover, Houghton, et al. (13) showed that, with suitable reduction of image data and with appropriate feature extraction, several vehicles can be tracked concurrently on a highway network.

In conclusion, the results of the literature review indicate that a variety of systems have been developed for measuring traffic data such as volumes, speeds, headways, lane occupancies, and junction turning counts. However, the detection of left-turning vehicles at intersections is needed.

## OBJECTIVES AND SCOPE

Image analysis is one of the subjects in the area of computer vision. An image is a two-dimensional array of pixels, obtained by a sensing device which records the value of an image feature at all points. The goal of image analysis is the construction of scene descriptions on the basis of information extracted from images or image sequences (14). Over the past two decades, many techniques for analyzing images have been developed, and this subject has gradually begun to develop on a scientific basis. The main applications of image analysis include document processing, microscopy, radiology, industrial automation, remote sensing, and reconnaissance.

This study intends to determine the feasibility of applying the image analysis technique to solving the problem of detecting left-turning movements. This study also intends to provide engineers with a better understanding of increasing traffic complexity. The specific objectives of this study are:

1. To identify problems and needs associated with the development of automatic traffic data collection and analysis,
2. To determine the suitability of applying the image analysis technique in the area of traffic engineering,
3. To develop an algorithm and a computer program for detecting left-turning vehicles, and
4. To investigate the real time possibility for continuously analyzing images of left-turning movements.

This study does not intend to completely solve the two problems listed previously. Rather, this is a preliminary study to determine not only the suitability of applying the image analysis technique but also the possibility of real time analysis for solving traffic problems. Thus, this study emphasizes the left-turning movements during the daytime only. The left-turning movements at night, as well as the right-turning and through movements, are not included in this study.

## METHODOLOGY

### Optical Image Device System

In order to conduct this research, an optical image device system was chosen. Figure 1 depicts the structure of the four major items which are included in this image system. These four items are listed below:

1. CCD Video Camera: CCD stands for charge coupled device. The purpose of a CCD video camera for this study is to convert the light energy when taking pictures from the street scene into analog signals.
2. Interface Board: As shown in Figure 1, the interface board is not only an analog-to-digital converter but also a digital-to-analog converter. The interface board converts the analog signals taken from the CCD camera into digital signals which are sent to a computer. In addition, the interface board converts the digital signals stored in the display memory into analog signals to be displayed on an image monitor. A Truevision Advanced Raster Graphics Adapter 8 (TARGA 8) board from AT&T (15) was selected for this study.
3. Image Monitor: Through the interface board an image monitor displays not only the live images taken from a CCD

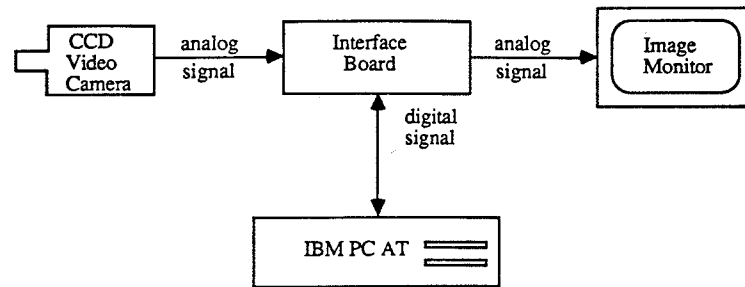


FIGURE 1 The structure of an optical image device system.

video camera but also the images stored in the computer memory. Thus, engineers can visualize the street traffic from the image monitor and make any adjustment if necessary. A black-and-white TV was chosen to be the image monitor for this study.

4. IBM PC/AT Computer: The purpose of an IBM PC/AT or equivalent is to store and/or analyze the digital data of images sent through the interface board. The PC/AT can also send the images stored in the computer memory through the interface board to be displayed on the image monitor. The reasons for selecting a PC/AT are its computing speed and capacity of its hard disk as well as its compatibility to the selected interface board.

#### Signal Lights of Left-Turning Vehicles

The driver's manuals in Canada state that a driver "should signal his intentions continuously and for a sufficient distance before making a turn" (16). The purpose of signaling before making a turn, for drivers in Canada, is the same as in other countries—safety. Thus, the left-turning vehicles which do not signal will be ignored in this study due to the limitations of the proposed method discussed below.

When pictures are taken of left-turning vehicles at intersections during the daytime, the signal lights of left-turning vehicles will likely be the brightest spots on the pictures. This study intends to search for these signal lights of left-turning vehicles in order to detect the left-turning movements. In other words, this study will utilize the image analysis technique to extract the left-turning lights from the pictures. However, "white noise," such as the reflection of vehicles from the sun, makes the search for signal lights difficult to obtain accurately.

Images are built up from individual dots, called picture elements or pixels. The display resolution is defined by the number of the dots in the picture, *i.e.*, by the rows or scan lines from top to bottom and the number of pixels from left to right in each line. The number of rows and the number of pixels in an image are determined by the computer capacity and the interface board.

As stated previously, a TARGA 8 board from AT&T was chosen for this study to digitize and store images. The TARGA 8 board supports an enhanced spatial resolution of up to 512\*482 pixels with 256 gray levels (8 bits per pixel), and captures images in real time: 1/30 second per frame. The 256 gray levels range from 0 to 255. The gray value of 0 indicates the dimmest level in an image. When the gray value increases, the brightness level of pixels increases. Thus, the

gray value of 255 represents the brightest level in the image. Therefore, this study will extract the spots or clusters in which these pixels have high values of gray levels. For instance, this study may search for the cluster in which most of the pixels have gray values of or above 240.

Figure 2 shows a black-and-white photo of left-turning movements at an intersection. In this picture, there are three left-turning cars whose signal lights are on. This example is to demonstrate what the pixels in an image are and what the gray values of the pixels are. Thus, this photo was digitized by using an optical scanner. The outputs from the scanner are gray levels of the pixels in the image. In order to obtain a hard copy from the computer printer, there are only 16 gray values from the scanner instead of 256 gray values from a TARGA 8 board. These 16 gray values range from 0 to 9, plus A to F. However, in order for human eyes to visualize the objects in the image, the sequence of gray levels has to be reversed. In other words, in this sample, the gray values 0 and F represent the brightest and the dimmest light levels, respectively. It is noticed that, in the TARGA 8 board, the gray values 0 and 255 represent the dimmest and the brightest light levels.

Figure 3 illustrates part of the output from the scanner. This figure includes the first two cars only. The boundaries of the two cars, the background of the scene, and the left-turn lights of both cars are delineated by a black pen. In this figure, the gray values of the hood and side doors for the white car are 0 and for the dark-blue car are D or E. On the other hand, the gray values for the shadows underneath these two cars are E and those for the pavement surface range from



FIGURE 2 An example of left-turning movements at an intersection.



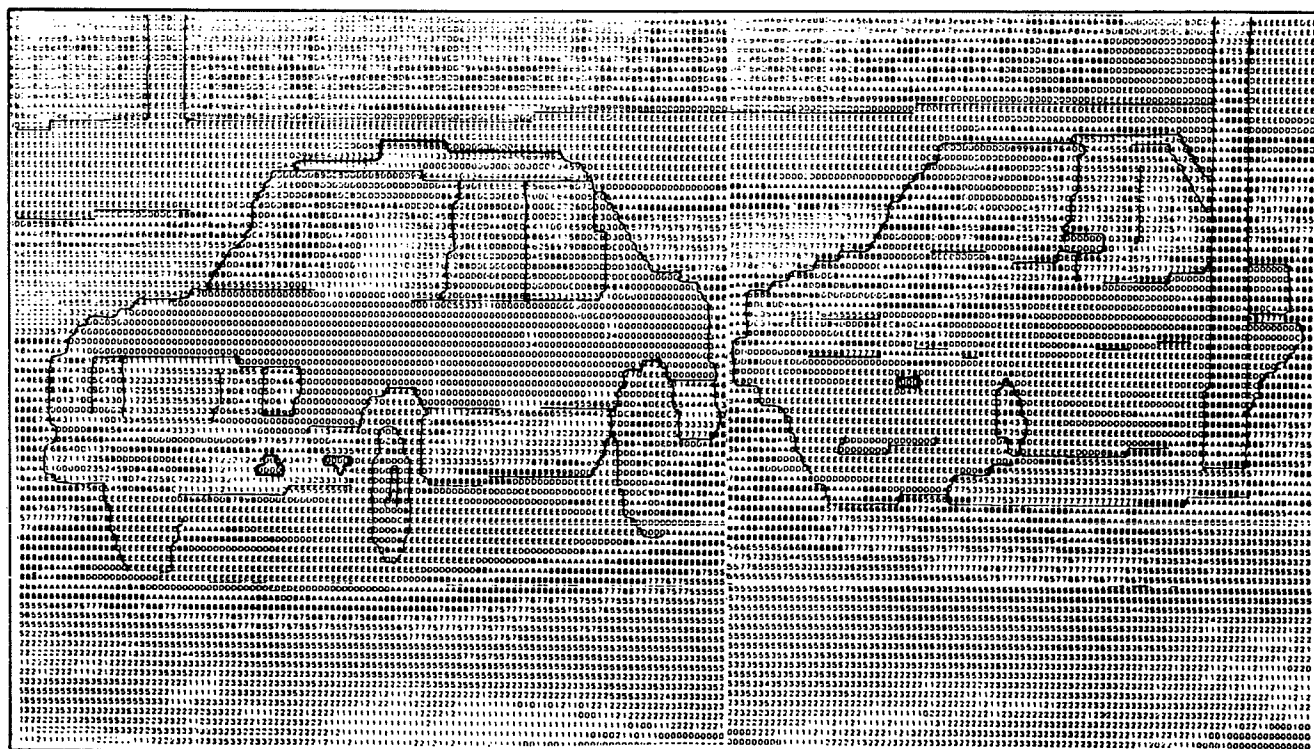


FIGURE 3 The gray values of two left-turning cars.

1 to B. Thus, it is feasible to group an image into several regions which share some values of a feature. Moreover, the gray values of the turning lights for both cars are 0. In addition, the neighborhood of the turning lights has gray values ranging from 1 to 8 for the white car and from 4 to D for the dark-blue car. Thus, the turning lights can be detected by finding bright groups of pixels and the abrupt discontinuities of such an image feature.

**Algorithm**

This study develops an algorithm to determine the number of left-turning vehicles by detecting their turning signals at intersections. Figure 4 depicts the flow chart of the algorithm. This algorithm is divided into three stages, which are discussed below.

*Stage 1: Environment Setting*

This stage of the algorithm fetches the relevant portion of traffic images into the RAM (Random Access Memory) of an IBM PC/AT. This stage also predetermines some factors of the environment. The four steps in this stage are as follows:

- Step 1: Fetch images from the camera to the TARGA 8 board. In this step, the TARGA 8 board is set in the live mode in order to continuously grab the images taken from a CCD video camera into the display memory on the board.
- Step 2: Fetch the relevant parts of the images into the RAM of a PC/AT computer. Here, the relevant parts of the images on the display memory are fetched into the RAM of the PC/

AT computer by discarding most of the images' background. The relevant part is the image portion containing the left turning vehicles and their surrounding background. This step significantly reduces the size of the traffic images.

Step 3: Determine the threshold of the bright pixels. In order to extract the pixels with high gray value, traffic engineers have to determine the threshold of the bright pixels, so that the pixels with a low gray value can be masked out in the next stage. From the results of this study, a gray value greater than or equal to 240 should be chosen as the threshold of the bright pixels.

Step 4: Determine the acceptable length of the bright segments. A bright segment is a group of consecutive pixels located in a line. In order to filter out white noise (*i.e.*, tiny bright spots in the image), traffic engineers predetermine the acceptable length of the bright segments. If the length of a bright segment is less than acceptable, then the bright segment is white noise and will be discarded. The acceptable length of the bright segments is the minimum length of the signal lights in the images. The acceptable length mainly depends upon the resolution of the images and the distance from the camera to the left-turning vehicles. Thus, engineers can output the gray values of pixels for several images at the beginning of the data collection. In this way, the length of the signal lights can be estimated.

*Stage 2: Image Processing*

During this stage, the pixels of the images obtained in stage 1 are scanned in order to locate all bright segments. In the meantime, any new segment will either create a new bright cluster or update an existing cluster. This stage principally

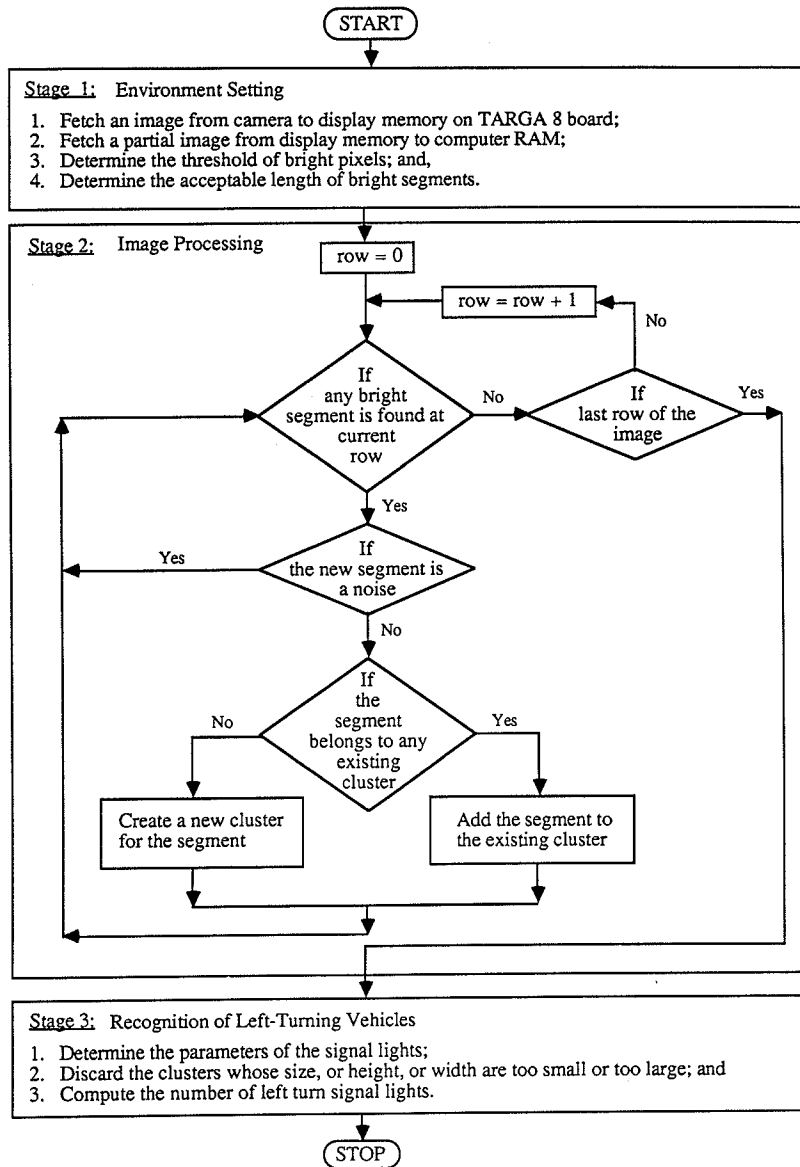


FIGURE 4 The flow chart of the algorithm.

processes the images in order to search for bright clusters, which are likely to include signal lights and other bright objects such as the reflection of vehicles. Four major steps are included in this stage and discussed below:

**Step 1:** If any bright segment is found at the current scanned row of the image, proceed to step 2. Otherwise, continue to scan pixel by pixel until the end of the row. A bright segment is a group of bright pixels which are next horizontally adjacent in a row of the image. When the present row is scanned pixel by pixel, and if a bright pixel is found, this pixel is considered as the left boundary point of a bright segment. Then, the algorithm searches for the right boundary point of the bright segment. A pixel is considered as a right boundary point if the following two pixels are not bright.

**Step 2:** This is to be undertaken if the new segment is noise. The length of a new bright segment is equivalent to its number of pixels. Because the two boundary points of the segment are known, the length of the segment can easily be computed.

In this step, if the length of the segment is less than acceptable, the segment is considered to be white noise and will be discarded; go back to step 1. Otherwise, proceed to the next step.

**Step 3:** After a bright segment is confirmed in step 2, step 3 examines the possibility that the segment may belong to an existing cluster. If so, the segment will be added to that existing cluster. Otherwise, a new cluster will be created for this segment.

Figure 5 illustrates the algorithm for determining whether the cluster belongs to any existing cluster. This algorithm does not search every existing cluster for making such a determination. Instead, it ascertains whether the pixels corresponding to this bright segment belong to any existing cluster. In Figure 3, the solid curved line is the boundary for the existing cluster *I*. The boundary is a curved line connecting all boundary points of bright segments. The dashed line is the current scan line, and pixels *a* and *b* are the two boundary points of this

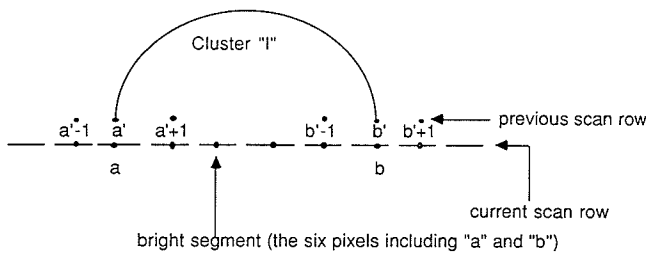


FIGURE 5 The algorithm for checking an existing cluster.

new segment. In addition, pixels  $a'$  and  $b'$  correspond to  $a$  and  $b$ , respectively. Pixels  $a' - 1$  and  $a' + 1$  are the two pixels adjacent to pixel  $a'$ , as are pixels  $b' - 1$  and  $b' + 1$  adjacent to pixel  $b'$ .

Since the images obtained by the TARGA 8 board are high resolution, with about 20K pixels per image, any bright object in the image will have a smooth boundary line. Therefore, this algorithm checks if any of pixels  $a' - 1$ ,  $a'$ , and  $a' + 1$ , or any of pixels  $b' - 1$ ,  $b'$  and  $b' + 1$  belong to an existing cluster. If so, the new segment belongs to the same cluster. If not, a new cluster will be created for the new segment. For instance, in Figure 3, pixels  $a'$ ,  $a' + 1$ ,  $b' - 1$ , and  $b'$  belong to cluster  $I$ . Thus, the new cluster, with the two boundary pixels  $a$  and  $b$ , belongs to cluster  $I$ .

Step 4: This step checks if the current scanned line is the last row of the image. If so, proceed to stage 3: this means that the current image has been completely processed. Otherwise, scan the next row of the image and go back to step 1 of this stage.

### Stage 3: Recognition of Left-Turning Vehicles

This stage sets up three parameters in order to search the obtained clusters for recognition of the left-turn signal lights of vehicles. This stage includes three steps, as follows:

Step 1: Determine the parameters of the signal lights. The bright clusters obtained from stage 2 include left-turn lights and other bright objects such as the reflection of vehicles from the sun. This step sets up three sensitive parameters for every cluster in order to distinguish left-turn lights from the other bright objects. These three parameters are size, width, and height of a cluster. The size of a cluster is the number of pixels contained in that cluster; the width of a cluster is the number of pixels for the longest row in that cluster; and the height of a cluster is the number of pixels of the longest column in that cluster. However, the results of this study indicate that the parameter of size is the most sensitive among the three parameters.

Step 2: Discard the clusters whose values of parameters are not within the acceptable range. Traffic engineers have to predetermine the acceptable range for each parameter, which depends on the resolution of the images and the distance from the camera to left-turning vehicles. In order to predetermine the acceptable range, engineers can output the gray values of several images of traffic scenes. The size, width, and height of all signal lights are counted. Thus, the acceptable range for each parameter is between the lower bound and upper bound of each parameter. This step checks the parameters of

each cluster to determine if the values of the three parameters fall into the three acceptable ranges, respectively. Since the size parameter is the most sensitive, the sequence for checking is (a) size, (b) width, and (c) height. So, if the size of a cluster is out of the acceptable range, the cluster will be discarded immediately, and it will not be necessary to check the other two parameters.

Step 3: Compute the number of left-turn lights. After Step 2 is completed, the clusters remaining are left-turn lights. This step counts the number of clusters for estimating the number of left-turning vehicles.

### Time Complexity of the Algorithm

Intuitively, the time complexity of an algorithm is the required running time of the algorithm when solving a problem. When an algorithm manipulates data, the manipulation can be broken into many elementary operations or groups of elementary operations, such as comparisons, additions, and multiplications. Thus, the time complexity of the algorithm is defined as the total number of operations for processing input data and producing output information when solving the problem.

Let  $T(n)$  denote the time complexity of an algorithm where  $n$  is the size of the input data of a problem. This is because  $T(n)$  is usually a function of  $n$ . The big- $O$  notation is also used to describe the relationship between  $T(n)$  and  $n$ . This relationship is a good indicator in evaluating the effectiveness and the power of an algorithm. For instance, if an algorithm with  $T(n) = O(n^2)$ , this indicates that the upper bound of the computer time increases as a square when the size of the problem increases. If an algorithm with  $T(n) = O(n^3)$ , it means that the upper bound of the computer time increases cubically when the value of  $n$  increases. Obviously, the algorithm with  $T(n) = O(n^2)$  is superior to the algorithm with  $T(n) = O(n^3)$ . This is because the former algorithm requires less computer time than the latter when  $n$  is sufficiently large. Furthermore, the latter algorithm becomes extremely expensive, even computationally unfeasible, (compared to the former algorithm) when solving a huge problem.

The purpose of this section is to prove that the

Time Complexity of the Proposed Algorithm =  $O(n)$

where  $n$  is the size of an image. The size of an image is represented by the number of pixels in this study. The time complexity of the algorithm is analyzed according to the three stages discussed below.

Stage 1: Environment Setting. In this stage, the algorithm digitizes the image from the camera, stores the digitized image in the display memory of the TARGA 8 board, and then grabs the relevant part of the image from the display memory to the computer RAM. The required computer running time in this stage is linear (Note: This statement has been verbally confirmed by AT&T.). Thus, the time complexity of stage 1 of the algorithm is  $O(n)$ , when  $n$  is the number of pixels of the image.

Stage 2: Image Processing. This stage horizontally scans each row of the image pixel by pixel to locate all the bright segments on the current row. If a bright segment is found, the algorithm makes one comparison to determine if the segment is white noise. If the segment is not white noise, the

algorithm makes at most six comparisons to determine if the segment belongs to any existing cluster. Therefore, the time complexity of this stage is  $O(n)$ .

Stage 3: Recognition of Left-Turning Vehicles. Let  $m$  denote the number of bright clusters. In this stage, the algorithm needs at most three comparisons for each cluster to determine if the cluster is a left-turn light. These three comparison are the three parameters (i.e., size, width, and height of the cluster). Thus, the time complexity of this stage is  $O(m)$ . However,  $m < n$  is true. Therefore, the time complexity of this stage is also  $O(n)$ .

In conclusion, the time complexity of the proposed algorithm is  $O(n)$  where  $n$  is the number of pixels per image. This indicates that the upper bound of the computer time increases proportionally to  $n$  when the size of the images increases. This also indicates that the algorithm is efficient and powerful.

## EXPERIMENTAL DESIGN AND ANALYSIS

### "Left-Turn Detection" Software

Using the derived algorithm, this study has developed micro-computer software, named "Left-Turn Detection." The main objective of this software is to detect left-turning vehicles at intersections. This software uses a DOS 3.1 operating system for IBM PC/AT computers. This software was written in the C computer language. The C language is suitable for this study because it can manipulate bits and is compatible with the TARGA supporting software (17) provided by AT&T.

Left-Turn Detection software utilizes a C86 compiler from Computer Innovations Inc. Because an image file can take up to 256K-byte memory space, the default value for stack and heap is changed to 320K-bytes. In order to speed up the program, this study takes advantage of the 80286 microprocessor by switching the compiler option to 80286 (the program will only run on 80286 microprocessors). Furthermore, the running time of this software is linear, i.e., the running time is proportional to the size of the input image file. This is the best feature of the program, that it conforms to a real time computer system.

### Real-Time Analysis

The application requires a real time system. This means the response time of the computer system has to be tied to the time scale of events occurring outside the computer. The computer must be able to accumulate, process, and output data within a critical, specified time period. In order to accomplish the real time analysis, this system requires guaranteed response from each part of the computer system—peripherals, processor, operating system, and the users' program—since the time headway of left-turn vehicles approaching an intersection is about 2 seconds per left-turn lane during peak hours. Thus, if the computer system can capture, process, and output the result within one second, the system is a true real time system for this study.

The TARGA 8 board takes up to just 1/30 second to capture an image from a video camera to the display memory of the board. The required computer time is small because of

the memory arrangement known as Row Addressable RAM (RARAM): a scheme that enables a row of 4000 bits to be moved into the memory at once, instead of 1 bit or 1 byte at a time. By using the same memory management, the image is transferred from the display memory of the TARGA 8 board to the memory of the IBM AT.

The image in RAM is processed by the algorithm. As discussed previously, the complexity of the algorithm is linear, i.e., if the image has  $n$  bytes, then the running time is  $O(n)$ . The second stage of the algorithm has to go through the memory location byte by byte to process the image. This processing takes longer than the transferring. For example, if the algorithm processes 20K bytes, then for every byte the memory access time would be around  $5 \times 10^{-5}$  seconds and the total computer time would be about 1 second.

In order to investigate the possibility of real time analysis, left-turning movements were tape-recorded by using a CCD video camera at the intersection of Sherbrooke and St. Mathieu Streets in downtown Montreal, Quebec. Thirty images of left-turn movements from different cycles of traffic signal timing were selected randomly for conducting this experiment.

The conditions of the experiment were as follows:

1. The images are black-and-white, and the gray level of each pixel ranges from 0 to 225;
2. After the preprocessing procedure in stage 1 of the algorithm, each image has 80 rows and every row has 256 pixels (i.e., each image has about 20K pixels);
3. The microprocessor used is Intel 80286; and
4. The threshold of gray level to distinguish bright pixels from the images is 240.

These thirty images were input into the Left-Turn Detection Software for detecting left-turning movements. The average running time for these thirty samples was 0.85 seconds. The standard deviation for this average value is 0.09 seconds. Thus, the algorithm is able to complete the analysis of an image within 1 second. Furthermore, the time for capturing an image and sending it to the RAM of the IBM PC/AT is about 0.05 seconds per image. This computer time is controlled by the TARGA 8 board. Hence, the total average computer time for each image is about 0.90 seconds. It is still less than 1 second. Therefore, the computer system developed is a real time system.

One second of computer time is a desirable amount of time to analyze an image on the basis of real time study for the intersection with one left-turn lane at each approach. Thus, a fraction of 1 second may be the time limit for the intersections with two exclusive or mixed left-turn lanes at each approach. Since the Intel 80386 processor is about two to three times faster than the Intel 80286 processor, by using the 80386 processor a real time computer system for this study can still be achieved.

### Accuracy of the Algorithm

The same thirty images of left-turning movement were also chosen to determine the accuracy of the algorithm in detecting left-turning vehicles at intersections. The set of samples was taken from a mixed left-turn and through lane at a busy intersection during peak hours on sunny weekdays. These thirty

images showed the different combination of cases in vehicle types, vehicle makers, number of vehicles in the queue, and percentage of left-turning vehicles in the queue. These images were input to the Left-Turn Detection software. The results of the computer output were compared with the results of human observation on an image monitor. The comparison indicates that the accuracy of the algorithm for these thirty images was about 80% for detecting the left-turning vehicles whose signal lights were on. Since not every driver uses turning signal lights, the reliability of the algorithm is further reduced.

The 80% accuracy indicates that the developed image analysis system has not reached the implementable stage yet. Improvements in both hardware and software are needed for the developed system. However, the difficulties of attaining greater accuracy are discussed below:

1. **Reflection of Small Objects:** The first stage of the algorithm has discarded most of the background, such as buildings and sidewalks, as well as the through vehicles in the adjacent lanes from the image. Thus, there are about 20K pixels remaining in each image in the second stage of the algorithm. These partial images still contain many objects. On sunny days, objects such as headlights, bumpers, windshields, windows, roofs, frames, wheel covers, and door handles on the vehicles all reflect sunlight into the camera. Thus, the bright clusters in the second stage of the algorithm include signal lights of vehicles plus all kinds of reflection. The third stage of the algorithm intends to distinguish the reflections from the signal lights by examining the size, width, and height of the cluster. However, it is extremely difficult for the algorithm to delete the reflections with sizes and shapes similar to those of the signal lights.

2. **Variation of the Size of Signal Lights:** There are two factors affecting the variation of the size of the vehicle signal lights. They are the size of the vehicle itself and the distance of the vehicle light from the camera. The size of the signal lights varies according to vehicle size and manufacturer. In general, large cars have larger signal lights than small cars, and European cars have slightly larger lights than Japanese and domestic cars. In addition, trucks, buses, and trailers have different sizes of signal lights among themselves and in contrast to other passenger cars. Furthermore, the longer the distance of a vehicle from the camera, the smaller the size of the lights present in the image. In other words, the queue position of the vehicles significantly affects the size of the signal lights shown in the image. Due to the wide variation in sizes, it is difficult for the algorithm to recognize the signal lights accurately.

3. **Blinking of the Lights When Turned On:** From observation, signal lights of vehicles blink from fifteen to twenty times every 10 seconds when the lights are on. The blinking of lights affects not only the size of the lights but also the gray values of pixels of the lights. The effect of blinking has made it even more difficult to detect the left-turn movements.

The possibilities for improving the developed image analysis system are as follows:

1. **Hardware Aspect:** A color interface board such as TARGA 16 will be recommended. TARGA 16 requires 2 bytes to present every pixel. For every pixel in TARGA 16,

5 bits are assigned to each of the primary colors (*i.e.*, red, green, and blue). Thus, TARGA 16 can display about 32,000 different colors. Since the color of signal lights of vehicles is yellow, it would be a great incentive in solving the problems listed above to examine the color of bright clusters. However, the disadvantage of using TARGA 16 is that the size of the images stored in the computer is larger than that in the TARGA 8, and thus more computer time is required.

2. **Software Aspect:** The third stage of the algorithm has to be modified. To solve the problem of the blinking lights, two or three consecutive images will be compared to determine the number of left-turning vehicles. Furthermore, a new parameter in the shape of the cluster might also help distinguish the lights from the reflection of small objects. The parameter, or shape of the cluster, is the combination of the two parameters, width and height of the cluster.

## CONCLUSIONS

Image analysis is one of the subjects in the area of computer vision. The image analysis technique includes two main procedures: image processing and pattern recognition. This technique has been successfully applied to several areas, such as document processing, microscopy, radiology, industrial automation, remote sensing, and reconnaissance. This study intends to detect the left-turning movements at intersections using the image technique. An optical image analysis system—including a CCD video camera, interface board, image monitor, and an IBM PC/AT—was used to conduct this research. The conclusions drawn from this study are as follows:

1. The problems in the current method of left-turn data collection are that there is no turning movement data and no wheel-oriented data.

2. An efficient and powerful computer algorithm was developed for this study. The algorithm includes three stages: environment setting, image processing, and recognition of left-turning vehicles.

3. The developed algorithm has been proven as linear. In other words, the time complexity of the algorithm  $T(n)$  equals  $O(n)$  where  $n$  is the size of the image.

4. Software called "Left-Turn Detection" was developed for IBM PC/AT's. This software was written in the C computer language.

5. The computer system developed for detecting left-turning movements is a real time system. The average computer time for the Left-Turn Detection software to analyze an image is 0.85 seconds. The computer time to grab an image from the camera to the computer RAM is 0.05 sec. Thus, the total average computer time for each image is about 0.90 sec. Hence, this computer system is capable of detecting left-turning movements on a real time basis.

6. The accuracy of the algorithm is about 80% for detecting left-turning vehicles whose signal lights are on. Some improvements in both hardware and software are needed in order to reach the implementable stage in the near future.

## ACKNOWLEDGMENT

This study has been sponsored by the Natural Science and Engineering Research Council of Canada.

## REFERENCES

1. P. C. Box. Traffic Studies. In *Transportation and Traffic Engineering Handbook*, 2nd ed., ITE, 1982, pp. 519-554.
2. K. W. Dickinson and R. C. Waterfall. Image Processing Applied to Traffic: 1. A General Review. *Traffic Engineering & Control*, Vol. 25, Jan. 1984, pp. 6-13.
3. D. M. Braston. A Method of Collecting Data on Speeds and Headways on a Motorway. *Traffic Engineering & Control*, Vol. 16, Oct. 1975, pp. 430-432.
4. J. B. Garner and J. Uren. A Digital System for Urban Traffic Flow Data. *Traffic Engineering & Control*, Vol. 19, Aug. 1978, pp. 389-391.
5. L. J. Mountain and J. B. Garner. Application of Photography to Traffic Survey. *The Highway Engineer*, Nov. 1980, pp. 12-19.
6. R. Ashworth. A Videotape-recording System for Traffic Data Collection and Analysis. *Traffic Engineering & Control*, Vol. 17, Nov. 1976, pp. 468-470.
7. A. Polus, M. Livneh, and S. Borovsky. A Multi-purpose Portable Data Collection System. *Traffic Engineering & Control*, Vol. 19, March 1978, pp. 123-125.
8. H. J. Wootton and R. J. Potter. Video Recorders, Micro Computers and New Survey Techniques. *Traffic Engineering & Control*, Vol. 22, April 1981, pp. 213-215.
9. R. Ashworth and A. Kentros. Performance Characteristics of an Ultrasonic Detector for Measurements of Vehicle Occupancy. *Traffic Engineering & Control*, Vol. 17, Dec. 1976, pp. 502-504.
10. E. E. Hilbert, et al. *Wide-area Detection System Conceptual Design Study*. Report FHWA-RD-77-86. FHWA, U.S. Department of Transportation, Feb. 1978.
11. A. P. Schlutsmeyer, et al. *Wide-area Detection System (WADS)*. Report FHWA-RD-82-144. FHWA, U.S. Department of Transportation, Sept. 1982.
12. K. W. Dickinson and R. C. Waterfall. Image Processing Applied to Traffic: 2. Practical Experience. *Traffic Engineering & Control*, Vol. 25, Feb. 1984, pp. 60-67.
13. G. S. Houghton, L. S. Hobson, and R. C. Tozer. Automatic Monitoring of Vehicles at Road Junctions. *Traffic Engineering & Control*, Vol. 28, Oct. 1987, pp. 541-543.
14. A. Rosenfeld. Image Analysis: Progress, Problems, and Prospects. In *Proceedings Pattern Recognition*, IEEE, 1984, pp. 3-12.
15. *Truevision Advanced Raster Graphics Adapter, TARGA 8 User Guide*. AT&T Electronic Photography and Image Center, Indianapolis, Ind., 1985.
16. *Driver's Handbook*. Gouvernement du Quebec, Ministere des Communication, 1983, pp. 122.
17. *Truevision Advanced Raster Graphics Adapter, TARGA 8, M8, 16, 24, and 32, Software Tools Notebook*. AT&T Electronic Photography and Image Center, Indianapolis, Ind., Nov. 1986, Release 3.0.

---

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

# Some Properties of Macroscopic Traffic Models

PAUL ROSS

Three "equations of state" are required to describe the traffic fluid. The first is volume = speed · density, and the second is the continuity of vehicles. There are at least four options for the third equation: (1) the deterministic speed-density model, (2) the equilibrium speed-density model, (3) the Payne model, and (4) the Ross model. Two restrictions on the space step  $DX$  and time step  $DT$  apply to numerical integrations of all four models. There is an additional restriction on  $DT$  that applies to the last three models and a special restriction on the "anticipation" term in the Payne model. The time required to perform numerical integrations of all four models is shown to be inversely proportional to the square of the length of the smallest feature represented. A general form for the "relaxation time" in the three non-deterministic models is derived. It is argued, on the basis of experience with the Ross model, that although a dependence upon speed is "correct," setting the relaxation time constant is adequate for most traffic purposes. The relationship between relaxation time and lost time at signals in the Ross model is shown to be linear.

This paper deals with five topics related to the macroscopic (speed, volume, density) representation of traffic. The topics are: (1) categorization of traffic formulations into four classes, (2) permissible step sizes when numerically integrating the traffic formulations, (3) execution time for numerical integrations, (4) form of dependence of the "relaxation time," a parameter in three of the formulations, upon average traffic speed, and (5) relationship between "lost time" and the relaxation time parameter in one of the models.

## TYPES OF MACROSCOPIC TRAFFIC MODELS

In this paper, the term "traffic dynamics model" or "traffic formulation" means a complete set of relationships between traffic volume, average traffic speed, and traffic density. Such relationships may be looked upon as the "equations of state" of the traffic fluid.

Three relationships are required. The first relationship is inherent in the definitions of traffic volume, speed, and density:

$$Q = kv \quad (1)$$

where

$$Q = Q(x,t) = \text{traffic volume (veh/hr) at location } x \text{ and time } t,$$

$k = k(x,t) =$  vehicular density (veh/mi) at location  $x$  and time  $t$ , and

$v = v(x,t) =$  space-mean speed (mi/hr) at location  $x$  and time  $t$ . (Proof that this is the harmonic mean of the "spot" speeds is provided elsewhere (1).)

A second relationship, the continuity of vehicles, was pointed out by Lighthill and Whitham (2):

$$\partial k / \partial t + \partial Q / \partial x = S(x,t) \quad (2)$$

where

$\partial / \partial t$  and  $\partial / \partial x$  indicate partial differentiation with respect to time  $t$  and with respect to location along the road  $x$ , respectively, and  $S(x,t) =$  source strength of vehicles from ramps, parking lots, etc., which may be negative (veh/mi-hr).

Equations 1 and 2 are fundamental. All traffic models that deal with volume, speed, and density must incorporate them or equivalent relationships.

At least four possibilities for the third relationship have been proposed, as follows:

### Deterministic Speed-Density Hypothesis

The deterministic speed-density traffic formulation states that the average traffic speed is a function of traffic density. [Greenshields (3) was the first to hypothesize that average traffic speed is a deterministic function of density, but innumerable investigators have followed his lead. A summary is provided elsewhere (4). The most recent authoritative work to adopt this approach is the Highway Capacity Manual (5), in its treatment of freeways.]

$$v = v(k) \quad (3)$$

The traffic-free speed is  $v(0)$ ;  $v(k_{\text{jam}}) = 0$ , where  $k_{\text{jam}}$  is the so-called "jam density" of vehicles; and  $\max[k \cdot v(k)]$  is the roadway capacity. The precise dependence of  $v$  upon  $k$  is not important to the arguments in this paper.  $v(k)$  need not be single-valued except at  $k = k_{\text{jam}}$  and  $k = 0$ .

### Equilibrium Speed-Density Hypothesis

The equilibrium speed-density formulation states that there is an equilibrium speed, which is a function of density, to which actual speeds relax. [The author's experience is that the equilibrium speed-density hypothesis is accepted by traffic researchers but has never been specifically proposed in the

literature. It is a logical implication of the work of Prigogine and Herman (6), who postulated that there is an equilibrium distribution of traffic speeds to which traffic relaxes.]

$$\partial v/\partial t + v \partial v/\partial x = [F(k) - v]/T \quad (4)$$

The terms on the left of the equal sign give the acceleration of traffic with respect to the moving traffic stream;  $F(k)$  is the equilibrium speed, and  $T$  is the so-called relaxation time (the parameter which controls how quickly traffic returns to its equilibrium speed). Actual free speed, maximum flow, and density at  $v = 0$  are variable depending upon the boundary conditions.

### Payne's Hypothesis

To overcome the tendency to spontaneously lock up in the deterministic and equilibrium speed-density hypotheses, Payne added an anticipation term (7, 8):

$$\partial v/\partial t + v \partial v/\partial x = [F(k) - v]/T - (g/T)\partial k/\partial x \quad (5)$$

The anticipation constant is  $g$ . The net effect of Payne's addition is to cause traffic to accelerate when  $\partial k/\partial t$  is negative—i.e., when the traffic anticipates lower density ahead. Otherwise, Payne's formulation is quite similar to the equilibrium speed-density hypothesis.

### Ross's Hypothesis

In a recent paper (9), Ross claims that the three traffic formulations listed above are all grossly unrealistic. He proposes that the third defining equation is

$$\partial v/\partial t + v \partial v/\partial x = [F - v]/T \quad (6)$$

where  $F$  is the free speed on the roadway and is explicitly not dependent on  $k$ . Roadway capacity and jam density are accounted for by separate constraints; jam density flow is constrained to be incompressible.

### INTEGRATION STEP SIZE

No matter which of the four traffic formulations is used, integration is frequently necessary. The simplest integration method is to step along the roadway using the chosen model to evaluate the volume, density, and speed at points a distance  $DX$  apart and to repeat that process every  $DT$  hours. The speed of the integration process is inversely proportional to the size of  $DX$  and  $DT$ ; it is therefore important to make  $DX$  and  $DT$  as large as possible without compromising the accuracy of the integration. We investigate what restrictions there should be on  $DX$  and  $DT$  to ensure accurate integration.

Continuity of vehicles (equation 2) imposes some fundamental constraints on the integration of all four traffic formulations. Equation 2 can be rewritten in terms of finite differences:

$$Dk_i = S_i DT - (Q_i - Q_{i-1}) (DT/DX) \quad (7)$$

where  $Dk_i$  is the change in density at location  $i$  between times  $DT$  apart,  $S_i$  is the traffic source strength at location  $i$ , and  $Q_i$  is the volume at location  $i$ .

The integration error will usually be small compared to  $k_i$  if  $Dk_i$  is small compared to  $k_i$ . Divide both sides of equation 7 by  $k_i$ :

$$Dk_i/k_i = (S_i/k_i) DT - (v_i - v_{i-1})(DT/DX) \quad (8)$$

where  $v_i$  is the volume at location  $i$ .

If both terms on the right are individually small,  $Dk_i/k_i$  will automatically be small. ( $Dk_i/k_i$  will also be small if the two right-side terms nearly cancel one another, but an integration scheme that relies on two large terms nearly cancelling one another would be impossible to implement.) Specifically:

$$DT \ll k_i/S_i \quad (9)$$

$$DX/DT \gg F \quad (10)$$

where  $F$  is the free speed on the roadway.

Since the source/sink volume is  $S_i \cdot DX$ , not  $S_i$ , restriction 9 can be rewritten:

$$DX/DT \gg \text{Source volume}/k_i \quad (11)$$

Inequalities 10 and 11 must both be satisfied if the integration is to give relatively accurate results. Inequality 10 means that the time step,  $DT$ , must be small enough and the space step,  $DX$ , large enough so that vehicles cannot cross an appreciable fraction the space step in one time step. Inequality 11 means that the time step must be small enough and the space step large enough that source/sink flows do not appreciably alter the number of vehicles in any space step during a time step. These conclusions are based entirely on equation 2 and, therefore, apply to all traffic formulations.

Since the third equation of the deterministic traffic formulation does not involve differentiation, it has no effect on the size of  $DX$  or  $DT$  in the deterministic model. Restrictions 10 and 11 are the only restrictions on  $DX$  and  $DT$  in the deterministic formulation.

In the equilibrium, Payne, and Ross formulations, the third equation of state does involve partial differentiation and, therefore, has an effect on the allowable sizes of  $DX$  and  $DT$ . The third equations in these formulations are similar enough to one another for the convergence properties under numerical integration of all three formulations to be analyzed at the same time. Consider the third equation of state from Payne's formulation converted to finite difference form:

$$Dv_i = [F(k_i) - v_i](DT/T) - v_i(v_i - v_{i-1}) (DT/DX) - (g/T)(k_i - k_{i-1})(DT/DX) \quad (12)$$

There are no fixed values of  $DX$  and  $DT$  which guarantee that  $Dv_i$  will be small compared to  $v_i$ ; we must settle for the weaker condition that  $Dv_i$  be small compared to the free speed,  $F$ . Again, each term individually must be small (compared to  $F$ ).

The condition on the first term is

$$| [F(k_i) - v_i](DT/T) | \ll F \quad (13)$$

which is always satisfied if

$$DT \ll T \quad (14)$$

Since a term with similar convergence properties occurs in the equilibrium and Ross formulations, restriction 14 applies equally to the equilibrium, Payne, and Ross formulations.

The second term in equation 12 is automatically small com-



TABLE 1 CONDITIONS ON INTEGRATION STEP SIZES  $DX$  AND  $DT$  FOR THREE TRAFFIC SIMULATION PROGRAMS

Restriction	KRONOS (Deterministic)	FREFLO (Payne)	RFLO (Ross)
10. $DX/DT \gg F$	68 mi/hr ?? 63 mi/hr	630 mi/hr >> 63 mi/hr	200 mi/hr ?> 63 mi/hr
11. $DX/DT \gg$ source/ $k$	68 mi/hr ?> 20 mi/hr	630 mi/hr >> 20 mi/hr	200 mi/hr >> 20 mi/hr
14. $DT \ll T$	Not applicable	0.000167 hr << 0.001875 hr	0.0005 hr << 0.0060 hr
16. $DX/DT \gg$ $g k_{jam}/TF$	Not applicable	630 mi/hr >> 68 mi/hr	Not applicable

NOTE: Restriction number refers to inequality number in the text. “??” means the restriction is not satisfied; “?>” means the restriction is marginally satisfied.

pared to  $F$  if restriction 10 is satisfied; no new restriction is needed.

The third term, the anticipation term, appears in Payne’s formulation only.

$$|(g/T)(k_i - k_{i-1})(DT/DX)| \ll F \quad (15)$$

Condition (15) is guaranteed if

$$DX/DT \gg g k_{jam}/TF \quad (16)$$

Restriction (16) applies to the Payne formulation only.

How well are these restrictions on  $DX$  and  $DT$  obeyed in practice? KRONOS (10) is the only recent simulation program to use the deterministic speed-density formulation. FREFLO (11) is the only simulation program to use the Payne formulation. RFLO, now under development at the Federal Highway Administration, uses the Ross formulation. (No example of a simulation using the equilibrium speed-density formulation could be found.)

It is assumed that the maximum average speed,  $F$ , is about 63 mi/hr. It is further assumed that typical source/sink flows  $\sim 100$  veh/hr and worst case (smallest) traffic densities  $\sim 5$  veh/mi, implying source flow/ $k \sim 20$  mi/hr.

In the KRONOS program, the space step,  $DX$ , defaults to 100 feet and the time step,  $DT$ , defaults to 1 second, yielding  $DX/DT = 68$  mi/hr. Table 1 shows how these values relate to restrictions 10, 11, 14, and 16.

In the FREFLO program, the space step,  $DX$ , is the link length;  $DX = 0.1$  mi can be postulated.  $DT$  is one-tenth of the travel time on the shortest link (0.00017 hours, with our assumptions), automatically making  $DX/DT$  ten times  $F$ . The relaxation time,  $T$ , is proportional to  $DX$  and inversely proportional to the roadway capacity; for capacity = 2000 veh/ lane-mi,  $T = 0.001875$  hr. The anticipation constant,  $g$ , is proportional to  $DX$  and the roadway capacity; in these conditions,  $g = 0.028$  mi<sup>3</sup>/hr. Jam density is 143 veh/lane-mi—say, 286 veh/mi on a two-lane roadway. The term  $g k_{jam}/TF$  evaluates to 68 mi/hr. (The similarity to  $DX/DT$  in the KRONOS program is coincidental.) Table 1 shows how these values relate to the applicable restrictions.

In the RFLO program,  $DX = 0.1$  mi;  $DT = 0.0005$  hr = 1.8 sec; ( $DX/DT = 200$  mi/hr);  $T = 0.0060$  hr. Table 1 relates these values to the applicable restrictions.

It is obvious from inspecting Table 1 that the KRONOS program does not satisfy restriction 10. This implies poor representation of density where it is changing rapidly. KRONOS is marginal with respect to restriction 11, implying that its representation of low-density traffic with comparatively large source/sink volumes is theoretically unsound. (Note, however, that the author’s experience is that, although one wants

the one-step change in any computed quantity to be less than 10% or so, integration steps that can, in theory, allow changes of 20 or 30% to work very well in practice. This is due to the fact that 20 or 30% changes only appear at such abrupt discontinuities that they rarely occur in real traffic.)

RFLO, with  $DX/DT$  less than four times the free traffic speed, is marginal with respect to restriction 10. Problems in the representation of density have not been detected in practice, but the possibility should be noted.

FREFLO satisfies all conditions well. In fact, the time step,  $DT$ , could probably be doubled or tripled in FREFLO without noticeable loss in accuracy.

#### MINIMUM EXECUTION TIME

Although the accuracy of the numerical integrations increases as  $DX$  becomes larger, precision increases as  $DX$  becomes smaller. If one wishes to simulate the effects of very small geometric features (such as an intersection wherein opposite streets are misaligned by, for example, 20 ft), one must make  $DX$  small (for example, 5 ft). Restrictions 10 and 11 both require that  $DT$  be made small proportionately as  $DX$  is made small. Since the number of space steps simulated is inversely proportional to  $DX$  and the number of time steps is inversely proportional to  $DT$ , the total computation time must be inversely proportional to the square of  $DX$ .

This conclusion—that minimum computation time is inversely proportional to the square of the length of the smallest feature simulated—applies to all four traffic formulations.

#### RELAXATION TIME: $T$

The three non-deterministic traffic formulations all use a quantity called the relaxation time, which has been symbolized by  $T$  here. The original formulation of the equilibrium speed-density hypothesis (equation 4) allows that  $T$  is probably a function of average traffic speed ( $v$ ), but determining that functional dependence was beyond the scope of the original paper, which opted for the simplifying assumption that  $T$  is independent of  $v$ . The question now is, What is a good functional form for relaxation time,  $T(v)$ ?

A small value for  $T$  means that the average traffic speed,  $v$ , relaxes to its equilibrium value quickly—that is, traffic acceleration is inversely proportional to  $T$ . Since a good deal is known about the acceleration of traffic as a function of speed, a functional form for  $T(v)$  can be deduced.

The average power used to accelerate vehicles is

$$P = m \cdot v \cdot \text{acceleration} \quad (17)$$

where  $m$  is the average mass of vehicles.  $T$  is inversely proportional to the acceleration:

$$T = c(mv + b)/P \quad (18)$$

where  $c$  is a constant to be determined, and where  $b$  is a small number added to account for the fact that accelerations at  $v = 0$  are not infinite (i.e.,  $T \neq 0$ ) but rather are limited by pavement friction and driver discomfort.

$P$  is not the engine power used, but rather the power used to accelerate the vehicle after rolling and wind resistance have been overcome. The wind and rolling resistance are approximated by constant forces, so that the power used is linear in  $v$ . The dependence of  $T$  on  $v$  is therefore approximated by

$$T(v) = (c_1 + c_2 v)/(1 - c_3 v) \quad (19)$$

The three  $c$ 's can be roughly estimated. We note that, when traffic is traveling at its maximum possible speed, it cannot accelerate further—i.e.,  $T \rightarrow \infty$  as  $v \rightarrow v_{\max}$ . Estimates are that the average maximum speed of the North American traffic stream is in the range 95 to 100 mi/hr and the average speed on level freeways is 63 mi/hr. A rough estimate of  $c_3$  is therefore  $0.67/F$ , where  $F$  is the free, desired speed on the road.

The remaining two parameters,  $c_1$  and  $c_2$ , can be chosen by noting that—in simulations with  $T$  held constant— $T = 0.030$  hr produces realistic traffic performance at speeds approaching the free speed ( $F$ ), and  $T = 0.0053$  hr produces signal discharge flows that represent about 2.1 seconds of lost time per green.

We conclude that the relaxation time,  $T(v)$ , can be approximately represented by

$$T = (0.0053 + 0.0047 v/F) \text{ hr}/(1 - 0.67 v/F) \quad (20)$$

At least three assumptions have been made in the above derivation:

1. The effects of traffic mix, roadway grade, and driving conditions on  $T(v)$  are represented by making the constants which multiply  $v$  inversely proportional to  $F$ .

2. Deceleration behaves essentially the same as acceleration.

3. Traffic speed will never approach  $F/0.67$ .

These assumptions are mathematically convenient and not obviously wrong. The argument applies equally to the equilibrium, Payne, and Ross traffic formulations.

Extensive simulations with the Ross formulation indicate that there are no startling differences between the  $T(v)$  variable as described above and  $T = \text{constant}$ —only subtle differences in the acceleration of traffic back to its desired speed upon leaving a bottleneck. The details of such accelerations have never been an important traffic issue. It is the author's conclusion that setting the relaxation time  $T = \text{constant}$  is adequate for most traffic purposes using the Ross formulation. This conclusion has not been tested for the equilibrium or Payne formulations.

### “LOST TIME” AT SIGNALS

The relaxation time,  $T$ , affects all aspects of traffic behavior in the three formulations where it is used. Its most obvious

effect is on “lost time” at signals. Because signal lost time is well known (12), this dependence can be used to estimate an appropriate value for  $T$ .

Consider a traffic signal with, for example, 0.01 hours of red alternating with 0.01 hours of green (cycle length 0.02 hr = 72 sec). It is straightforward to use any of the traffic formulations to simulate such a traffic condition. The results of one such simulation with the Ross model are shown in Figure 1.

When the simulation of Figure 1 is extended for a long time, a standing queue forms at the signal and the downstream flow stabilizes at 933 veh/hr. This implies an effective green time of  $933/2000 = 0.466$  hrs/hr or 33.6 sec/cycle. Lost time is: (36 sec of actual green per cycle) - (33.6 sec of effective green per cycle) = 2.4 sec/cycle.

Similar simulations can be repeated using different values of relaxation time and noting the resulting lost times. The relationship between relaxation time and lost time in the Ross formulation is shown in Figure 2. Figure 2 applies to the Ross formulation only.

### SUMMARY

This paper has discussed five related topics. The conclusions are:

1. Three equations of state are required to describe the traffic fluid. The first is volume = speed · density, and the second is the continuity of vehicles. There are at least four options for the third equation: the deterministic speed-density model, the equilibrium speed-density model, the Payne model, and the Ross model.

2. Two restrictions on the space step  $DX$  and time step  $DT$  apply to numerical integrations of all four models. There is an additional restriction on  $DT$  that applies to the last three models and a special restriction on the anticipation term in the Payne model.

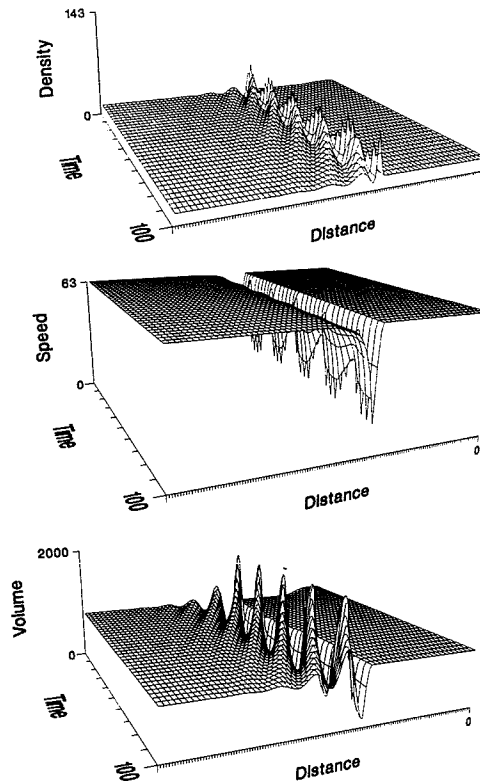
3. The time required to perform numerical integrations of all four models is inversely proportional to the square of the length of the smallest feature represented.

4. A general form for the relaxation time in the three non-deterministic models is derived. It appears, on the basis of experience with the Ross model, that although a dependence upon speed is “correct,” setting the relaxation time constant is adequate for most traffic purposes.

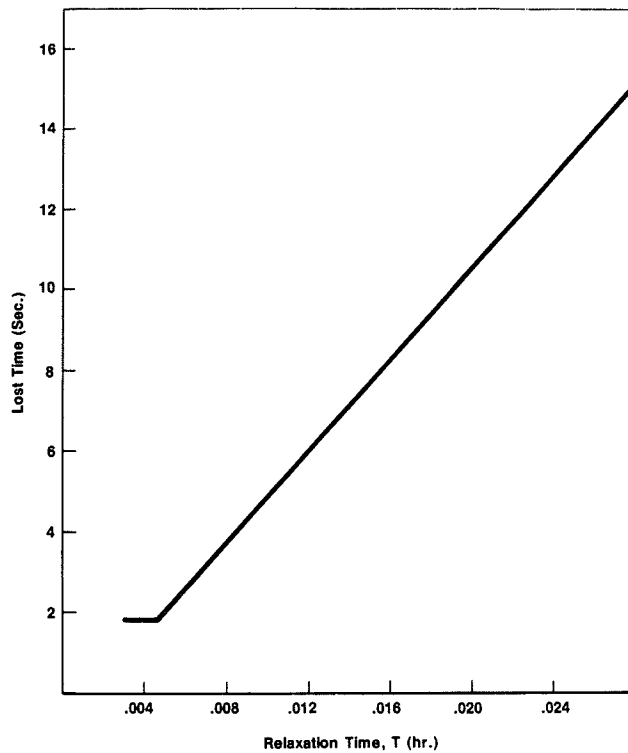
5. The relationship between relaxation time and lost time at signals in the Ross model is linear.

### ACKNOWLEDGMENT

The author gratefully acknowledges the help of Osama Sharaf-Eldien and Walter Hunter, of SRA Technologies. Sharaf-Eldien provided all the material relating to the current computer implementation of FREFLO and helped to interpret it. Hunter set up the perspective presentation in Figure 1. This work was done in the course of the author's employment with the Federal Highway Administration. However, FHWA accepts no responsibility for the accuracy or reliability of any statements made herein. Nothing in this paper constitutes an FHWA policy or standard.



**FIGURE 1** Density, speed, and volume as represented by the Ross traffic formulation at a traffic signal with actual green = 36 sec/cycle, actual red = 36 sec/cycle. Demand volume rises from 800 to 950 veh/hr. Traffic flows from right to left. Distance is in units of 0.1 mi (13 mi total). Time runs from 0.00 to 0.10 hr (five cycles), back to front. Jam density of vehicles is 143 veh/lane-mi. Free speed is 63 mi/hr.



**FIGURE 2** Signal lost time as a function of relaxation time,  $T$ , in the Ross model. Free speed = 63 mi/hr. Cycle length = 0.02 hr with 50% actual green. Lost times less than the integration step size (1.8 sec) are not observed.

## REFERENCES

1. J. G. Wardrop. Some Theoretical Aspects of Road Traffic Research. *Proc., Institute of Civil Engineers, Part II*. Vol. 1, No. 2, 1952, pp. 325-378.
2. M. J. Lighthill and G. B. Whitham. On Kinematic Waves II-A Theory of Traffic Flow on Long Crowded Roads. In *Special Report 79: An Introduction to Traffic Flow Theory*, HRB, National Research Council, Washington, D.C., 1964, pp. 8-35.
3. B. D. Greenshields. A Study of Traffic Capacity. *HRB Proc.*, Board 14, 1934, pp. 448-474.
4. D. L. Gerlough and M. J. Huber. Traffic Flow Theory. In *Special Report 165*, TRB, National Research Council, Washington, D.C., 1975.
5. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
6. I. Prigogine and R. Herman. *Kinetic Theory of Vehicular Traffic*. American Elsevier, New York, N.Y., 1971.
7. H. J. Payne. Models of Freeway Traffic and Control. *Proc., Simulation Councils*, Vol. 1, No. 1, Simulation Councils Inc., La Jolla, Calif., 1971.
8. L. Isaksen and H. J. Payne. Simulation of Freeway Traffic Control Systems, The Mathematics of Large-Scale Simulation. *Proc., Simulation Councils*, Vol. 2, No. 1, Simulation Councils Inc., La Jolla, Calif., June 1972, pp. 35-42.
9. P. Ross. Traffic Dynamics. *Transportation Research*, in preparation.
10. P. G. Michalopoulos. Dynamic Simulation Program for Personal Computers. *Transportation Research Record 971*, TRB, National Research Council, 1984, pp. 68-79.
11. E. B. Lieberman, B. Andrews, M. Davila, and M. Yedlin. *Macroscopic Simulation for Urban Traffic Management*. Reports FHWA RD-80-113, -114, and -115. FHWA, U.S. Department of Transportation, 1982.
12. A. K. Gilbert. Signalized Intersection Capacity. *ITE Journal*, Vol. 54, No. 1, Jan. 1984, pp. 48-52.

---

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

# Generating Partial Origin-Destination Tables for Streamlined Application of Corridor Models

SAM YAGAR

A method is described for generating origin-destination (OD) matrices for the subset of drivers who are likely to divert in response to any changes to the traffic network, such as ramp metering for freeways. These ODs are generated automatically from responses to survey questionnaires handed out at strategic points in the network. The responses are factored up to represent 100% of the observed flow on critical links. The procedure also preloads onto the network any link volumes that are not represented by factored survey responses, and are presumably insensitive to any control schemes that might be considered. Preloading of the less sensitive flows, rather than the existing procedure of creating and assigning pseudo-ODs, should reduce the work required of the analyst and lead to improved predictions of flows and queues.

Assignment-based traffic models require origin-destination (OD) information in the form of OD tables or matrices. Further, dynamic assignment models such as CORQ (1) require the input of set OD matrices to represent the trip demands in each of a series of contiguous time slices (2).

Field data are obtained by surveying drivers and must include at least the trip path and trip OD for each surveyed driver. Usually, only information from drivers whose trip paths are sensitive to the effects of any proposed traffic controls are required. However, most assignment models require that this information be presented to them in the form of OD matrices.

One of the most time-consuming tasks in traffic studies is the conversion of raw field data into these OD matrices. This is complicated by the fact that full OD matrices must generally contain some pseudo-ODs in order to represent all of the flows on the network. These pseudo-ODs represent trips which are generally not sensitive to any control strategies that might be tested and therefore need not really be represented by the user survey that is conducted to obtain the OD information.

This paper describes the structure of a preprocessor for producing OD matrices for use by corridor assignment models such as CORQ2 (3).

## RATIONALE

Kuwahara and Sullivan (4) have identified two primary problems which are encountered when one attempts to convert roadside survey data into OD matrices:

1. Double-counting: Trips which pass through more than one survey location may lead to errors when scaling the survey responses to represent the total flow.

2. Leaky screenlines: It is often physically impractical to set up a survey location at every possible crossing point of a screenline.

In their paper, Kuwahara and Sullivan proposed five methods for overcoming these problems. All of their solutions, however, involve estimating the probability that a trip between one origin and one destination will take a specified path. These probabilities will generally be difficult to generate accurately in real situations. Since the outputs of their procedure depend so heavily on these uncertain probabilities, so will the final assignment. This paper outlines a different approach.

Historically, the ODs obtained directly from survey responses have been factored up to match the flows and queues observed on the network, as all trips on the network could not be sampled. Inevitably, there are flows remaining on the network which are not explained by the factored survey responses, yet the models require full ODs. The traditional approach to this discrepancy has been to create pseudo-ODs traveling between fictitious or irrelevant origins and destinations to compensate for any such flows. By doing this, the analysis program will always be able to perform the assignment to a network which is initially empty.

The CORQ2 program does not require that the network be initially empty before assigning any ODs to it. Instead, it uses a set of "deterministic preloads" (DPs) in addition to the OD matrices. DPs are fixed "deterministic" flows loaded onto the network before the OD demands are assigned. The use of DPs allows the analyst to use OD matrices containing only the network sensitive demands that are of interest, while still representing the effect of the remaining flows on the network.

The procedure of producing ODs and DPs by the preprocessor outlined herein, followed by assigning them by CORQ2, obviates the leaky-screenline problem. The creation of pseudo-ODs added no more relevant information to the analysis. Rather, it merely created additional work, both at the preprocessor stage and at the assignment/analysis stage. Indeed, creating and assigning pseudo-ODs actually confounds the analysis, as the pseudo-ODs are given equal weight with the survey ODs in an automated assignment procedure. The use of DPs allows the more sensitive drivers to be assigned last, as the network assignment approaches its equilibrium. This improves

Department of Civil Engineering, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.

the modeling realism. If the survey locations have been chosen so that the people surveyed are the ones who will be sensitive to any traffic management strategies, then the proper information required for an analysis has been obtained. In reality, the drivers who create the rational equilibrium are those who are most sensitive to link costs. Therefore, they should be the last ones assigned. This is accomplished by preloading the insensitive drivers.

The concept of using DPs rather than pseudo-ODs was proposed in an earlier paper (5). The rationale behind DPs is to permanently load onto the network first those flows which will not change their routes as a result of control strategies. These "permanent" flows are preloaded onto the network. Then, assignment is performed for those trips which are sensitive to any network changes. In this way the assignment procedure is most sensitive to the most sensitive trips.

Through the use of DPs, the preprocessor and CORQ2 reduce the number of steps and the amount of computational effort required to simulate the operation of a network. Rather than expending the effort of removing the additional flows, creating pseudo-ODs for them, and then reassigning them to the network, CORQ2 simply preloads them onto the network before any ODs are assigned. Thus the effects of these flows on the network will be present regardless of the control strategies tested.

The preprocessor also deals with the problem of double-counting. The rest of this paper is devoted to the description of the preprocessor and the method that it uses to scale up the survey responses to create OD matrices that represent all of the assignable trips once and only once. As discussed above, the DPs will always be present on the links and need not be assigned.

## PROCEDURE

The preprocessor attempts to factor up the survey responses at each survey location in order to represent all of the counted trips passing through that location. If the drivers passing through

a survey location are likely to have to respond to a traffic management strategy that will be tested using the ODs, such as the potential metering or closure of an on-ramp to a freeway, then it is necessary to have the sensitive flows at that location represented by ODs. Since all of the vehicles on a link would have to change paths if the link were closed, it is customary to factor survey responses for a link up to 100% of the link's flow. When the factoring is completed, these flows are removed from the network and placed into the OD matrices. Any flows remaining on the network, and not represented by the assignable OD matrices, are considered to be deterministic preloads.

The preprocessor requires the following input data:

- network topology (how links connect)
- link cost vs. flow relationships
- list of origin nodes and destination nodes
- observed flows and queues on links
- placement of survey handout locations
- first and last survey numbers handed out at each location
- total flow passing through each surveyed link
- traces of paths followed by representative samples of the drivers using the survey links

A returned survey questionnaire provides a trip trace from which can be gleaned the origin, trip path (nodes or links), destination, time of departure, and a list of all survey locations passed. The map from a typical questionnaire for this purpose is shown in Figure 1.

Given the above data, the preprocessor examines each questionnaire response. It determines the survey location and the handout time based on a coded number assigned to each individual handout. Starting at this location, it traces backward and forward along the stated path and generates a list of the links traveled. It also searches for the appropriate origin and destination nodes. A coded sequence of nodes/links for the trip traced in Figure 1 is illustrated in Figure 2. The latter represents the form in which the driver-traced trip of Figure 1 is fed to the preprocessor.

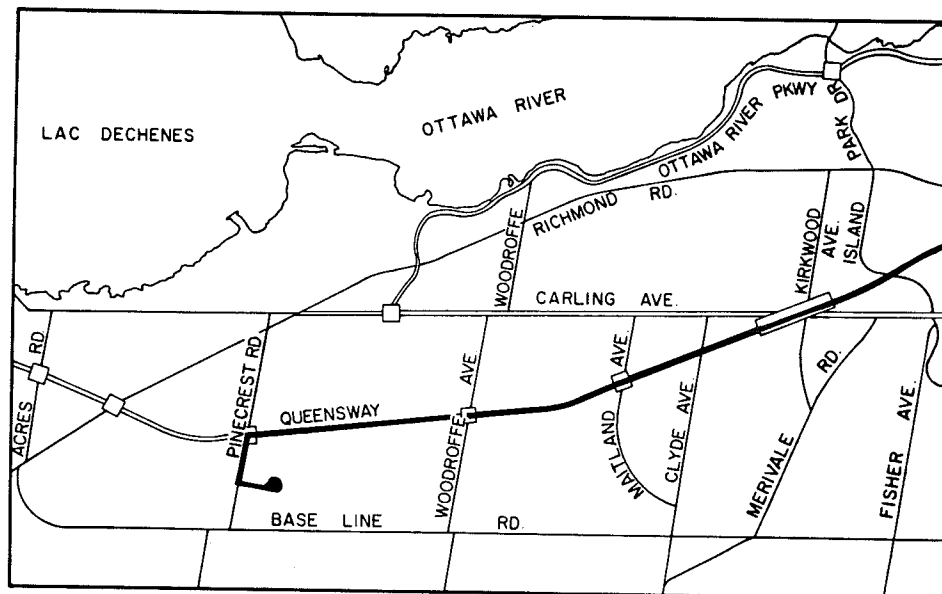


FIGURE 1 Driver's trace as shown on the survey questionnaire.

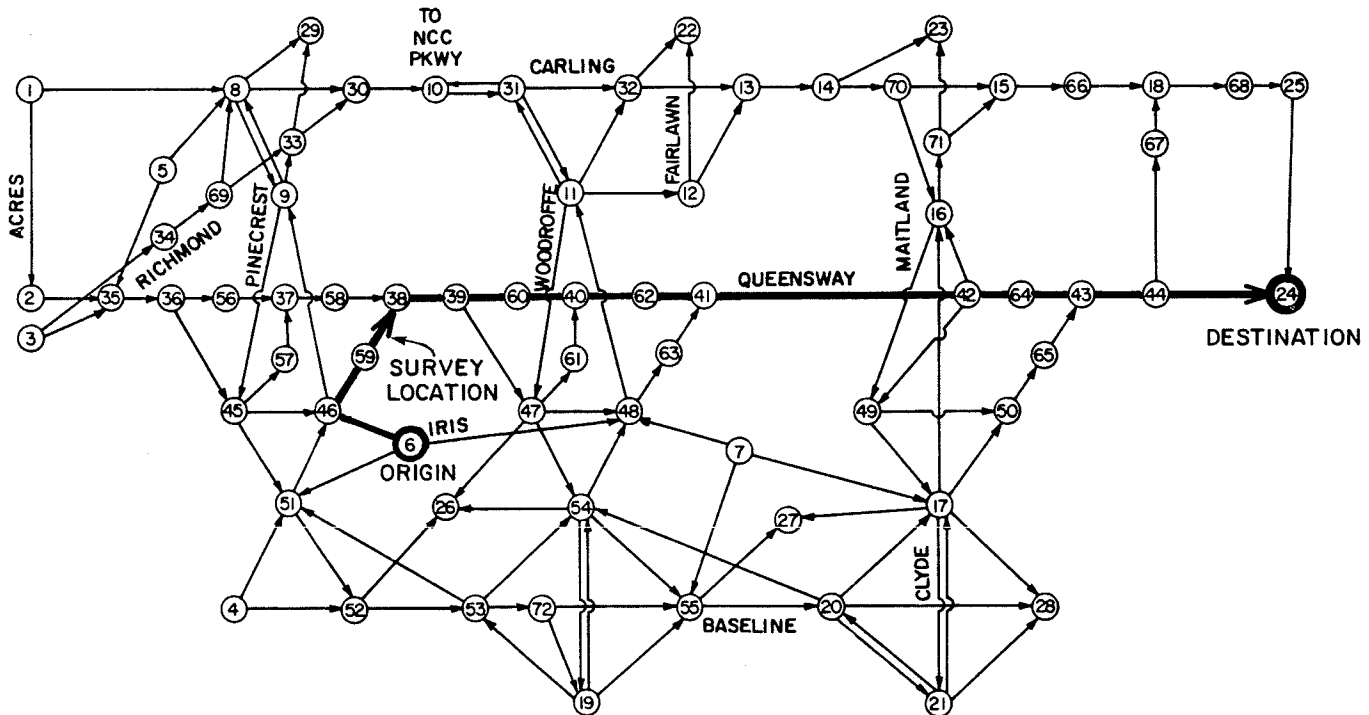


FIGURE 2 Driver's trace on link-node diagram.

Generally, the analyst will have coded the path as starting at one of the prescribed origin nodes and ending at one of the prescribed destination nodes. Since the path specified by the motorist may not exactly coincide with the network as defined, the analyst may have to make some decisions regarding what origin, destination, and intermediate trajectory should be chosen. If the start or end of the trip is not well enough defined that it automatically occurs near a terminal node, then the analyst will have to be careful as to which one is selected.

As an example, if the motorist's returned questionnaire in Figure 1 had only indicated that the start of the trip was at the Pinecrest interchange, the analyst would have to decide whether to assign that trip to origin node 9 or origin node 6. These are the origin nodes closest to the location at which the trip specified by the driver started. The node sequence from origin node 9 would be 9,45,57,37,58,38, . . . , rather than the sequence 6,46,59,38, . . . , from origin node 6 that was illustrated in Figure 2.

The choice of the origin node will clearly be an issue here in terms of the analyses to be performed, especially since this questionnaire will generally be factored up to represent several vehicles in the OD matrix. For example, if the interchange at Pinecrest were to be removed, then motorists from origin node 9 would likely divert along Carling Avenue, while motorists from origin node 6 would likely divert either to Baseline Road or to the Woodroffe on-ramp to the Queensway. The CORQ2 preprocessor has several routines for estimating appropriate origin/destination nodes when the driver has not specified trip-end at origin and destination nodes. These are based on the network connectivity and topology. It is noted that not all network nodes are origin and/or destination nodes.

After the entire path for a trip has been defined, from the origin node through a series of links to the destination node,

the time of arrival at each node is calculated so that the preprocessor knows in which timeslice the trip traveled on each link. The timeslice in which the trip originates is also calculated. The preprocessor lists the survey locations through which this trip passed and the timeslices in which the respective survey locations were passed, and stores these on a temporary file for later use as described below. It also updates the number of returned questionnaires from each survey location.

After all of the questionnaires have been processed, there is enough information available to calculate the scaling factors for each response. At this point, the responses are read back in from the temporary file, factored up, removed from the observed flows and queues on the appropriate links (in the appropriate timeslices), and added to the OD matrices. The OD matrices and the remaining deterministic link preloads are then written out to files for use by CORQ2.

### SCALING FACTORS

If each surveyed motorist were to pass through only one survey location, then the factoring procedure would be simple. Simply dividing the total flow at each location by the number of responses attributed to location would give the correct scaling factor.

For an idealized situation in which there is no interaction between survey locations, the simple relationship  $F = V/R$  holds, where:

- $F$  = the factor for a survey location,
- $V$  = the total volume at the survey location, and
- $R$  = the number of respondees (returned questionnaires) from the survey location.

For example, if there is a flow of 300 vehicles on the surveyed link in a timeslice, and 50 questionnaires are returned, then every respondee is assumed to represent  $F = 300/50 = 6$  vehicles. As usual, it is assumed that the returned questionnaires form a representative sample of the trips of all of the motorists using the surveyed link.

In practice, there may well be trips which pass through more than one survey location. This is where the scaling becomes difficult. If a vehicle passes through one survey location with a simple scaling factor of 4, for example, and another location with a scaling factor of 6, then how many vehicles does this respondee really represent on each of the traveled links? This is the basic problem encountered when attempting to factor up the survey responses to the total OD matrix. It cannot be solved exactly in mathematical terms, because it cannot be fully resolved theoretically.

When some trips pass through two or more survey locations, we are not guaranteed an exact solution to the problem. The procedure used in the CORQ2 preprocessor is summarized below. A full description of the procedure is beyond the scope of this short paper. Each trip passing through  $N$  survey locations increments counters so that  $1/N$ th of the effect of the response is assigned to each survey location. This has roughly the same effect as setting the factor for that trip equal to the average of the factors of the  $N$  locations. While this procedure is quite robust, it can still be aided by the user through proper selection of survey locations to minimize the number of trips passing through more than one survey location. It also helps if drivers' response rates at the various survey locations are about the same.

## DISCUSSION OF RESULTS

Using the preprocessor output as input data to a dynamic corridor assignment model such as CORQ2 should serve to reproduce the flows and queues observed in the field. However, the simulated results will vary somewhat from those observed for the following reasons:

1. The survey responses returned represent only a strategically chosen sample of the users of the network.
2. Flows which pass through more than one survey location may not be factored up precisely as discussed above in the description of the scaling algorithm. Approximation is necessary.
3. Different people have different sets of values and measurements, and they will not necessarily choose the same path between a given origin and destination at a given time, i.e., they will not necessarily utilize the path selected by CORQ2, for example. This is a problem encountered by all assignment techniques. However, the use of a shortest path algorithm in producing the OD matrices for assignment by a shortest path algorithm is a form of pre-calibration, and should help to reduce the error caused by individual driver preferences.

## REFERENCES

1. S. Yagar. CORQ—A Model for Predicting Flows and Queues in a Road Corridor. *Transportation Research Record* 533, TRB, National Research Council, 1975, pp. 77–87.
2. S. Yagar. Dynamic Traffic Assignment by Individual Path Minimization and Queueing. *Transportation Research*, Vol. 5, Pergamon Press, 1971, pp. 179–196.
3. S. Yagar. Characteristics of a Freeway Corridor Model. *Procs., Engineering Foundation Conference on Management and Control of Urban Traffic Systems*, Henniker, N.H., Eng. Found. Publication No. 87-14, June 1987, pp. 161–170.
4. M. Kuwahara and E. C. Sullivan. Estimating Origin-Destination Matrices from Roadside Survey Data. *Transportation Research*, Vol. 21B, No. 3, 1987, pp. 233–248.
5. S. Yagar. Partial Synthesis of Traffic Demands to Facilitate Application of Corridor Models. *R.T.A.C. Forum*, Vol. 5, No. 2, 1983, pp. 31–37.

---

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



# Modeling the Effect of Traffic Signal Progression on Delay

KENNETH G. COURAGE, CHARLES E. WALLACE, AND RAFIQ ALQASEM

The Highway Capacity Manual (HCM) offers a new model for assessing the effect of signal progression on delay at a signalized intersection. This paper discusses a comparison between the HCM progression model and the progression model used in TRANSYT, a signal design and evaluation program which has been in use for several years. The TRANSYT-7F program is used to compare the delays estimated for various qualities of progression with the delays estimated for random arrivals. Comparisons are made on a single pair of links under controlled conditions and on a network of 85 links under simulated field conditions. It was demonstrated that the two techniques agree quite closely. It was also observed that the platoon ratio,  $R_p$ , as defined in the HCM, provides a better predictor of progression quality with heavy traffic volumes. The TRANSYT results suggest that a wider range of progression adjustment factors exist than the HCM recognizes, and that some extrapolation of the HCM values may be warranted to cover exceptionally good and exceptionally poor progression. An independent indicator of progression quality was also developed and tested. It is derived from the ratios of bandwidth measured on the time-space diagram and is therefore termed the "band ratio." The advantage of the band ratio is that, unlike the platoon ratio, it may be computed without field studies. From the studies reported in this paper, it appears that the band ratio may be used as a cost-effective substitute for the platoon ratio for most purposes.

The need to coordinate the operation of two or more traffic signals which operate in close proximity is self evident. A wealth of literature exists on the subject of coordinated signal systems. A variety of techniques, ranging from simple graphic approaches to microscopic computer simulation programs, is available to the analyst. Each technique deals with some aspect of the system performance, expressed in terms of delay, stops, bandwidth efficiency, or other measures of effectiveness.

The most recent entry in the field of traffic analysis models is the 1985 Highway Capacity Manual (HCM) technique for determining delay at signalized intersections (1). This technique recognizes the axiom that delay at any given signal is influenced by the quality of traffic progression from its neighbors. A progression adjustment factor,  $PF$ , is given in table 9-13 of the HCM. The  $PF$  is a scalar multiplier which increases or decreases the delay as a function of the progression quality, the degree of saturation, and the type of control equipment (pretimed, traffic actuated, etc.). The values contained in HCM table 9-13 range from 0.40 to 1.85, indicating that the quality of progression, as viewed by the HCM, exerts a substantial effect on the delay at a signalized intersection.

Other traffic control system models which have been in use for several years also recognize the effect of the quality of progression on delay. One such model, the Traffic Network Study Tool (TRANSYT) (2), has been used widely in several countries. This paper will compare the TRANSYT and HCM progression modes.

In the following discussion, the computational aspects of the two models will be compared. The results will then be examined on a single pair of links with controlled conditions and on a network with approximately 85 links with varying quality of progression.

The scope of this paper is limited to the effects of progression. No comparisons of the absolute values of delay are appropriate to the methods used in this study. This is simply a comparison of the relative degree of improvement attributable to progression as seen by two different analysis models.

## BACKGROUND

The TRANSYT model was developed initially by Dennis I. Robertson in 1968. Subsequently it has been improved primarily by the Transport and Road Research Laboratory and others in several nations. It has been extensively tested and used throughout the world for design and evaluation of traffic signal timing.

The program has evolved substantially since its original development and several versions have been released. The specific version of TRANSYT used in the study was TRANSYT-7F (3). The HCM has also evolved since its first release in 1950. The 1985 version incorporates significant enhancements over its predecessors, especially in the analysis of traffic signal operations.

## COMPUTATION OF DELAY

There are some important similarities and differences between the TRANSYT and HCM delay models. They are similar in the sense that they both use variations of the general two-component delay model originally proposed by Webster (4). In this model, delay is expressed as the sum of two separate functions:

$$D = d_1 + d_2 \quad (1)$$

where

$D$  = the delay per vehicle (seconds);

$d_1$  = the delay which would result if the traffic volumes were uniform from cycle to cycle; and

TABLE 1 COMPARISON OF TRANSYT AND HCM DELAY MODELS

Delay element	TRANSYT	HCM
Uniform Delay ( $d_1$ )	Arrivals are projected on each step of the cycle (may 60 steps per cycle) from the previous intersection. Queues are stored on the red and released on the green to be projected to the next signal on a step by step basis.	The average delay per vehicle is computed based on the assumption of uniform arrivals. No consideration is given to adjacent signals.
Random and saturation delay $d_2$	An empirically derived formula is applied. The formulation differs somewhat between the two models. The degree of saturation exerts the strongest influence in both cases.	
Quantification of progression quality	There is no progression adjustment in TRANSYT. The detailed treatment of arrivals in the uniform delay computation accounts for the effect of progression.	Progression quality is grouped into one of five categories. Assessment may be subjective or based on the observed proportion of vehicles arriving on the green.
Progression adjustment (PF)		An adjustment factor is determined by table look up based on the arrival category and the degree of saturation.

$d_2$  = the additional delay which results from variability of volumes throughout the analysis period. This is generally referred to as the "random and saturation" delay.

On the other hand, the two models differ in the way that both terms are derived and applied. A detailed comparison of the computational aspects of the two models is given in table 1. The main difference evident from this comparison is that TRANSYT accounts for the effect of progression by dividing the cycle into as many as 60 equal time steps and performing a discrete analysis of traffic flow for each time step. The HCM, on the other hand, makes the original computations with no consideration of progression, then performs a final progression adjustment based primarily on arrival type and degree of saturation.

The determination of the arrival type requires some further consideration. The arrival type is the sole measure of progression quality. It must be assigned a value between 1 and 5. Higher values indicate better progression. The middle of the range (*i.e.*, Type 3) indicates the neutral condition resulting from random (or uniform) arrivals.

Progression quality is difficult to assess subjectively. The HCM provides some guidance here in the form of a "platoon ratio,"  $R_p$ , which reflects the proportion of vehicles arriving on the green relative to the proportion of green time given to the approach.

Table 9-2 in the HCM suggests a relationship between platoon ratio and the arrival type. The platoon ratio is defined in the HCM as:

$$R_p = PVG/PTG \quad (2)$$

where  $PVG$  is percentage of vehicles arriving during the (effective) green; and  $PTG$  is percentage of the cycle that is green for this movement.

Since the comparisons between these two models must be made on a quantitative basis, the platoon ratio will be used as an indication of arrival type for purposes of this paper.

## ANALYSIS PROCEDURE

Both the platoon ratio and the progression adjustment factor must be obtained from TRANSYT before any comparisons may be made with the HCM procedure. Neither of these items are direct outputs of the TRANSYT program. The derivation of both quantities required some innovative applications of TRANSYT combined with some external programming for data reduction purposes. In neither case was the TRANSYT model modified in any way. Instead, maximum use was made of the graphics data file (GDF) produced by TRANSYT-7F for analysis purposes. The graphics data file is described in Appendix D of the *TRANSYT-7F User's Manual* (3).

The complete analysis procedure is illustrated in figure 1, which shows the data flow through the various computational steps. To compare the effect of progression in TRANSYT, it is necessary to have two TRANSYT runs which are identical in all respects, except that one of the runs must have the progression linkages established and the other must have them removed. This is accomplished by a "delinking" process which will be described later.

TRANSYT outputs, in the form of GDFs, were obtained for both conditions of progression (linked and delinked). At

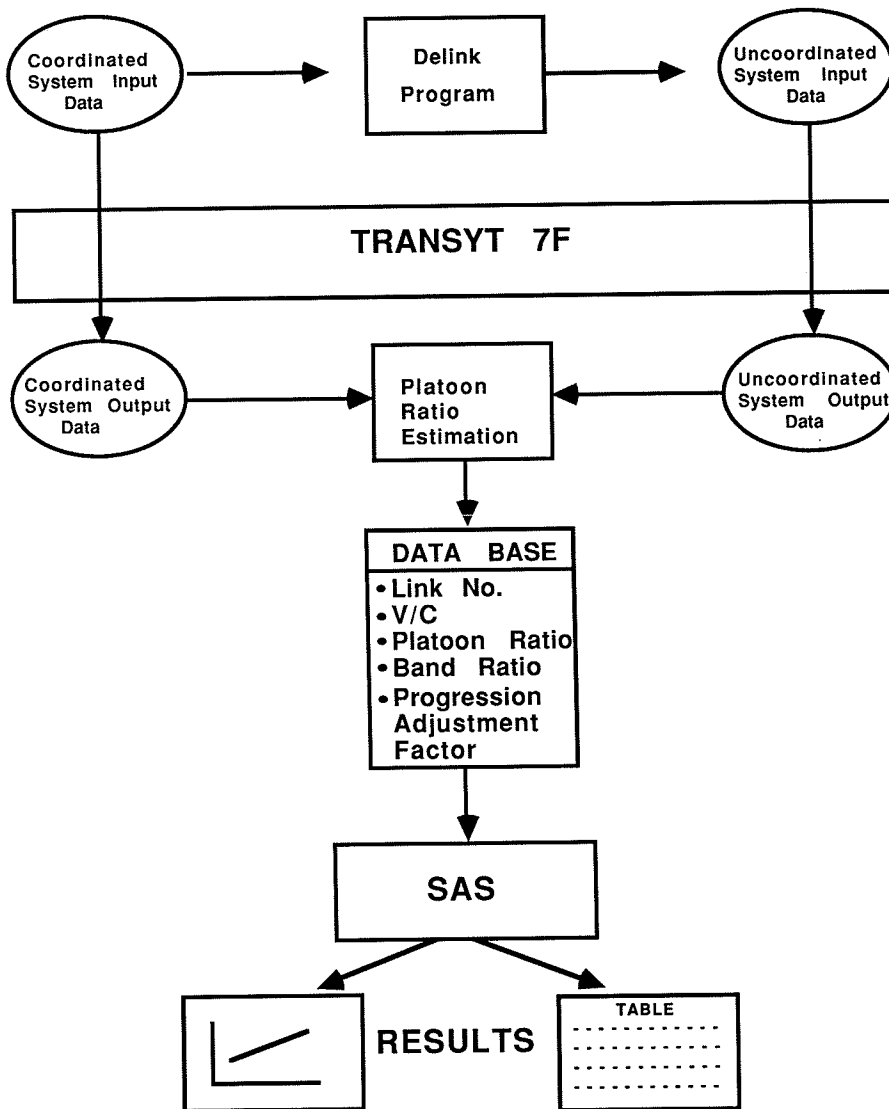


FIGURE 1 Data reduction and analysis procedure.

this point, a separate data reduction program [a modified version of the Platoon Progression Diagram (PPD) program, which is described in Appendix I of the *TRANSYT-7F User's Manual* (3)] was run to determine the platoon ratios by simply accumulating the arrivals on the red and green phases on a step-by-step basis. This information was obtained from the stopline flow profile data contained in the GDF.

Since the GDF also contains delay information for each link, the progression adjustment factor was easily determined by dividing the computed delay for the linked operation by the computed delay for the delinked operation. Similarly, the other independent variables, degree of saturation, was obtained directly from the GDF. All of the data items generated by the data reduction program were placed in a data base for analysis by Statistical Analysis System (SAS) (5) to produce the results which will be presented later.

**THE BAND RATIO**

The platoon ratio,  $R_p$ , is essentially a field measurement. This places some limits on its value as an element of capacity

analysis because it is possible to measure only for existing conditions. The platoon ratio is also costly and time consuming to measure. An alternative measure which could be applied without specialized field studies would be a definite asset.

The time-space diagram (TSD) provides a good starting point for the derivation of a progression quality measure because the TSD represents progression quality graphically in terms of the relative widths of progression bands.

Assume for the moment that traffic approaches a signal with one of two platoon densities which are represented by the relative proportion of vehicles entering at the upstream signal on the artery and on the cross street. Then:

$P_a$  = the proportion of traffic entering upstream from the artery, and

$1 - P_a$  = the proportion of traffic entering upstream from the cross street.

With this simplifying assumption, the TSD for a single link would appear as shown in figure 2. The arrivals on the green

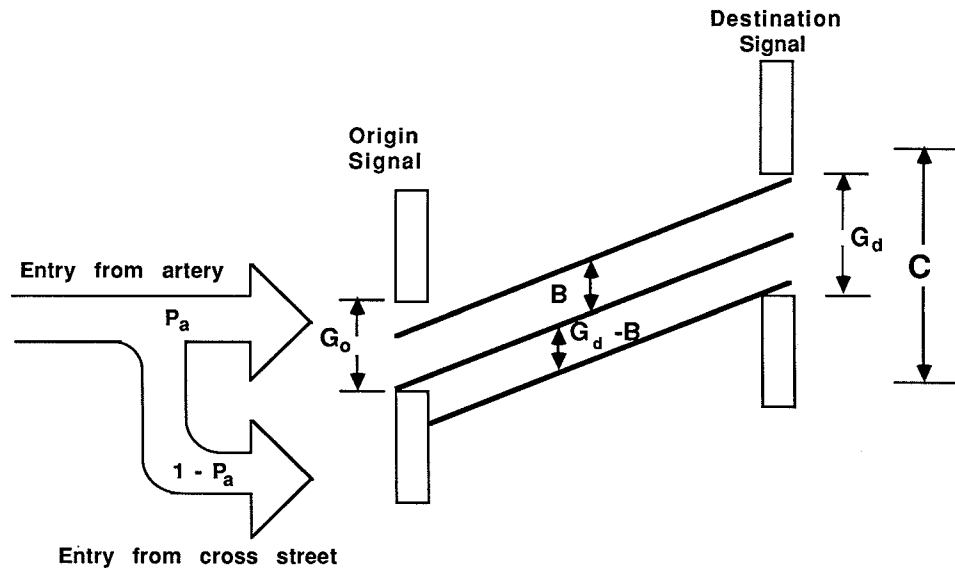


FIGURE 2 Time-space diagram of a single link with progression.

phase at the downstream intersection would be at relative density  $P_a$  within the band and at relative density  $1 - P_a$  outside of the band. Now, let:

- $C$  = the cycle length,
- $B$  = the band width,
- $G_o$  = the green time at the origin signal and
- $G_d$  = the green time at the destination signal.

Then the proportion of vehicles arriving on the green at the downstream signal within the band will be:

$$P_1 = P_a \cdot \frac{B}{G_o} \quad (3)$$

The proportion of vehicles arriving on the green at the downstream signal outside of the band will be:

$$P_2 = (1 - P_a) \frac{G_d - B}{C - G_o} \quad (4)$$

So the total proportion of vehicles arriving on the green at the downstream signal will be the sum of the two proportions just computed.

Now, the platoon ratio is defined by the HCM as the proportion of arrivals on the green relative to the proportion of green time available. So, an estimator of the platoon ratio would be given as:

$$R_b = \frac{PVG}{PTG} = \frac{P_1 + P_2}{G_d/C} \\ = \frac{C}{G_d} \left[ \frac{P_a B}{G_o} + \frac{(1 - P_a)(G_d - B)}{C - G_o} \right] \quad (5)$$

$R_b$  will be called the band ratio for the remainder of this discussion.

One of the objectives of this paper will be to determine how well the band ratio serves as a quantitative indication of the quality of progression. In other words, is it possible to make a realistic assessment of arrival types given only the traffic volumes and the time-space diagram?

## THE DELINKING PROCESS

TRANSYT's traffic simulation model has gained a well-deserved reputation for the realism with which it macroscopically models traffic flow in a coordinated system. The macroscopic model, although not as ultimately realistic as a stochastic microscopic simulation model, is necessary to TRANSYT because it is used identically in the optimization process.

The value of the model is primarily due to the propagation of traffic from multiple upstream sources (links) to downstream movements (also links) and the dispersion of traffic from link to link. The user establishes the link-to-link relationships through data inputs. Specifically, for each interior link, at least one, and up to four, upstream links are identified as source, or feeder links to the current link. Indeed, it is this link-to-link relationship, which "coordinates" adjacent signalized intersections.

Coordinated link flows propagate along the assigned "paths" passing through the platoon dispersion model. The relative position of the green phase at the downstream intersection affects the number of vehicles queued, and thus delayed.

A way of approximating uncoordinated operation is to "delink" the link-to-link relationships. This is easily accomplished by simply deleting the upstream input link number(s) and volume(s) from the link data cards in the TRANSYT input file. The length and primary link speed are retained to enable measures of effectiveness (MOEs) such as total travel, total travel time, and fuel consumption to be comparable.

The delinking process was accomplished in this study using the DELINK program. DELINK is quite simple to use. The network is coded normally with the link-to-link connections in place. DELINK locates all links at each intersection to be delinked as well as the inputs of that intersection to its neighbors and removes the data from the fields which represent the input links. For those familiar with the TRANSYT-7F coding scheme, the values in fields 7, 8, and 10-15 of the link data card (type 28) are deleted. The speed in field 9 is retained.

The DELINK program was used in this study to transform

many "coordinated" links, with platooned arrivals randomized by link, to completely randomized arrivals on all links. This is analogous to many links having arrival types varying randomly from 1 to 5, to a fixed index of 3 on all links.

The DELINK program has been incorporated, along with a number of other useful programs, into a "Signal Utility Package" (SIGUTIL) available from the McTrans Center.

**STUDY RESULTS**

This study addresses two specific questions:

1. How does the progression adjustment factor (*PF*) computed by TRANSYT compare with the *PF* computed by the HCM based on platoon ratios and volume/capacity ratios estimated by TRANSYT; and
2. How well does the band ratio,  $R_b$ , proposed earlier in this paper, serve as an estimator of the platoon ratio,  $R_p$ ?

The first question will be addressed using a single link pair with controlled volumes and offsets to produce the full range of simulated conditions. The second question will use a more extensive network which was analyzed by TRANSYT using input data from the field.

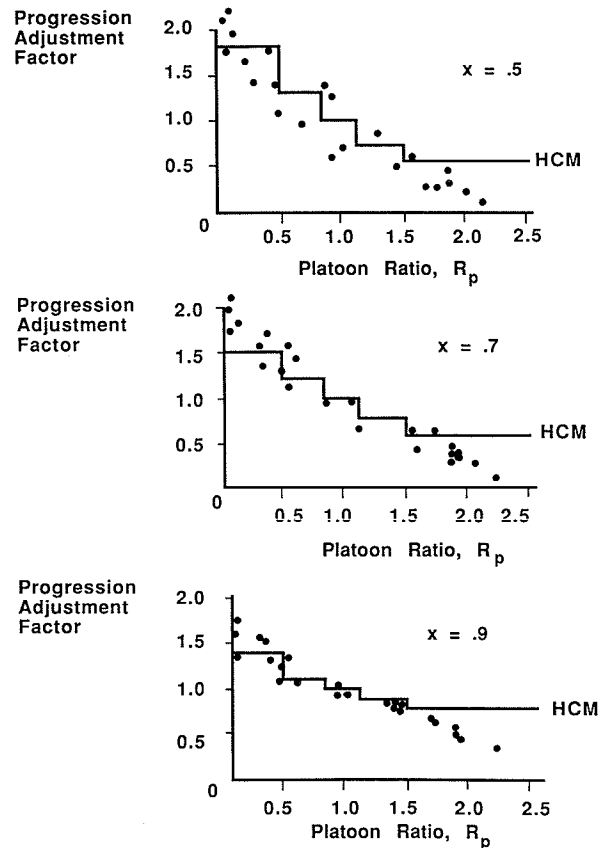
**The Single Link Pair Study**

A pair of links was created hypothetically using two interconnected signals. The TRANSYT runs were made for three traffic volume levels representing 50, 70, and 90 percent saturation. There were no turning movements. These represent mid-range values for each of the three degrees of saturation represented in the progression adjustment factor table in the HCM. The signal timing was based on a 60-second cycle with 50 percent green time. The controller offset was varied by 5-second intervals throughout the cycle. The forward and reverse direction links were given unequal lengths for added variability in the data.

By this method, 72 observations ( $12 \times 2 \times 3$ ) were created for platoon ratio, band ratio, coordinated delay, and uncoordinated delay. The progression adjustment factor, *PF*, was determined for each observation by dividing the coordinated delay by the uncoordinated delay.

The results of this study are shown in figure 3. The observations of progression factor are plotted against platoon ratio separately for each of the three saturation levels. The progression factors computed from the HCM are shown on each plot. The HCM progression factors were obtained from HCM table 9-13 as a function of arrival type estimated from the platoon ratios using HCM table 9-2. Since TRANSYT models pretimed control explicitly, the pretimed control section of HCM table 9-13 was used to determine the HCM progression factors.

There are three observations which stand out clearly on figure 3. The first is that there is excellent general agreement between the progression factors computed by the HCM and by TRANSYT for each of the three saturation levels. The "staircase" function of the HCM method provides a very close visual fit to the data points which were obtained from the



**FIGURE 3** Platoon ratio versus progression adjustment factor for pretimed control.

TRANSYT runs. This finding lends credibility to both methods, since the two were developed independently.

A second look at figure 3 suggests that the saturation level affects the predictability of the progression adjustment factor. At high levels of saturation, as indicated by the plot for  $X = .9$ , the data points adhere very closely to the line which represents the HCM results. At low saturation levels, represented by the plot for  $X = .5$ , the data points still follow the HCM function, but with much more dispersion. At the saturation level represented by  $X = .7$  the dispersion is clearly between the two extremes. This suggests that the platoon ratio is a better predictor of the quality of progression at higher levels of saturation.

The third observation apparent from figure 3 is that the range of progression adjustment factors computed by TRANSYT exceeds the range specified in the HCM at both ends of the scale. This suggests that, from TRANSYT's point of view, extremely good progression will reduce delay by a greater amount than indicated by the HCM. It also suggests that extremely bad progression will cause more delay than the HCM would predict.

The implications of this observation are most important in the area of good progression. The minimum value of the progression factor in HCM table 9-13 for pretimed control is 0.53. TRANSYT, on the other hand, estimated values much lower than this when progression was "perfect." Because of this study's controlled conditions, it was possible to achieve better progression in the computer than would normally be

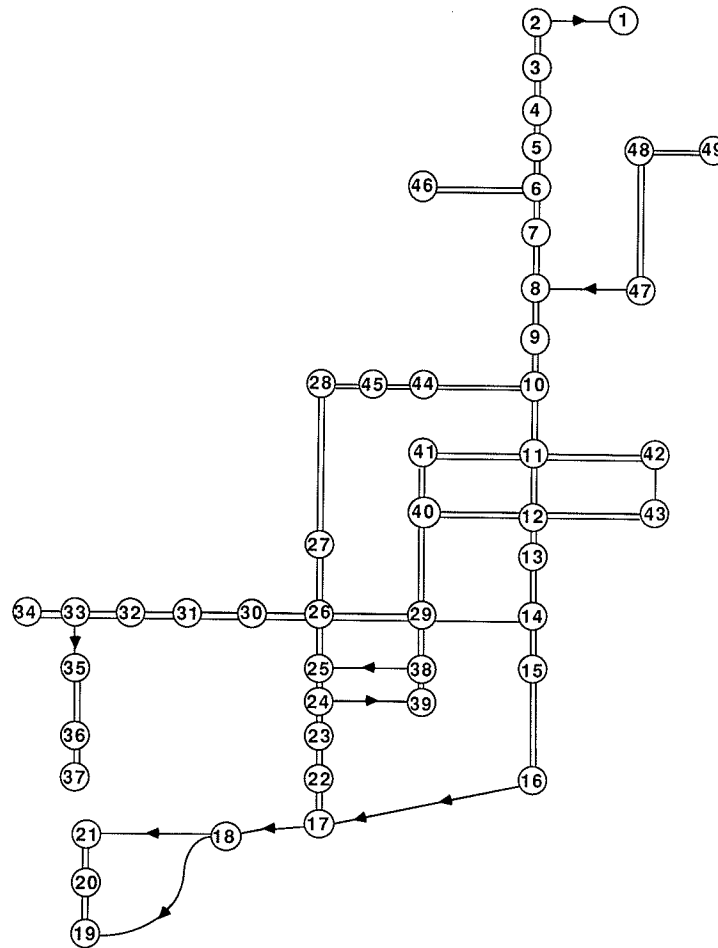


FIGURE 4 TRANSYT-7F network configuration for progression delay study.

observed in the field. However, it is reasonable to conclude from these results that when truly "ideal" progression exists in the field, for example at intersections with extremely short spacing and no significant entry from cross-street turning movements, then the delay is probably being overestimated by the current HCM method.

#### Network Study Description

While highly controlled conditions were appropriate for the previous analysis, an actual network with field data can better assess the value of the quantitative relationships between the variables. The network chosen for this study contains 49 intersections and 85 links as figure 4 illustrates. The information was obtained from an available TRANSYT data set coded for Port Huron, Michigan. The network configuration and signal phasing reflect actual field conditions. The traffic volumes were, however, increased selectively to create a wider variety of saturation levels. The controller offsets were manipulated artificially also to ensure a wide range of progression quality. Twenty-four TRANSYT runs were performed on this network with the controller offsets established randomly. By this process, 2,040 separate observations were generated. The degree of saturation ( $v/c$  ratio) ranged from .06 to .93, with

a mean value of .63. The progression quality, as indicated by the PF, ranged from .22 to 2.81, with a mean value of 1.06. This suggests that the randomized controller offsets have produced neutral progression on the average.

#### Platoon Ratio vs. Band Ratio

Each of the 2,040 observations included a platoon ratio,  $R_p$ , and a band ratio,  $R_b$ . To assess the value of  $R_b$  as an estimator of  $R_p$ , these two quantities were plotted and a regression analysis was performed.

The results are presented in figure 5. The relationship between the two progression quality indicators is plotted to illustrate both the central tendency and the dispersion inherent in this relationship. It is quite apparent from figure 5 that a strong relationship exists. A visual inspection suggests a linear relationship with a 1:1 slope and zero intercept. The equation obtained by linear regression is:

$$R_p = 0.99R_b + 0.012 \quad (6)$$

The correlation coefficient,  $r^2$ , for this model is 0.55, indicating that approximately 55 percent of the variation may be explained by the model. This is not exactly a deterministic relationship; however, considering the relative ease with which

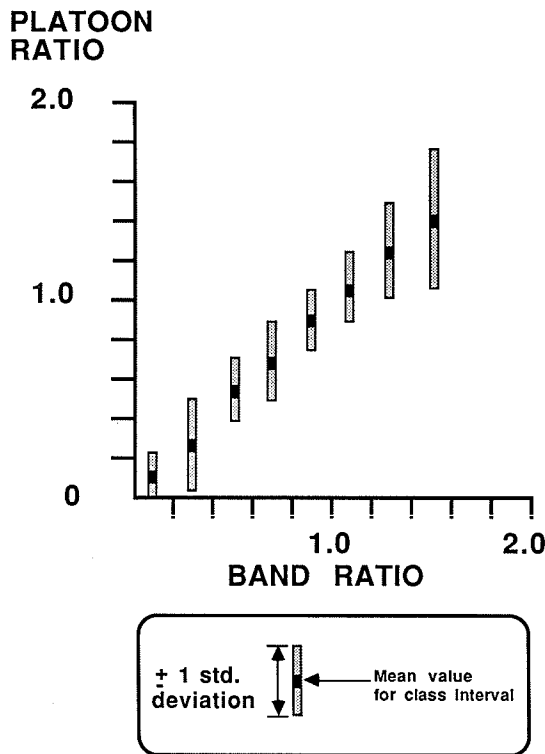


FIGURE 5 Platoon ratio versus band ratio as computed by TRANSYT.

the band ratio may be determined compared to the platoon ratio (which requires field studies), it is reasonable to conclude that the band ratio could be used as a cost-effective substitute for the platoon ratio for most purposes.

**Progression Factor Comparison**

The progression factor for each observation was computed in two ways. The ratio of linked delay to delinked delay gave the progression factor as computed by TRANSYT. The HCM progression factor was determined from HCM table 9-13, based on the value of the platoon ratio (also a TRANSYT computation). The error between the two values was defined by

$$E = \frac{PF_T - PF_H}{PF_T + PF_H} \times 200$$

where

$E$  = percent error referenced to the average value of the two estimates,

$PF_T$  = progression factor computed by TRANSYT, and

$PF_H$  = progression factor computed by HCM.

Figure 6 illustrates the distribution of the error. The mean error for the entire sample was 2.36 percent, which indicates a very small average discrepancy between the two methods. The standard deviation, on the other hand, was 29 percent, indicating that individual errors were sometimes substantial.

Neither the degree of saturation nor the progression quality showed a significant effect on the magnitude of the error.

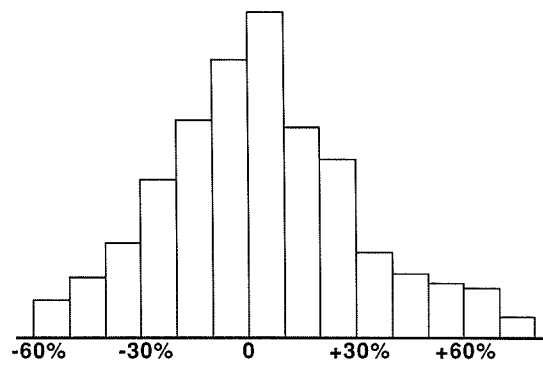


FIGURE 6 Distribution of error in progression factor computations.

**Progression Factor vs. Platoon Ratio**

The comparison of the platoon ratio and the progression factor was originally carried out for a single link pair and illustrated in figure 3. A more quantitative study of this relationship was performed on the network data. The results are shown in figure 7, which shows the mean and standard deviation of the progression factor computed by TRANSYT as a function of the platoon ratio. The regression equation was

$$PF = 2.19 - 1.14 R_p$$

The correlation coefficient,  $r^2$ , was .51. These results suggest that the platoon ratio is indeed an indicator of the quality of progression in a coordinated signal system, as suggested by the HCM.

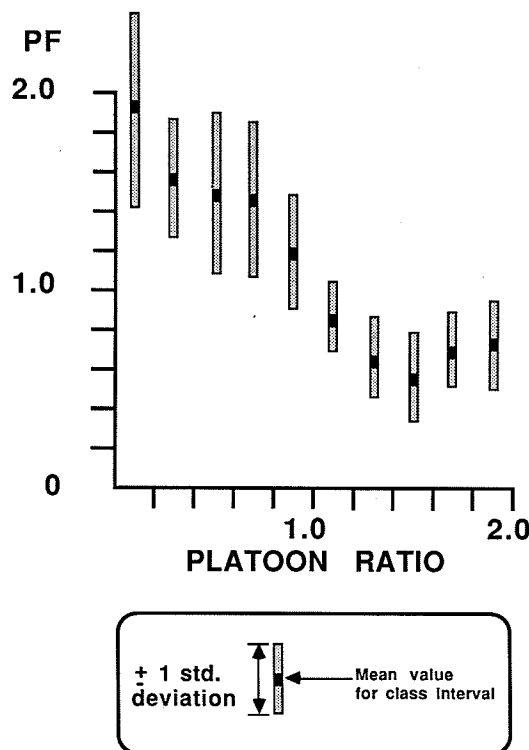


FIGURE 7 Progression factor versus platoon ratio as computed by TRANSYT.

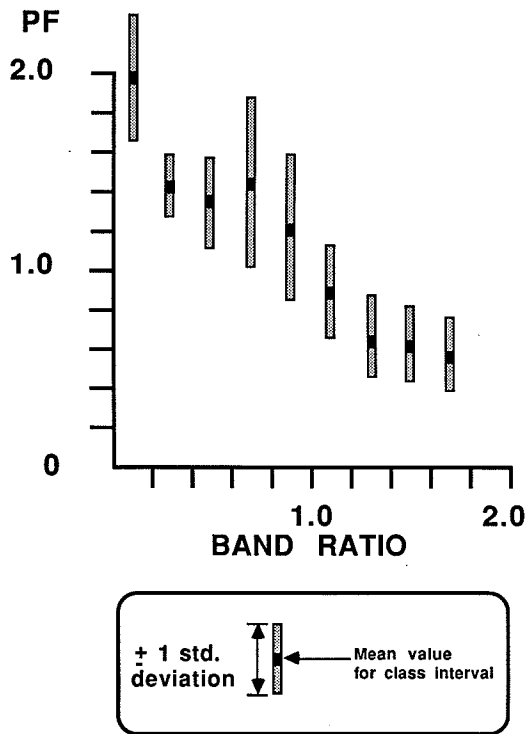


FIGURE 8 Progression factor versus band ratio as computed by TRANSYT.

#### Progression Factor vs. Band Ratio

It has already been established that the platoon ratio and the band ratio are strongly correlated. Therefore, it is a reasonable hypothesis that the band ratio could function as a practical surrogate for the platoon ratio in quantifying the arrival type. The relationship between progression factor and band ratio is shown in figure 8, which has the same format as the progression factor-platoon ratio relationship shown in figure 7. On the surface, the two relationships appear to be very similar. The regression equation for the band ratio is

$$PF = 2.26 - 1.20 R_b$$

which agrees very closely with the platoon ratio equation. The correlation coefficient was .32 compared to a value of .51 for the platoon ratio equation. This suggests that, while field measurement of platoon ratio provides a more reliable indicator of progression quality, the band ratio should offer an acceptable substitute when field data are not available.

A further test of the validity of the band ratio was performed by comparing the progression factor estimated from HCM table 9-13 using the computed values of the platoon ratio and the band ratio. The band ratio produced the same estimated progression factor as the platoon ratio in 69 percent of the cases. Of the remaining 31 percent, the band ratio produced closer agreement with the TRANSYT progression

factor in 19 percent of the cases, and the platoon ratio produced better agreement in the final 12 percent. This offers further support for the band ratio as an estimator of progression quality.

#### CONCLUSIONS

Within the limitations of this study, the following conclusions are offered.

There appears to be excellent general agreement between the HCM method and the TRANSYT model on the effect of the arrival type (as measured by the platoon ratio) on the progression adjustment factor. The average discrepancy between the two models was extremely small, although larger differences were observed on individual data points. The platoon ratio is a better predictor of the arrival type when the saturation level is high. The TRANSYT model suggested that some extrapolation of the current values in HCM table 9-13 may be desirable to account for extremely good and bad progression.

The band ratio,  $R_b$ , developed in this paper is considerably easier to measure than the platoon ratio,  $R_p$ , defined in the HCM. The platoon ratio is based on field measurements, whereas the band ratio is based on simple bandwidth ratios taken from the time-space diagram. A comparison of these two measures suggests that the band ratio is an adequate predictor of the platoon ratio for most purposes. It is highly cost effective and provides estimates of the quality of progression for hypothetical situations where the platoon ratio cannot be measured.

It is important to remember that the data presented in this paper are the result of a modeling application and do not represent actual field observations of the relationships which are reported. They are, however, based on a model which has been extensively used and accepted throughout the world. The agreement between TRANSYT and the HCM reported herein should be considered as a "plus" for both techniques.

#### REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
2. D. I. Robertson. *TRANSYT: A Traffic Network Study Tool*. Road Research Laboratory Report LR 253. Ministry of Transportation.
3. C. E. Wallace, K. G. Courage, D. P. Reaves, G. W. Schoene, G. W. Euler, and A. Wilbur. *TRANSYT-7F User's Manual*. FHWA, U.S. Department of Transportation, Washington, D.C., Revised July 1987.
4. F. V. Webster and B. M. Cobbe. *Traffic Signals*. Road Research Technical Paper No. 56. London, 1966.
5. *SAS Program and Documentation*. SAS Institute, Cary, N.C. 1985.



# Validation of Saturation Flows and Progression Factors for Traffic-Actuated Signals

PANO D. PREVEDOUROS AND PAUL P. JOVANIS

A need for empirical validation of the 1985 Highway Capacity Manual (HCM) arises from its development with limited field data and its adoption of radically new procedures. Chapter 9 of the Manual, Signalized Intersections, contains some of the more extensive changes and is the subject of field validation. Emphasis is placed on comparisons of field saturation flows, delays, and progression factors with those estimated by the manual. Data are obtained from ten intersection approaches for twenty-five 15-minute analysis periods. Saturation flows in the field are significantly higher than those in the manual. A value of nearly 2,000 vehicles per hour of green per lane (vphgpl) is consistently observed at field sites where the HCM estimates values of 1,800 passenger cars per hour of green per lane. The difference in saturation flows significantly affects delay estimates: the delays predicted using the value of 2,000 more closely match field delays than do the estimates using 1,800. Our conclusion is that an ideal saturation flow equal to 2,000 (rather than the HCM value of 1,800) is recommended for the analysis of high-design intersections similar with those analyzed in this project. The progression factors estimated from our data are statistically significant and differ significantly from HCM values. Our platoon factors are lower than the HCM, indicating that greater reductions in delay are necessary than are currently provided in the progression factor adjustments. In addition to being lower than current HCM values, our progression factors are much less sensitive to the platoon ratio (a relative measure of the percent arrivals on green). It appears that our field data for actuated controllers is almost totally inconsistent with the current HCM.

In August 1985, the Transportation Research Board published a revised Highway Capacity Manual (HCM) (1). This third edition of the manual contains significant revisions to the previous manual (2) and several entirely new procedures. Many of the most significant changes can be found in Chapter 9, "Signalized Intersections."

Signalized intersections are among the most complex elements in a traffic system. This complexity has, in turn, necessitated a complex methodology to conduct a proper analysis of signalized intersections. This methodology can be used to assess level of service relationships for both existing intersections and proposed designs.

HCM Chapter 9 defines capacity as a function of the saturation flow(s) and the green split ( $g/C$ ). The level of service is a function of the average stopped delay per vehicle. The methodology employs a series of worksheets and tables to

arrive at delays and the corresponding levels of service. One of the key variables in this process is the percentage of all vehicles in the movement arriving during the green phase; this determines the so-called progression factor. The progression factor, as well as the degree of saturation, has the largest effect on level of service.

A detailed literature review was conducted of research related to traffic signal delay estimation and progression (3). While there are many simulation-based approaches to delay estimation—including TRANSYT-7F, PASSER II-84, and others—there are virtually no other procedures available to estimate delay in the variety of conditions actually observed in the field. Previous research in Kentucky (4) and a more recent study by a consultant (5) are the only published attempts to measure saturation flows in the field, other than the measurements at some 40 intersections that were undertaken as part of the formative research on the 1985 HCM. With the exception of ongoing research at the Texas Transportation Institute (6), no attempt has been made to validate the progression factors in the 1985 HCM.

There is a clear need for empirical validation of the 1985 HCM, particularly the chapter on signalized intersections and most particularly on delay estimation. This research attempts to respond to that need.

## STUDY OBJECTIVES

The overall goal of the research which forms the basis of this paper is the evaluation of the fit of the 1985 Highway Capacity Manual to the analysis of signalized intersections in Illinois (3). The following specific objectives are undertaken to achieve this goal.

The first objective is the investigation of the difference between the field-measured and the HCM-estimated saturation flows. The saturation flow is an important component of the volume to capacity ratio, which strongly influences predicted delays. The implications of differences in saturation flows are explored by comparing delays obtained with the field-measured and HCM-estimated saturation flows.

The second objective is the identification of changes, if necessary, in the progression factor, to better reflect the effect on delay of actuated signal system progression. The progression adjustment is one of the most important factors in Chapter 9 of the HCM because it can reduce the estimated delay by 50 percent or can increase it by 100 percent.

## STUDY DESIGN

The design sought to explicitly control for factors that were exogenous to the main study objectives. Rather than conduct a very broad study of saturation flow adjustments, an early decision was made to narrow the study scope to a limited but frequently occurring condition. To compare field measured and model estimates of saturation flows, we chose intersection approaches with 12-foot lane widths, left turn lane, no local buses, no parking, right turn channelization, and flat grades. These criteria controlled for all adjustments to saturation flow, except the percentage of heavy vehicles. Thus our observations of saturation flow are clustered around a narrow range of conditions but ones that occur frequently in practice.

The study of progression adjustment factors is limited to through lanes at these high-design intersections. These intersections are frequently controlled by traffic-actuated equipment, therefore, we believed that identification of sites for data collection would be comparatively easy. By examining through lanes only, a much more precise analysis could be conducted of the effect of progression.

In sum, the authors adopted the approach of studying a limited set of problems at great depth rather than a broader but less detailed assessment of many factors. The joint objective of studying the progression adjustment factors and accuracy of lane saturation flow estimates led to a study design that emphasized high-design, suburban-type intersection approaches. The approaches are excellent test sites for progression adjustment studies and also offer the availability of measured saturation flow data. This judicious selection of project scope allows us to gain the most information from each site visit.

## DATA COLLECTION

### Required Data

The data required to meet the analysis objectives of the research are summarized in Table 1. The first seven data items are needed as input to the signalized intersection procedure in the 1985 HCM. We have chosen to count arrivals by cycle according to the method proposed by Berry (unpublished data). Once the cycle-by-cycle counts are obtained, they can easily be aggregated to 15-minute flows.

Data on geometrics, activity pattern, and traffic composition are needed to develop adjustments to saturation flow. By using a careful study design and site selection, all of these adjustment factors equal 1.0 except for the heavy vehicle adjustment factor, which is typically very close to 1.0 for the study's sites.

The duration of green and cycle length are required signal timing parameters. Because the study sites have actuated signal control, these data must also be collected each cycle. It is expected that both green times and cycle lengths will vary significantly at each site because of the control equipment's response to traffic fluctuations.

The last two sets of data are needed to evaluate the HCM procedure. Elapsed discharge times are used to measure saturation flow in the field. These measured values are compared to HCM estimates for each site. Stopped vehicle counts are needed to provide field delay data for comparison with HCM estimates of delay. A 20-second sampling rate seemed reasonable for the cycle lengths that were observed at the sites (typically 120 seconds).

A potential problem in the counting of arrivals is the left

TABLE 1 DATA COLLECTED DURING SITE VISITS

DATA	PURPOSE
1. VOLUME COUNTS: the number of arrivals per cycle.	Obtain the peak hour, the peak 15 minutes and the peak hour factor (PHF) for the analysis of the intersection and the determination of its performance (LOS).
2. GEOMETRICS: lane widths, grades.	Obtain the $f_w$ and $f_g$ coefficient for the saturation flow.
3. ACTIVITY PATTERN : CBD or non CBD location, parking activities, bus stops.	Obtain the $f_a$ , $f_p$ and $f_{bb}$ coefficient for the saturation flow.
4. TRAFFIC COMPOSITION: number of heavy vehicles in the total of arrivals.	Obtain $f_{HV}$ coefficient for the saturation flow.
5. DURATION OF GREEN	Obtain the $g/C$ ratio to calculate capacity.
6. CYCLE LENGTH	Obtain the $g/C$ ratio to calculate capacity and delay.
7. ARRIVALS DURING GREEN	obtain the platoon ratio ( $R_p$ ) to determine the progression factor.
8. ELAPSED TIMES of discharge the 4th and 10 <sup>th</sup> vehicle in the queue	Obtain the field measured saturation flow.
9. STOPPED-VEHICLE counts every 20 seconds.	Obtain the field measured delay per vehicle.

turning vehicles. Vehicles which arrive at the end of a queue that extends beyond a left turn lane may have the intention to turn left. Because of the lack of space to enter the lane, they cannot have immediate access to the turn lane. These vehicles have to be counted as vehicles demanding through movement service because, as long as they are in the through movement lane, they contribute to the queue length and the amount of delay in the through lane. After the queue starts dissipating during the green, these vehicles may shift to the left turn lane. This movement is captured in the field by direct measurement both of the delay and of the saturation flow after the shift occurs. This is an approximate procedure because the vehicles that shift cause only partial delays to through vehicles; they are not discharged by the through lane. Within the queue, they also result in gaps that can reduce the measured saturation flows. For the approaches considered in this study, the left turn lanes were typically six to ten vehicles long, so the effect of the left turn "traps" is considered minor for both saturation flow and delay.

Another potential problem, which also arose during a session of the 66th TRB Annual Meeting (1987), is the definition and use of the term, "vehicle arrivals during green." Although the definition seems clear, the problem is how to successfully apply it to the variety of conditions observed in the field. We chose to define a vehicle arriving during green as any vehicle that joined the queue during green or passed across the stop-line unimpeded during green.

#### Data Collection Sites

To fulfill the needs of the study design, a certain type of intersection needed to be chosen. The desirable characteristics of the intersection are:

1. Existence of exclusive lane for right turns, so that the right turn movement does not interfere with the through movement;
2. Existence of left turn bay so that the left turn movement will not interfere with the through movement;
3. Existence of significant rush hour volume so that a meaningful peak period can be identified;
4. Existence of actuated signal operation;
5. A non-central business district (CBD) location so that the analysis will not need to be included in the analysis; and
6. Low pedestrian volume in order to have minimum interference both with the traffic and with the field crew.

The field data collection session should be held under mild and dry conditions during normal weekdays. A summary of the ten intersection approaches used for collection is contained in the detailed project report (3).

#### Data Collection Methods

There are two feasible methods to collect the field data; first, by filming the chosen approach; and second, by collecting data manually using a team of individuals. There are advantages and disadvantages associated with each method.

The filming alternative requires a camera and film, as well as some experience in its appropriate location and use. This

approach requires the least number of persons in the field; two persons are sufficient. On the other hand, to obtain a sufficient field of view to estimate delays, an elevated vantage point has to be available. After the films have been developed, significant manpower is required to translate pictures (frames) into data.

The second alternative is the collection of data manually in the field by a team with specific assignments. One member of the team counts the arrivals and the arrivals during the green and also records the heavy vehicles in the traffic. One or two other members count the stopped vehicles in specific time intervals. One member takes the appropriate measures with respect to the saturation flow and the last member counts the cycle lengths and indicates the beginning and end of the green to the volume counter. There could be an additional member assigned to call out the time intervals but this member can be substituted by a tape player. Therefore, the total size of the crew, for data collection from one approach, is four to six persons. Detailed crew assignments are described in reference (3).

The advantages of this method are the flexibility, speed, and convenience which it offers with respect to the positioning in the field and the starting of the data collection session. Additionally, the data obtained are readily available for analysis.

The shortcomings of this method lie in the human factor and the associated potential for errors, especially if the field crew members are inexperienced and/or ill trained. These shortcomings can be largely alleviated with careful training and direct supervision in the field. This alternative was finally chosen as most appropriate, especially considering the tight schedule of the project.

#### Summary

While the data collection methods did not use sophisticated technology, they are carefully managed to reduce errors and assure accuracy. The procedures meet the needs of the validation requirements for the existing procedures in the HCM.

#### DATA ANALYSIS

##### Structure of the Analysis

The structure of the analysis follows generally the statement of objectives from this paper's first section. The accuracy of saturation flow estimates is first assessed along with their implications for delay estimation. The last phase of the analysis is the study of the progression factor and its validity for the conditions observed in the field.

Before presenting specific results, it is useful to discuss several important aspects of actuated signal operation that have affected the analysis. Actuated signalization results in cycle lengths with considerable variation. One must be careful in determining the peak, 15-minute volume because the interval of volume measurement should end at the same time as a signal cycle. This is because all the field measurements of flow and delay are made per cycle. This results in analysis periods that are close to, but not exactly equal to, 15 minutes (e.g., analysis periods of 13.6 to 16.4 minutes). This use of

approximate 15-minute analysis intervals must be made to obtain the most precise comparisons with field observations. Of course, the appropriate adjustments are made to flow data to convert to hourly rates. For the rest of this paper, the term 15-minute analysis periods refers to the approximate 15-minute analysis period. The actual length of the analysis period is explicitly stated only when necessary.

### Saturation Flow Accuracy and Its Implications for Delay Estimation

The estimation of the saturation flow( $s$ ), from field-measured headways indicates that the true saturation flow is considerably higher than the saturation flow suggested in the HCM. The HCM suggests a saturation flow of 1,800 passenger cars per hour of green per lane (pcphgpl) while the average saturation flow across all sites results in a value of 1,995 vehicles per hour of green per lane (vphgpl) (Figure 1). Because the sites have been selected to control for major geometric and surrounding characteristics, it is possible to obtain this average saturation flow value for all six approaches where measurements were taken. We were able to collect discharge rates only at six of the ten sites we visited, because of lack of data

collectors. The confidence interval for this estimate at the 99-percent level of significance is + 98 vphgpl, which indicates that the estimate is reliable and most certainly higher than the value suggested in the HCM. Because of the significant differences in the saturation flows, we conduct several subsequent analyses comparing results for  $s_0 = 1,800$  and  $s_0 = 2,000$ . We have taken the liberty of converting 1,995 vphgpl to 2,000 pcphgpl because nearly all adjustments of saturation flow equal 1.0.

An efficient way to check the implications of alternative values of the saturation flow is to plot field-measured delays and delay estimates using each of the two values. This plot is presented in Figure 2. The diagonal of the graph indicates the ideal line which corresponds to the location of points, had the estimated delays been identical to the field-measured delays. The legend in the middle of the graph explains the points obtained for the two levels of saturation flow ( $s = 1,800$  and  $s = 2,000$ ). The regression lines corresponding to the alternative levels of saturation flow indicate that estimates obtained with  $s = 1,800$  are consistently higher than the field-measured delays. This nearly constant bias of 5 seconds or more exists over the entire range of delay values.

The estimated delays with  $s = 2,000$  are closer to the field-measured ones, especially for delays less than 30 seconds per

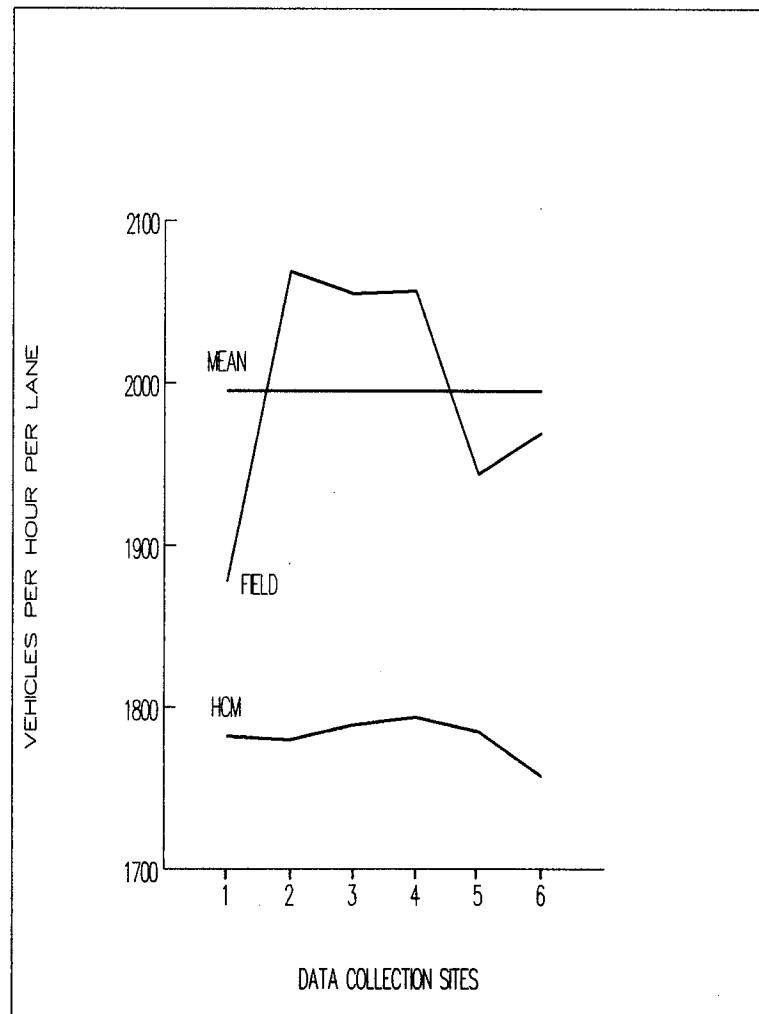


FIGURE 1 Average saturation flow for six sites.

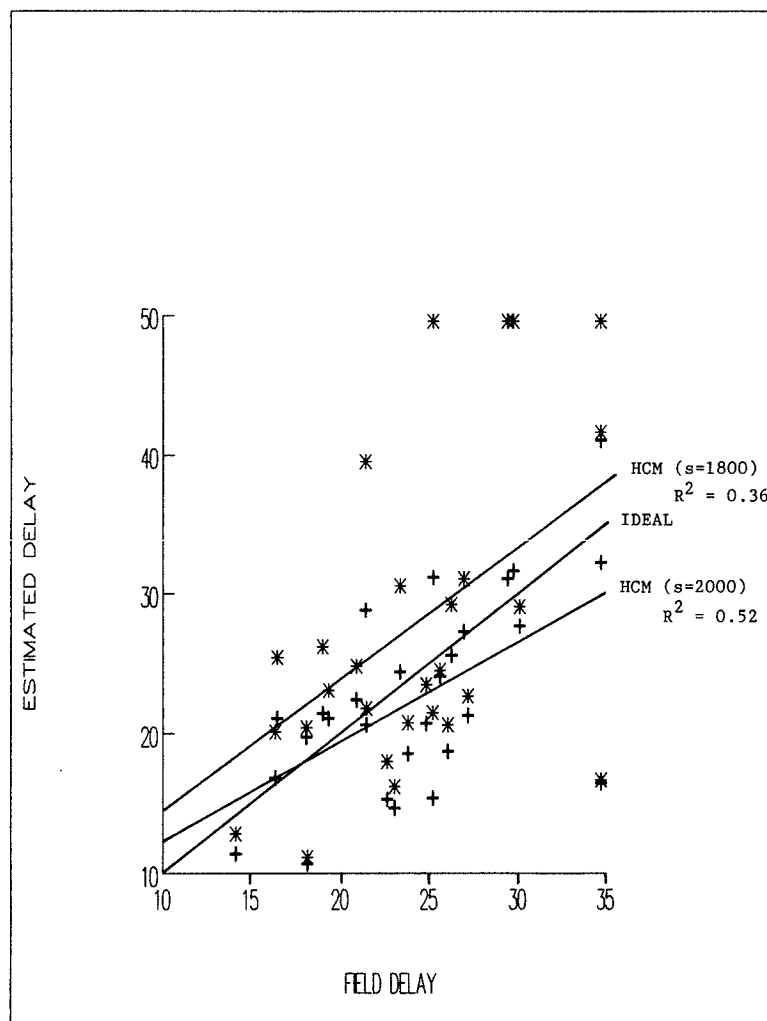


FIGURE 2 Estimated and field measured delays.

vehicle, where most of the data are clustered. There is a slight tendency to underestimate the delays at high delay levels. The goodness of fit ( $R^2$ ) of the line corresponding to the estimates with  $s = 2,000$  is much better than for  $s = 1,800$ . The parameter for the constant term in the regression model is insignificant, indicating a failure to reject the null hypothesis that the line starts from the intersection of the axes.

Further insight into the estimation accuracy can be achieved by investigating the effect of the degree of saturation on the estimated delays, particularly on the error in estimation. The degree of saturation is a major determinant of delay and an indicator of utilization. It is important that the delay model accurately predicts delays at high degrees of saturation, as these are likely to reflect peak period conditions. The level of service assessment under these conditions is very important, as these sites are likely to be candidates for design improvement.

The saturation flow has a direct effect on the degree of saturation; the higher the saturation flow, the lower the resulting degree of saturation. For example, when studying the same approach, the degree of saturation resulting from a saturation flow equal to 2,000 vphgpl will be equal to nine-tenths of the degree of saturation resulting from a saturation flow equal to 1,800 vphgpl.

Let us examine the graphs where the percent estimation error is plotted against the degree of saturation for these two levels of saturation (Figures 3 and 4). The “% ERROR” in estimation is defined as:

$$\% \text{ Error} = 100 \frac{([\text{Estimated delay}] - [\text{Field delay}])}{[\text{Field delay}]}$$

When the saturation flow is equal to 1,800 a large range of errors results (–55 percent to +105 percent). The distribution of errors is rather biased: for low levels of saturation the delays are underestimated while for high levels of saturation the delays are overestimated.

When the saturation flow is equal to 2,000, a much smaller range of errors results (–55 percent to +35 percent). The distribution of errors is unbiased because across the range of the degree of saturation the underestimated delays balance the overestimated ones.

Similar conclusions are reached if one plots the magnitude of the error against the degree of saturation. Errors as large as two levels of service are observed in these figures.

There is conclusive evidence to support the use of a saturation flow value of 2,000 in place of the HCM value of 1,800 for high-design intersections. Because of the significance of this finding, subsequent studies of the progression factor with

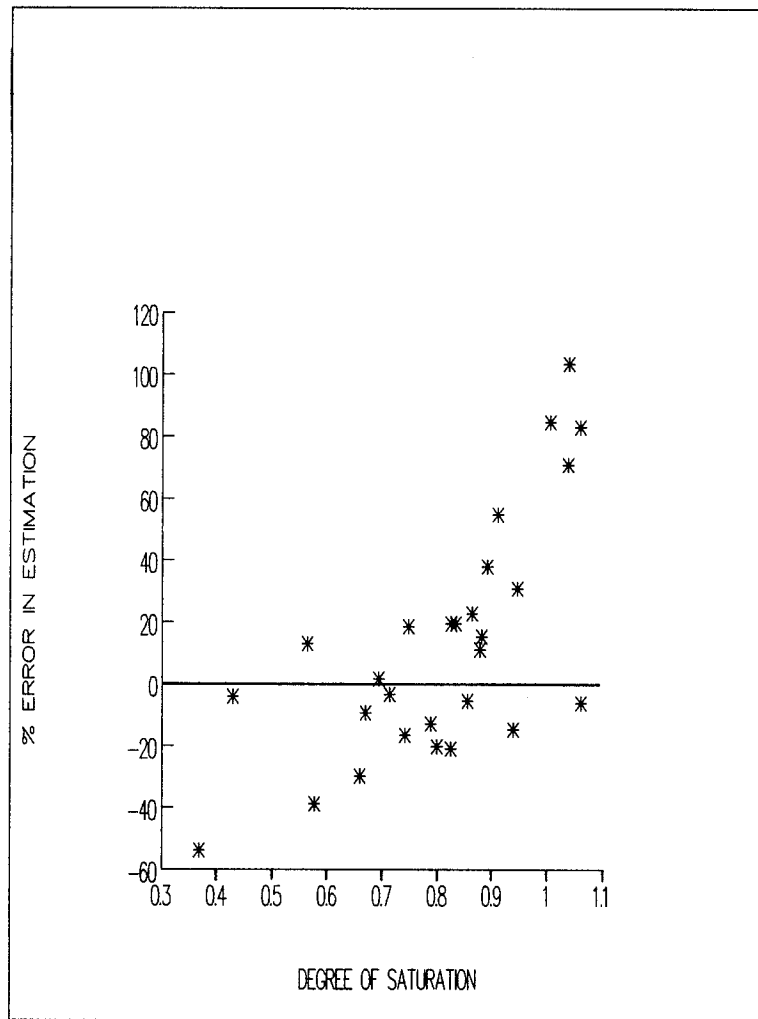


FIGURE 3 Estimation error and saturation ( $s = 1,800$ ).

the more accurate saturation flow value,  $s = 2,000$  vphgpl, were conducted.

### Progression Factor

The progression factor ( $PF$ ) is a factor adjusting the delay to incorporate the benefits of a good progression (*i.e.*, many arrivals during green) and the losses of poor progression (*i.e.*, many arrivals during red). The progression factor has a potentially large effect on delay estimates because it is an adjustment that is applied directly to the delay. Its magnitude can result in halving or nearly doubling the delay for random arrivals. It is the only adjustment factor in HCM Chapter 9 that is applied to delay directly; all other adjustments are inputs to the delay equation, not a final adjustment to it.

To obtain a qualitative comparison of our field data with HCM values, Figures 5 and 6 are constructed. The figure illustrates the delays that result from a given percent of arrivals during green for each 15-minute period. The slope of the lines reflect the effect of platooned arrivals on observed delay (Figure 5) or estimated delay (Figure 6). The authors' interpretation of these data is that the HCM model illustrates a much

more direct relation between platooned arrivals, as expressed by the percent of arrivals during green, and delay than does our field data. Estimates of the progression factor from the study data are expected to differ from those in the HCM. (And they do.)

The authors' approach to the analysis is to create models in which the progression factor is the dependent variable while the degree of saturation, the percent of green, the platoon ratio and the percent of arrivals during green are all candidate independent variables. The estimation equations for the progression factor are derived from HCM equations. The starting point is the adjustment to random arrivals to reflect the effect of platooning:

$$\text{Field delay} = (\text{HCM delay}) PF \quad (1)$$

Solving for  $PF$ , the following is obtained:

$$PF = \frac{\text{Field delay}}{\text{HCM delay}} \quad (2)$$

Equation (2) expresses the progression factor as a function of the field delay, which is known by measurement in the field, and the estimated delay, which is calculated using the HCM method. Equation (2) thus allows us to calculate an

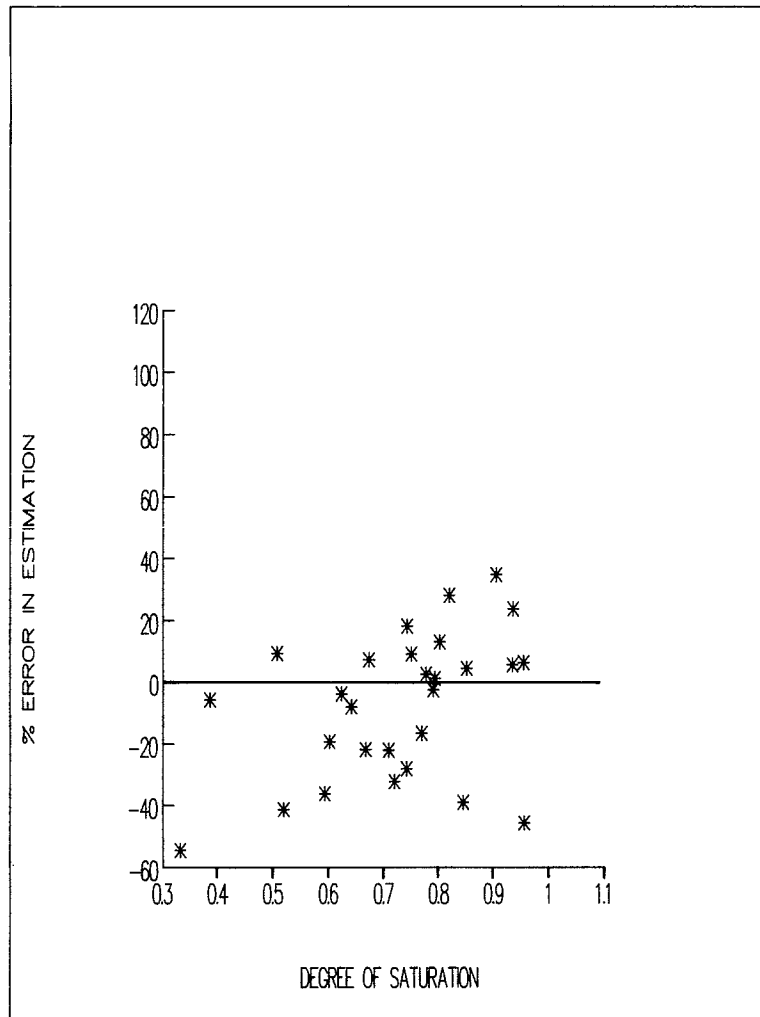


FIGURE 4 Estimation error for two levels of ideal saturation flow ( $s = 2,000$ ).

“observed” progression factor for each 15-minute period. These observed values can then be fit to a variety of models using standard regression techniques.

From the mathematical standpoint, there are several forms of models which can be used to analyze our data. The two forms we select are the linear and multiplicative, which generally are expressed as:

linear:

$$PF = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (3)$$

multiplicative:

$$PF = \alpha_0 X_1^{\alpha_1} X_2^{\alpha_2} \dots X_m^{\alpha_m} \quad (4)$$

The advantage of the multiplicative form is that it cannot result in negative estimates for the dependent variable; the linear model can. A negative value for the progression factor is counterintuitive because negative delays have no physical meaning. We explore both models and determine a preferred model based on the interpretability of parameter estimates and the goodness of fit. We use SPSS on an IBM PC to conduct the estimations. The multiplicative model is esti-

mated by taking the natural logarithm of both sides yielding a linear in parameter specification:

$$\ln PF = \ln \alpha_0 + \alpha_1 \ln X_1 + \alpha_2 \ln X_2 + \dots + \alpha_m \ln X_m \quad (5)$$

Note that both model forms treat the independent variables as continuous, rather than discrete, ranges (as in Table 9-13 in the HCM). This is advantageous because it allows us to estimate one model using all our data. It is valid if we expect the independent variables to have the same effect on the progression factor over its entire range of values. If we do not believe this, a piecewise model for the progression factor can be constructed and model parameters compared to the full model. Constraints on sample size require that we estimate one model over the entire range of the progression factor, leaving piecewise tests for future research.

Considering the theoretical validity of the parameters, let us first discuss our expectation for the sign for the degree of saturation. At low to moderate levels of the degree of saturation, progression is expected to have a very strong effect. If small to moderate platoons consistently arrive on red (or

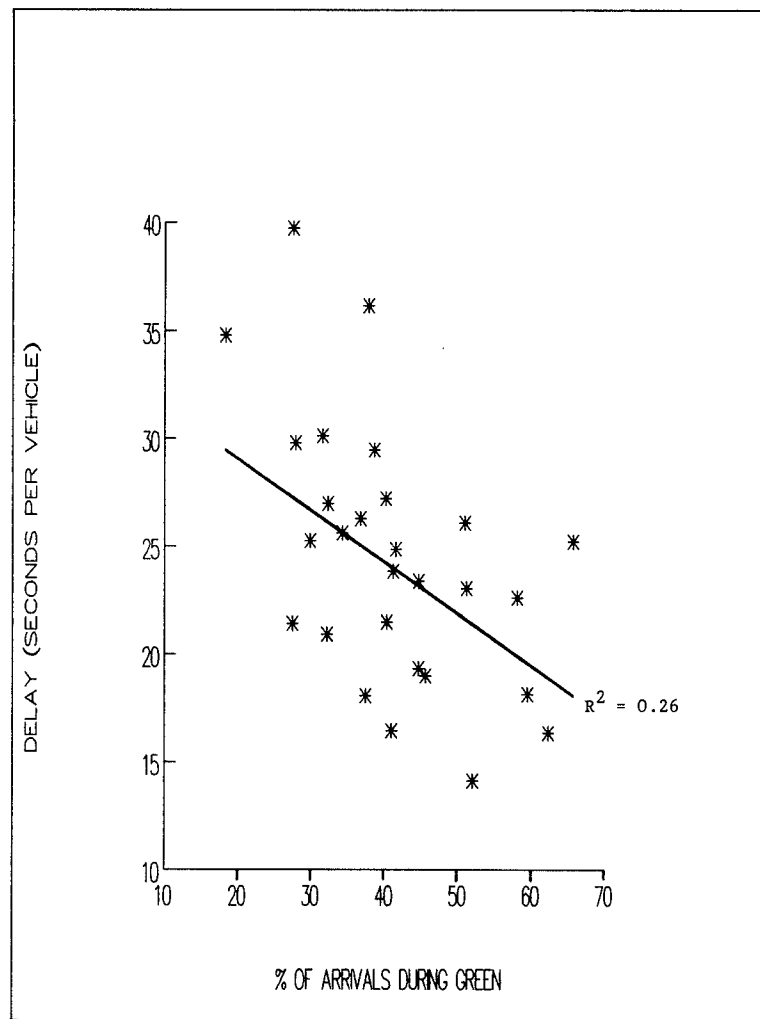


FIGURE 5 Arrivals during green and field delay.

green), significant increases (or decreases) in delay are expected, compared to random arrivals. This effect is expected to be the strongest for coordinated signals that have a fixed time relationship between the beginning and end of green and the time of platoon arrival. Over the span of 15 minutes, any variation in signal offset is likely to weaken the effect of progression.

For high levels of saturation ( $X > 0.9$ ) it is difficult to sustain good progression. Long existing queues might not dissipate totally before arrivals from upstream reach the intersection. In this case, the vehicles stop instead of moving smoothly through the signal. At high levels of saturation, progression is likely to have little or no effect. These considerations suggest that a non-linear specification may be best for the degree of saturation. Given the preliminary nature of our analysis, we leave other specifications for future research.

For the available data set, a negative sign is expected as representing the majority of the field conditions. The parameter may not be strongly significant if much of our field data represent nearly saturated conditions or reflect large changes in platoon size, shape, and timing.

Another potentially strong explanatory variable is the platoon ratio. The platoon ratio, ( $R_p$ ), is defined as the ratio of the percent of vehicles arriving during green over the percent

of time that the signal indication for the analyzed approach is green ( $I$ ). For this variable, a negative sign was expected because higher values indicate better progression and conceivably lower average delay.

In addition to models using degree of saturation and platoon ratio, a number of analyses were conducted using the percent of arrivals during green and the percent of green time as additional predictors. Several difficulties arose with the use of these last two variables. The percent green is collinear with both  $R_p$  and the percent of arrivals during green; therefore, its simultaneous incorporation in a linear model would be erroneous. Models that seek to estimate the progression factor as a function of the percent of arrivals during green and the percent green time had poorer explanatory power than those using the platoon ratio. Therefore, only results using the platoon ratio are included here.

Before proceeding with the results of the validation, plots of the individual major explanatory variables and the progression factor are presented. Figure 7 presents the relation between the field delay and the HCM delay before it is multiplied by the progression factor. There is a clear overestimation of the delays, therefore, the progression factor should be consistently less than 1.0.

Figure 8 indicates that the progression factor is relatively



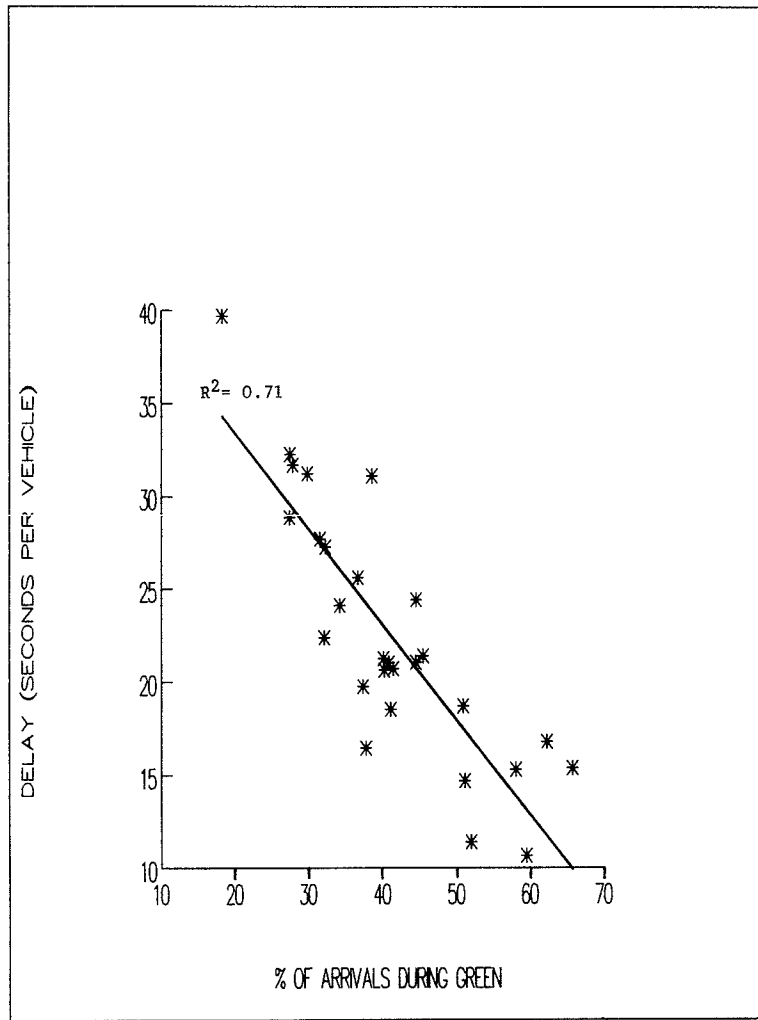


FIGURE 6 Relation of the field measured and of the estimated delays with the percentage of arrivals during green ( $s = 2,000$ ).

indifferent to the level of saturation. This result is somewhat counter to our expectations. The data contain significant scatter, particularly at high degrees of saturation ( $X > 0.75$ ), which contain data with both high and low values for the progression factor. We expect that the full regression models may show, at best, a weak association between the platoon factor and degree of saturation. It appears as though the platoons from the actuated systems that we studied vary greatly from cycle to cycle.

Figure 9 shows clearly that the progression factor decreases as the platoon ratio increases. The effect is strongly significant and is consistent with our hypothesis.

We tested more than 20 alternative model specifications and conclude that the following two models offer not only the best fit to our data but are also theoretically consistent. This fit was achieved after the exclusion of an outlier resulting from erroneous field measurement. The models are:

$$PF = 0.86 - 0.17 X - 0.24 R_p$$

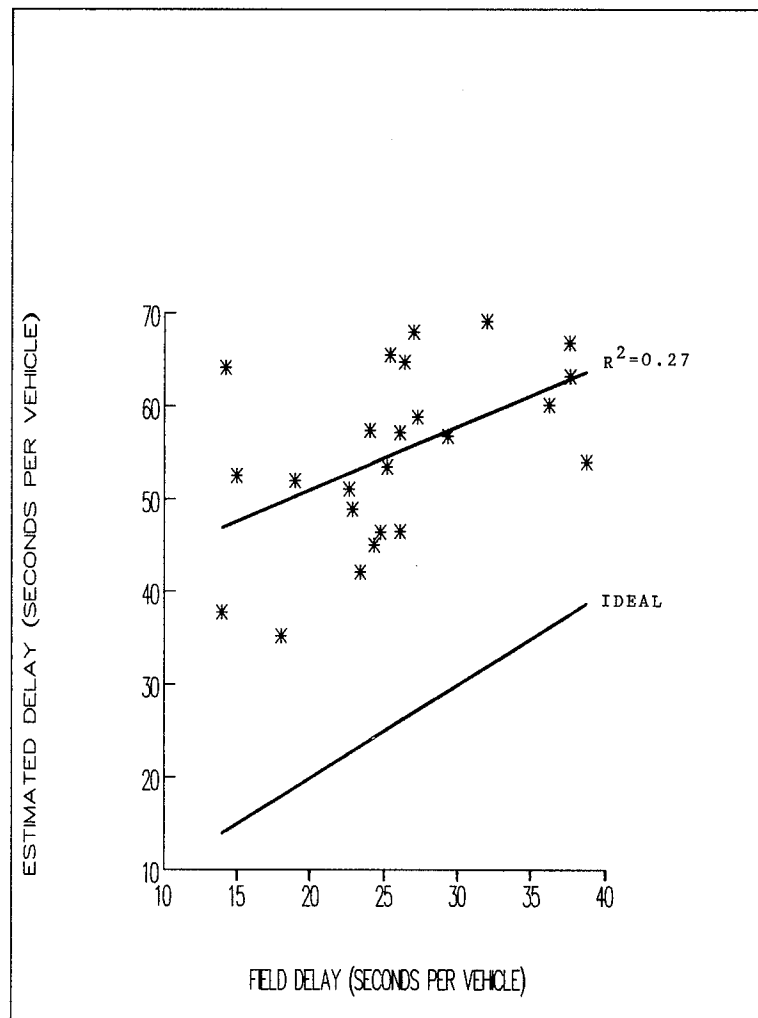
$$(R^2 = 0.41, F = 7.6) \quad (6)$$

$$PF = e^{-0.895 X - 0.474 R_p} R_p^{-0.650}$$

$$(R^2 = 0.40, F = 7.4) \quad (7)$$

These models were estimated from twenty-five independent approximately 15 minute intervals. Table 2 summarizes the results of the bivariate models and equations 6 and 7, including goodness of fit statistics and tests of significance for model parameters and the regression model as a whole.

The results obtained from both models indicate that the values of the progression factor in the HCM are substantially different from the values which are reflected in our data. Although our data set is quite small, the parameter estimates are statistically significant as are the models as a whole. A clearer perspective of the difference between our Model 2 and current HCM values is illustrated in Figure 10, which plots both on the same scale, along with the 95 percent confidence interval for the regression line. It is clear that the model estimated from our data is very different from the HCM values. For the Arrival Type 3, ( $0.86 < R_p < 1.15$ ) which is the most common observation in our data set, our estimates are roughly 60 percent less than the corresponding values in the HCM. This means that the delays estimated from our model are 60 percent less than the ones which are estimated using the  $PF$  values in the HCM. We are surprised at the magnitude of this difference but confident that the results accurately and significantly reflect field data.



**FIGURE 7** Relationship between the field delay and the HCM-estimated delay without the progression adjustment.

Arrival Type 3 is the most common type of arrival that we observed in our networks with actuated signals. Because the signal settings and offsets vary cycle by cycle, all types of arrivals are identified with each 15-minute analysis period. Individual cycles with excellent progression (Type 1) are often balanced by subsequent cycles with poor progression (Type 4 or 5). This is in direct contrast to coordinated pretimed systems, which have fixed cycle lengths, splits, and offsets throughout the analysis period. A greater uniformity in arrivals and platooning with pretimed control is expected. This causes us to question if one would ever consistently observe Type 1 or Type 5 (*i.e.*, extreme) arrivals with actuated systems. In fact, our study includes no data for Arrival Type 1 ( $0.0 < R_p < 0.50$ ) and only two observations for Type 5 ( $1.51 < R_p$ ).

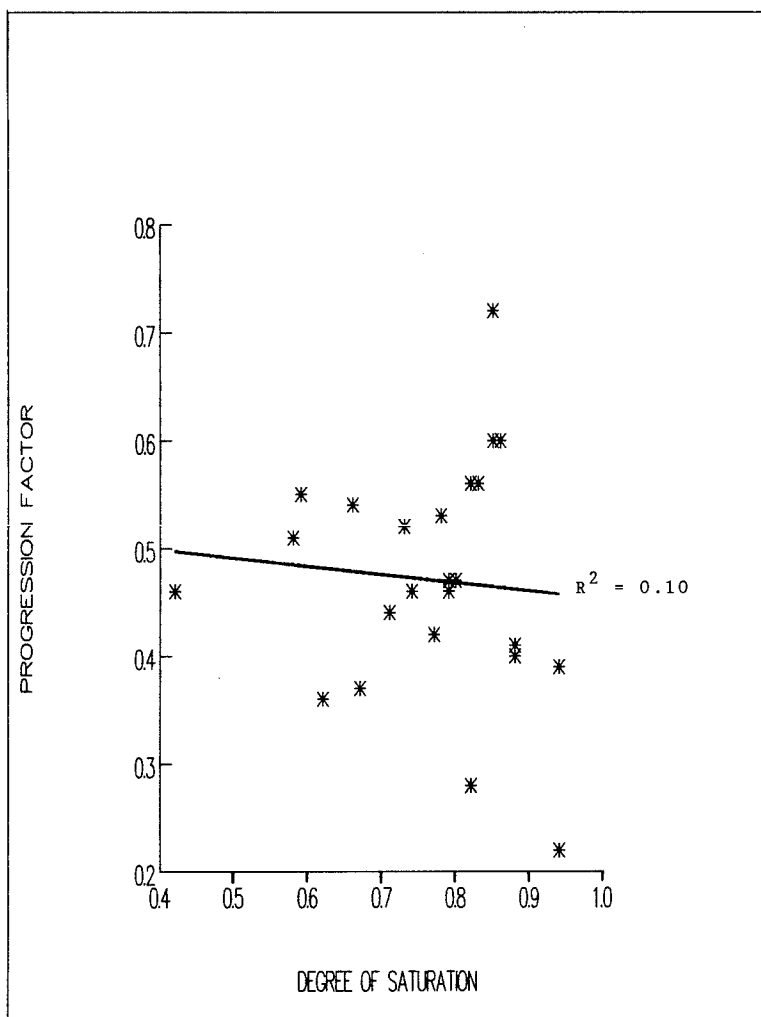
## CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

### Conclusions

Saturation flows observed in the field are significantly higher than those in the manual. A value of nearly 2,000 vphgl is

consistently observed at field sites where the HCM estimates values of 1,800 vphgl. The difference in saturation flows significantly affects delay estimates: the delays predicted using the value of 2,000 more closely match field delays than do the estimates using 1,800. Our conclusion is that an ideal saturation flow equal to 2,000 (rather than the HCM value of 1,800) is recommended for the analysis of high design intersections similar with those analyzed in our project.

The evaluation and calibration of the values for the progression factor consumed a significant amount of effort in all the stages of the project. Despite collecting data over a 4-week period at ten sites, data are restricted to random (Type 3) or nearly random (Type 2 and 4) platoon types. Only two observations of Type 1 or 5 platoons are contained in the data. The progression factors estimated from our data are statistically significant, however, and differ significantly from HCM values. Our platoon factors are lower than the HCM, indicating that greater reductions in delay are necessary than are currently provided in the progression factor adjustments. In addition to being lower than current HCM values, the study's progression factors are much less sensitive to the platoon ratio (a relative measure of the percent arrivals on green). It appears that this field data for actuated controllers is almost



**FIGURE 8** Relationship between the progression factor and the degree of saturation.

totally inconsistent with the current HCM. The authors believe that their values are more appropriate for high-design traffic actuated intersections and that they should be used in place of the current HCM values.

#### Recommendations for Future Research

The large differences in both saturation flows and the progression factor surprised even the current research team. This experience indicates the importance of exploring the validity of HCM procedures that are based on limited field data. During the course of the research, there were a number of instances in which additional studies seemed warranted. The following is a summary of those additional topics.

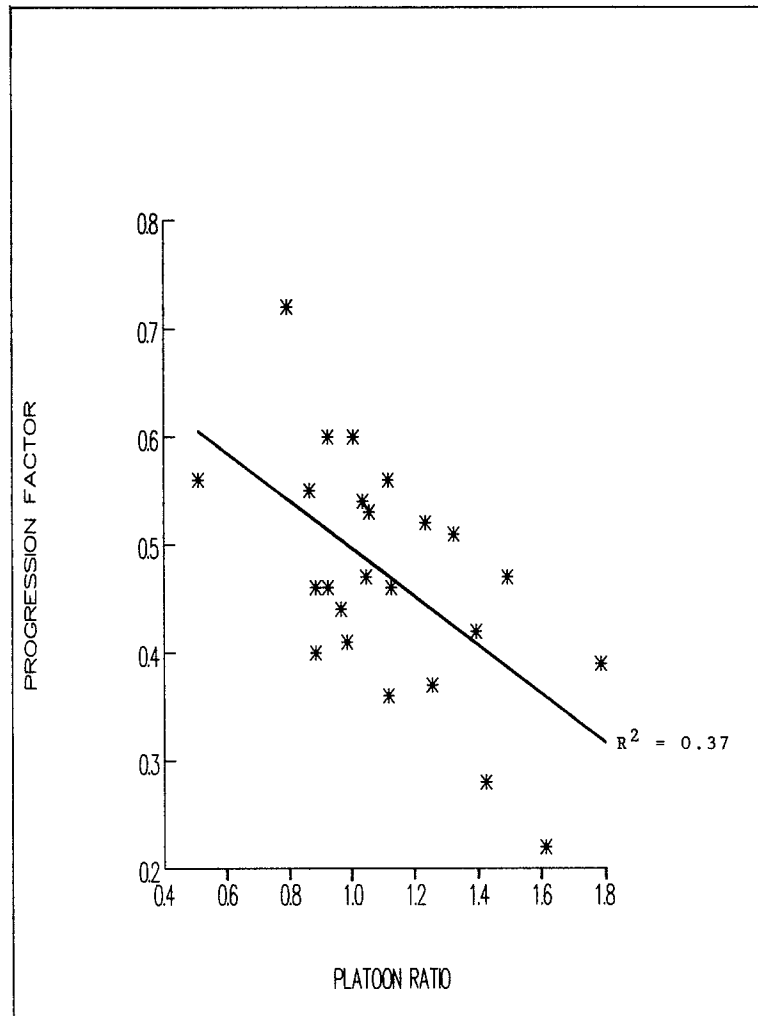
The measurement of saturation flows (and their effect on delay need to be extended) to a broader range of intersections and lane types. Additional studies could include measurement of saturation flows in smaller urbanized areas in similar geometric conditions. This would allow testing of regional differences in saturation flow rates. Given the importance of

saturation flow in delay estimation, this is a very useful project.

During the course of the current research, observations at left turn bays indicated that saturation flows for protected left turns may be higher than for through lanes. This is based on casual observation only; it would be instructive to conduct more detailed measurements.

Much more can be done to validate other adjustments to saturation flow. Nearly all of the saturation flow adjustments are candidates for further study, including: the effect of left turns, grades, lane width, parking, and local buses. Carefully designed site selection should allow for a more precise estimate of each of these effects.

Given the large differences that are observed between HCM delays and field data, it appears that much more needs to be done to better understand delays for systems of actuated controllers. The data set for this study contained very few 15-minute periods with Type 1 or Type 5 arrivals. Is this to be expected? Is this generalizable to other locations? Further study of actuated intersection delay is needed. Much more should be done to improve the accuracy of our delay esti-



**FIGURE 9 Relationship between the progression factor and the platoon ratio.**

**TABLE 2 SUMMARY OF PROGRESSION FACTOR REGRESSION MODELS**

Variable Name	Model 1	Model 2	Model 3 <sup>a</sup>	Model 4 <sup>a</sup>
	$\hat{\beta}$ (t)	$\hat{\beta}$ (t)	$\hat{\beta}$ (t)	$\hat{\beta}$ (t)
Degree of Saturation	-0.28 (0.5)	--	-0.17 (1.2)	-0.47 (1.8)
Platoon Ratio	--	-0.26 (3.7)	-0.24 (3.4)	-0.65 (3.2)
Constant	0.67 (3.7)	0.75 (9.2)	0.86 (6.9)	-0.90 (9.8)
R <sup>2</sup>	0.10	0.37	0.41	0.40
F Value for Significance of Model	1.1	13.6	7.6	7.4

<sup>a</sup>Model 3 is a linear additive specification, model 4 is multiplicative.

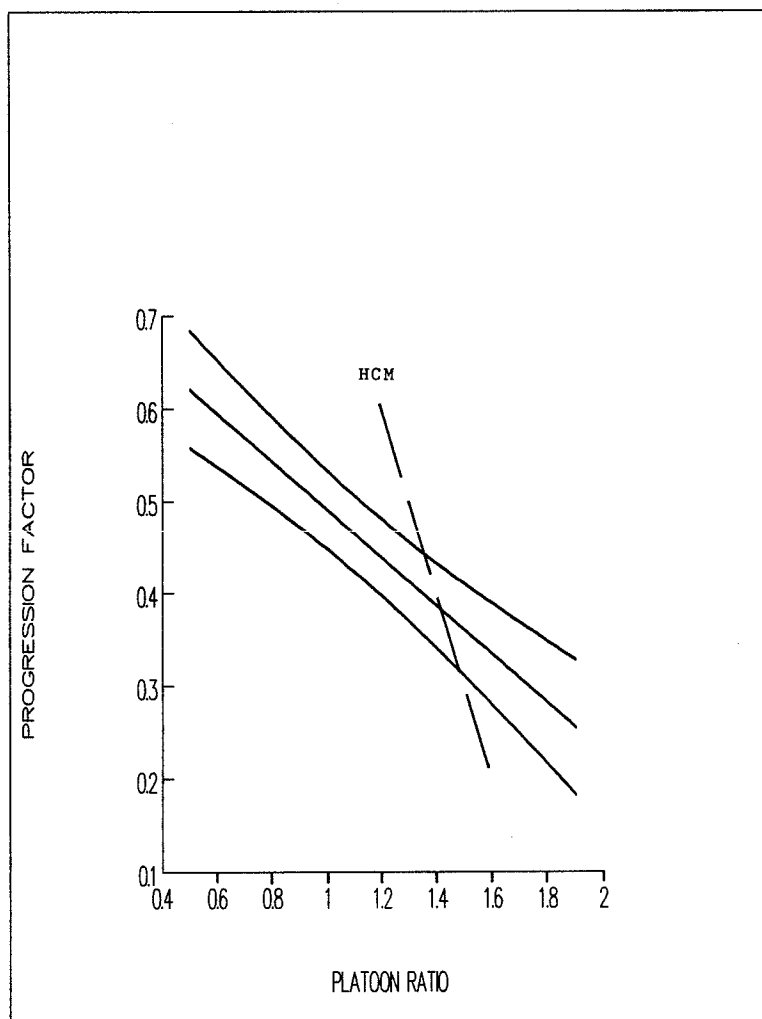


FIGURE 10 Comparison of values for the progression factor from this analysis and the 1985 HCM (confidence of estimate = 95%).

mates, as these have a direct relationship to the quality of resource allocation decisions.

#### ACKNOWLEDGMENTS

The authors gratefully acknowledge the support and cooperation of the Illinois Department of Transportation (IDOT) and the Federal Highway Administration. The authors wish to thank especially John Sanford of IDOT, who contributed many thought-provoking questions during the research in his role as technical monitor. This paper reflects the views and opinions of the authors, not the sponsoring agencies.

#### REFERENCES

1. Signalized Intersections. In *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985, 516 pp.
2. At Grade Intersections. In *Special Report 87: Highway Capacity Manual*. HRB, National Research Council, Washington, D.C., 1965, 411 pp.
3. P. Jovanis, P. Prevedouros, and N. Rouphail. *Design and Operation of Signalized Intersections in Illinois*. Final Report No. IHR 013. Illinois Department of Transportation; The Transportation Center, Northwestern University, April 1987, pp. 1-99.
4. K. Agent and J. Crabtree. Analysis of Saturation Flow at Signalized Intersections. Kentucky Transportation Program, University of Kentucky, Lexington, May 1982.
5. J. Zegeer. Field Validation of Intersection Capacity Factors. In *Transportation Research Record 1091*. TRB, National Research Council, Washington, D.C., 1986, pp. 67-77.
6. E. Chang and D. Fambro. Effects of Quality of Traffic Signal Progression on Delay. Final Report, NCHRP Project 3-28C. TRB, National Research Council, Washington, D.C., forthcoming.
7. D. Berry. Sensitivity of Overflow Delay to Arrival Patterns. Paper presented at the 66th TRB Annual Meeting. January 1986.

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

# Modeling of Shared Lane Use in TRANSYT-7F

CHARLES E. WALLACE AND FRANK J. WHITE

The TRANSYT family of programs continues to be one of, if not the most, widely used computer programs in the world for traffic signal timing and traffic flow analysis. In the past this program was excellent for modeling protected or unopposed traffic movements from separate lanes; however, it did not have the capability to model several different movements, for example, unprotected left turns and through movements from a shared lane. A project to incorporate a model to explicitly deal with this condition in TRANSYT-7F is described. The model is based on the 1985 Highway Capacity Manual. Its implementation in TRANSYT-7F and the user interface are reviewed.

The Traffic Network Study Tool (1), Version 7, Federal (TRANSYT-7F) (2) is one of the most useful tools available to the traffic engineer for traffic operations analysis and traffic signal timing optimization. The traffic model is a deterministic, macroscopic time scan simulation model, which is quite realistic in modeling homogeneous flows unencumbered by other traffic.

The applicability of the model was, in the past, somewhat limited when one traffic movement had to yield to another, for example, when permitted left turns were opposed by traffic traveling in the opposite direction. If the permitted left turn movement had its own lane or bay, this could be approximated roughly by estimating a reduction in the saturation flow rate to represent left turners yielding to opposing traffic and by delaying the start of effective green to reflect the delay due to the departure of the opposing through queue. However, if the unprotected left turn movement was from a common or shared lane with the through traffic, no reasonable way to model this condition existed in TRANSYT.

To overcome the limitations, the Federal Highway Administration (FHWA) awarded a contract to the University of Florida Transportation Research Center (TRC) to develop and implement Enhancements to the TRANSYT-7F Program. The specific objectives of the first phase of the project are summarized as follows:

1. Develop a permitted movement algorithm that will enable TRANSYT-7F to model permissive and protected plus permitted left-turn phasing, including "sneakers" that turn at the end of the permitted phase.
2. Develop algorithms that will enable TRANSYT-7F to explicitly model stop sign control and shared left and through lanes.

The first objective was reported earlier (3). The stop control model was handled in TRANSYT by simply estimating a minimum delay to sign controlled traffic, equal to the time requirements to decelerate from the coded speed to a stop, then accelerate to the speed when the permitted model indicated available capacity.

This paper deals with the selection and implementation of the shared lane model. It describes the development of the model and its implementation in TRANSYT-7F.

## EXISTING TRANSYT TREATMENT OF SHARED LANES

TRANSYT-7F, like its predecessors, is a deterministic macroscopic time scan simulation model intended specifically for unencumbered traffic flow. Traffic movements are modeled according to a very simple rule of flow. Periodic flow patterns are modeled in an upstream-to-downstream order. A flow pattern representing the periodic departure as a function of time during a typical cycle is

0, if the link's signal is effectively red;

$GO(t)$ , if the signal is effectively green and

$OUT(t)$  = a queue exists; and  $IN(t)$  if the signal is effectively green and no queue exists. (1)

where

$OUT(t)$  = the "output" flow rate at time interval  $t$ , in vph;

$GO(t)$  = the maximum flow, or "go" rate (vph), which is the saturation flow at time  $t$ ; and

$IN(t)$  = the arrival, or "input," rate (vph) at the reference point of the link at time  $t$ , which is a product of TRANSYT's platoon dispersion model, or a uniform rate for "external" links.

In TRANSYT's simulation model the IN pattern is known. It is predicted by the platoon dispersion model for "internal" links or is a uniform distribution for "external" links. Thus, the key to incorporating a shared lane model in TRANSYT is to calculate the GO pattern.

The earlier versions of the model had no provision for permitted movements, which must filter through opposing flows. As mentioned previously, recent work (3) has been undertaken to provide for permitted movements from exclusive lanes. Models were developed empirically from the data collected around the Washington, D.C., area and have been installed.

The shared lane analysis incorporates permissive turners

and is therefore an extension of this work. Analytical procedures and simulation were used in lieu of field calibration due to the great resource demand of the latter.

Prior to this work TRANSYT-7F was unable to treat shared lane flow explicitly. The user's manual (2) suggested using the shared stopline feature in which vehicles depart from the set of shared stopline links in the order in which they arrive, regardless of the link on which they arrive. All vehicles depart at the lesser of the rate at which they arrive or the saturation flow rate coded for the primary link (saturation flows were not coded for minor links).

Coded saturation flows could either be measured or estimated. Measurements must be taken under "typical" conditions. Because of the many factors which influence the saturation flow in a shared lane, the results would not be as reliable as saturation flow for exclusive through lanes, which are generally only governed by geometrics and driver behavior. Aside from direct measurements, "engineering judgment" could be used, but only when measurements or calculations were infeasible.

The saturation flow rate may also be calculated, for example using the Highway Capacity Manual (HCM) (4) procedure. This procedure is somewhat involved, but is possibly the most realistic approach available to obtain a reliable result. If the HCM procedure is used, however, signal timing parameters must be known. In a simulation they are known but must be estimated if an optimization is performed. Unless the initial guess proves correct, this requires a recursive approach that can be quite consuming in resources.

TRANSYT's simulation routine, progressing link by link in a downstream order, is not conducive to permitted flows, especially shared lane flow with permitted elements. Shared lane flow depends on a plethora of variables, the primary elements being opposing flow and left turn volume. Due to this dependence on opposing flow, shared lane saturation flow will vary on a step-by-step basis as opposing flow varies. Unopposed movements, such as exclusive through lanes, are only constrained by intersection geometry and driver behavior. Thus, the saturation flow for a given location is generally constant over time.

If an intersection has shared lanes opposing each other, calculation of saturation flows, particularly on a step-by-step basis, would require simultaneous processing of both approaches as current opposing flows are required to calculate up-to-date saturation flows. Such a computing procedure would require rewriting TRANSYT completely, which was beyond the scope of the FHWA project.

### Modeling Approach

Many analysis methods of traffic movements have attempted to equate vehicles to a standard, typically equivalent through passenger vehicles (ETVs), because they make up the dominant proportion of all movements. Additionally, through vehicles generally make maximum usage of the roadway when compared to oversized vehicles and turning vehicles, particularly those that must filter through an opposing flow. Saturation flows can be expressed in terms of ETVs and remain constant, independent of the traffic mix. Shared lane analysis uses the ETV concept.

Traffic in a shared lane takes two basic forms:

1. Protected left turns (exclusive phase): Vehicles in the shared lane may move freely without impedance from vehicles traveling in the opposite direction. Saturation flows can be readily measured for protected movements, and suggested rates can be found in the HCM (4). Protected shared lane saturation flow can be determined using these suggested values.

2. Permissive left turners: ETVs for the left turners are dependent on the opposing volume. Past research provides various forms for the relationship between permissive turners and opposing flow, ranging from a simple linear to a more complex exponential form. The saturation flow of shared lanes is also a function of the composition of traffic in the lane and the vehicles opposing them.

Many previous studies were investigated, including work by Akcelik in Australia (5, 6); Peterson et al. (7) in Sweden; City of Edmonton, Canada (8); Lin et al. (9); Lee et al. (10) at the University of Texas; and Fambro, Messer, and Anderson (11, 12). Most of these methods use at least a variation of the ETV concept and gap acceptance models developed by the authors or others, such as Drew (13).

Because a permitted movement model was being added to TRANSYT-7F at the time of this research, and the resources to add the shared lane capability were limited, it was decided to use an approach similar to the HCM.

### Model Considerations

An analytical model was developed by first establishing upper and lower bounds for shared lane saturation flow. A critical lane procedure was used to pinpoint a unique flow rate, within the previously developed bounds.

#### Bounds Solution

As a preliminary investigation of shared lane behavior, upper and lower bounds for the shared lane saturation flow,  $S_s$ , were addressed.

The saturation flow for a shared lane is primarily dependent on the mix of traffic in the lane, through and left vehicles, and the opposing volume, which interacts with the left turners. The obvious upper bound use of the lane occurs when it is used exclusively by through vehicles, resulting in a saturation flow of  $S_{ts}$ , the base saturation flow for through vehicles. Conversely, the lower bound occurs when the lane operates as an exclusive left turn lane with a saturation flow of  $S_l$ , the permissive left turn saturation flow rate. In both cases, however, the lane is not operating as a shared lane.

To investigate shared lanes they will be viewed from the perspective of left turners, as Lin viewed them (9). Given a through volume of  $V_{ts}$ , occupying the shared lane, what will be the left turn saturation flow,  $S_{ls}$ ? The shared lane saturation flow can be determined by simply adding  $S_{ts}$ , the left turn saturation flow from a shared lane and the through volume added to saturate the shared lane, which is equal to  $S_{ts}$ , the through saturation flow from a shared lane.

An upper bound solution occurs when the through vehicles sharing the lane do not utilize time when gaps occur in the opposing traffic, which otherwise would be available to the permissive left turners. In fact, it is assumed that the left turns

are distributed such that they arrive ideally to use all available gaps. It is assumed that all the opposing vehicles arrive together at their saturation flow rate, and move simultaneously with the through vehicles in the shared lane, which are also departing at their maximum flow rate. This allows the maximum number of left turners to use the shared lane.  $S_s$  is therefore the minimum of

$$S_{ts} = S_{to} - V_o, \text{ or } S_{to} - V_{ts} \quad (2)$$

where

$$\begin{aligned} S_{ts} &= \text{left turn saturation flow from a shared lane, vphg;} \\ S_{to} &= \text{base through saturation flow from a shared lane, vphg;} \\ V_o &= \text{opposing volume, vph;} \text{ and} \\ V_{ts} &= \text{through volume in the shared lane, vphg.} \end{aligned}$$

The shared lane saturation flow rate is the lesser of

$$S_s = S_{to} - V_o + V_{ts}, \text{ or } S_{to} \quad (3)$$

where  $S_s$  is shared lane saturation flow, vphg; and  $S_{to}$ ,  $V_o$ , and  $V_{ts}$  are as defined before.

If every through vehicle in the shared lane used an otherwise available opportunity for a left vehicle to turn, a lower bound solution would occur. The respective saturation flows are

$$S_{ts} = S_l - V_{ts} \quad (4)$$

and

$$S_s = S_l \quad (5)$$

where  $S_l$  is permissive left turn saturation flow, vphg; and  $S_{ts}$ ,  $S_s$ , and  $V_{ts}$  are as defined before.

### Critical Lane Analysis

Critical lane analysis was used for signalized intersections in the 1985 HCM. Earlier work by Messer (12) formed the foundation for the resultant procedures. Unfortunately, no documentation describing the development of the HCM procedures exists, other than the brief description in the manual itself. With this and the complexity of the shared lane procedures in mind, it was decided not to insert the copious HCM equations into TRANSYT directly, but to use the critical lane procedure to produce a more simplified approach.

Critical lane analysis assigns vehicles equally among available lanes on the basis of their ETV volume. This is predicated on the assumption that drivers strive to minimize delay in their travel.

When permissive turners are converted to ETVs, their numbers increase above their actual on-road volume. In effect, it is assumed that all drivers have "perfect knowledge" about existing traffic conditions and future events. That is, through vehicles arriving at an intersection on a multilane approach, which includes a shared lane, with vehicles already queued in each of the  $N$  available lanes, have  $N$  possible lane choices. The choice process can be reduced to two basic steps. First there is a choice between the  $(N - 1)$  identical through lanes. This is a rather trivial process; obviously the lane with the shortest queue will be chosen. Next the result of the first step is compared to the shared lane.

Choosing between the shared lane and the preferred through lane is not quite as straightforward. At this stage the concept

of ETV becomes important. Assuming that only passenger cars use the roadway, permissive turners must be converted to ETV on the basis of their opposing traffic. The actual shared lane queue on the roadway serves as an indicator to approaching drivers of the desirability of the lane. If it is longer than the favored through lane, for example, one would expect the drivers to reject the shared lane.

Problems arise when the shared lane queue is equal to or less than the best through-only alternative. The driver must make a subjective judgment on lane use. Intuitively one would expect the through lane to be favored when the actual queues are the same length, but the risk of joining the marginally shorter shared lane queue and being delayed by a left turner is high. As the actual queue length of the through lane increases, a point of equilibrium is reached when the driver is equally likely to select either lane. In theory this occurs when the ETV queue lengths are identical. Beyond this point the driver would be expected to join the shorter queue in the shared lane.

Of course drivers do not explicitly think in terms of ETV when making lane choices; this is a procedure developed to simulate lane choice. This concept, however, has intuitive appeal, aside from the fact that it is assumed arriving drivers know about the future opposing traffic while at the intersection.

The above interpretation of driver behavior forms the basis of the shared lane model developed below.

### Model Derivation

The calculation of shared lane flow involves two steps:

#### Step 1. Calculation of $V_{ts}$

It is assumed that vehicles are distributed evenly among the available lanes on the basis of their ETV. Therefore,

$$V_s (ETV) = (V_t + EL * V_l) / N \quad (6)$$

where

$$\begin{aligned} V_s (ETV) &= \text{volume in the shared lane, ETV/hr;} \\ v_t &= \text{through volume in all approach lanes;} \\ EL &= \text{through vehicle equivalent for opposed left turns;} \\ V_l &= \text{left turn volume, vph;} \text{ and} \\ N &= \text{number of lanes on the shared lane approach being analyzed.} \end{aligned}$$

A check should be performed to see if the shared lane is acting as a de facto left turn lane. The check is as follows:

$$\text{If } V_s (ETV) < EL * V_l, \quad (7)$$

then treat as an exclusive left turn lane.

In this case the lane saturation flow is equal to the permissive left turn saturation flow,  $S_l$ . Otherwise the following analysis procedure is used to determine shared lane saturation flow rates:

$$V_{ts} = V_s (ETV) - EL * V_l \quad (8)$$

where  $V_{ts}$  is through volume in the shared lane, vph; and the rest were defined previously.



Substituting  $V_s$  (ETV) from Equation 6 into Equation 8 gives

$$\begin{aligned} V_{ts} &= [(V_t + EL * V_t)/N] - (EL * V_t) \\ &= [V_t + EL * V_t * (1 - N)]/N \end{aligned} \quad (9)$$

### Step 2. Use $V_{ts}$ to Determine Saturation Flows

The available capacity of the shared lane, from  $S_{to}$ , is distributed among the several movements using the lane according to their respective volumes, in ETVs. The left turn saturation flow from a shared lane is as follows:

$$S_{ls} (ETV) = S_{to} * (V_t * EL)/V_s (ETV) \quad (10)$$

where  $S_{ls}$  is left turn saturation flow from a shared lane, ETVphg;  $S_{to}$  is base saturation flow for through vehicles, vphg; and the rest as previously defined.

Equation 10 gives the left turn saturation rate in terms of effective through vehicles. If Equation 10 is divided by the equivalent left turn factor,  $EL$ , as follows, the saturation flow can be expressed in the more familiar units of vphg:

$$\begin{aligned} S_{ls} &= S_{ls} (ETV)/EL \\ &= (S_{to} * V_t)/V_s (ETV) \end{aligned} \quad (11)$$

where  $S_{ls}$  is left turn saturation flow from a shared lane, vphg; and the rest were defined previously.

The through saturation flow rate from a shared lane is as follows:

$$S_{ts} = S_{to} * V_s/V_s (ETV) \quad (12)$$

where  $S_{ts}$  is through saturation flow from a shared lane, vphg; and the rest were defined previously.

The sum of the shared lane saturation flows, in units of ETVphg, is equal to the base through saturation flow rate,  $S_{to}$ , or

$$(S_{ls} * EL) + S_{ts} = S_{to} \quad (13)$$

Using a volume division of the available capacity is only an approximation, increasing in accuracy as the shared lane approaches its limit, saturation.

### Through Volume Equivalency and Effective Opposing Volume

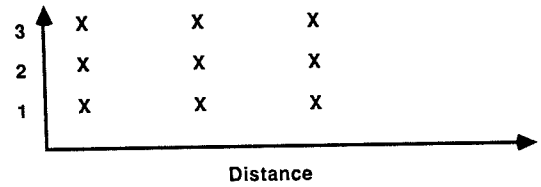
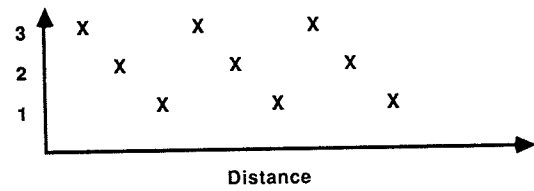
Through volume equivalency used for opposed turns was determined using the HCM procedure:

$$EL = S_{to}/(S_{to}) \quad (14)$$

where  $V_o < S_{to}$  and  $S_{to}$  is base protected left turn saturation flow rate, vphg.

The HCM uses  $S_{to} = 1,400$  vphg. In the calculation of this factor 1,800 vphg is used for  $S_{to}$ . A departure from the manual was made, however, in calculating the opposing volume,  $V_o$ .

The HCM does not differentiate between opposing flow as a function of the number of lanes,  $n$ , available. It is assumed that the opposing flow faced by permissive turners is the same, regardless of  $n$ . Figure 1 illustrates upper and lower bound solutions of the effect of the number of lanes available for



- Notes: 1. Vehicles are represented by the symbol X. In both cases these vehicles represent the total opposing flow,  $V_o$ , vph.  
2. Only three of  $n$  possible lanes are shown for clarity.

FIGURE 1 Effect opposing volume bounds; X = total opposing flow ( $V_o$ ) in vph; only three of  $n$  possible lanes are shown.

opposing traffic. At worst the opposing volume,  $V_o$ , could be distributed entirely uniformly, in such a way that the opportunity for left turners to filter through is the same as if only one lane were available. Conversely,  $V_o$  may be distributed in echelon fashion, such that the effective opposing volume is reduced to  $V_o/n$ . In the absence of specific theory or data, it was decided to use a simple average of these two extremes in calculating the opposing volume corrected for multiple opposing lanes,  $V'_o$ . That is:

$$\begin{aligned} V'_o &= [V_o + (V_o/n)]/2 \\ &= (V_o * (n + 1))/(2 * n) \end{aligned} \quad (15)$$

where  $V'_o$  is opposing volume corrected for multiple opposing lanes, vph; and  $n$  is number of opposing lanes.

No adjustment is necessary if the opposing approach itself includes a shared lane. Clearly the number of opposing lanes is less than  $n$ . This question is left to future research.

Second, the HCM only uses  $EL$  when permissive turners are filtering through opposing flow,  $g_u$ , which by this time is assumed to have returned to the random arrival rate, because opposing vehicles that queued during the preceding red period have cleared during  $g_q$ . Figure 2 illustrates the calculation of the unsaturated portion of the permissive green phase. This behavior is reasonable for an isolated intersection where vehicles will in fact arrive in a random manner. However, when signalized intersections are closely spaced and coordinated, random arrivals cannot be assumed, because platooning will tend to occur.

The platoon may arrive at any time during the cycle, assuming the traffic is, in fact, grouped; however, the step-by-step

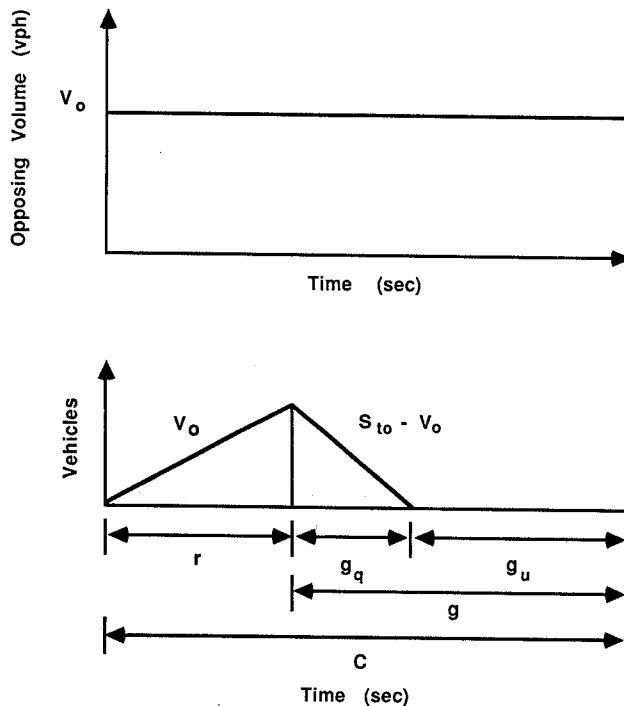


FIGURE 2 Calculation of unsaturated portion of permissive green phase.

calculations would require a recursive model which, as stated before, is beyond the scope of this work. The model thus assumes a "uniform" saturation rate, but the opposing flow rate is adjusted as described below.

The opposing flow used is the coded hourly volume  $V_o$ , multiplied by the  $C/g$  ratio, where  $C$  is the cycle length and  $g$  is the effective green time during which traffic in the shared lane may move. In effect, the opposing flow faced by the permissive turners has been evenly distributed over the available turning time, recognizing that this is an approximation.

#### Finalizing the Model

Calculations of the shared lane saturation flows were also modified for both multiple opposing lanes and signal timing parameters. The latter is necessary because the adjusted opposing flow ( $V_o$  above) moves only during its effective green; thus, the effective flow rate is found by multiplying  $V_o$  by the  $C/g$  ratio, or

$$V_{\text{oeff}} = V_o * (C/g) \quad (16)$$

where  $V_{\text{oeff}}$  is effective opposing volume, corrected for both multiple opposing lanes and signal timing.

The modified form of  $EL$  is thus

$$EL' = S_{to} / (S_{to} - V_{\text{oeff}}) \quad (17)$$

$EL'$  is substituted for  $EL$  in Equations 6 through 11 to give the resultant saturation flows from a shared lane incorporated into TRANSYT.

The final models are expressed as follows:

$$S'_{is} = S_{io} * V_i / V_s (ETV)' \quad (18)$$

and

$$S'_{is} = S_{io} * V'_{is} / V_s (ETV)' \quad (19)$$

where

$S'_{is}$  = modified left turn saturation flow from a shared lane, vphg;

$S_{is}$  = modified through saturation flow from a shared lane, vphg;

$V_s(ETV)'$  = modified volume in the shared lane, ETV/hr; and

$V'_{is}$  = modified through volume in the shared lane, vph; or

$$V_s(ETV)' = (V_i + EL' * V_i) / N \quad (20)$$

and

$$V'_{is} = V_s(ETV) - EL' * V_i \quad (21)$$

$V_i$  and  $S_{io}$  are as defined previously. Here  $S_{io}$  is the user coded saturation flow rate for the shared lane group, adjusted as appropriate for the number of lanes.

## IMPLEMENTATION INTO TRANSYT

A major aim of the shared lane enhancement was to ensure that minimal additional coding would be required by the user, and if possible all changes could be incorporated into the existing input data file. Additionally, the shared lane capability was only one component of a variety of changes being undertaken, so it had to be compatible with the other modifications, namely, permissible turners, stop-control, and sneakers. Through the implementation of a simplified analytical procedure, these aims were met, in addition to modeling the shared lane behavior somewhat realistically.

### User Interface

In regard to user input data, no new data are required to use the shared lane facility per se. All user interface either existed previously, or resulted from the addition of the permitted movement model. In short, the following three items summarize how the user informs TRANSYT-7F of a shared lane condition:

1. The shared stopline facility (card type 7) is used for the shared lane group, just as before. The through, unopposed movement must be the primary link and the permitted, opposed link (only one allowed) is one of perhaps several other secondary links.
2. The permitted opposed link is identified on the phase data card (card type 2X) with a negative number, as required for the permitted movement model. A secondary maximum flow rate to override the permitted movement model is optional (card type 29).
3. Sneakers and the opposing link numbers, with an optional percentage, are coded on the link data continuation card (card type 29).

## Model Implementation

The foregoing model was incorporated directly into TRANSYT's traffic simulation routine, in close coordination with the permitted movement model referred to previously (3).

To clarify how the permitted model is used, a typical simulation run of TRANSYT-7F will now have three iterations, or passes, through the simulation model (for "normal," non-shared lanes):

1. In Pass 1 all links are simulated, but permitted links are modeled with a uniform maximum flow rate based on the HCM.
2. In Pass 2 only permitted links are simulated, using the permitted movement model reported by Wallace (3). This provides a more realistic estimate of traffic performance as a function of opposing traffic.
3. Finally, in Pass 3 all nonpermitted links are resimulated to correct their flow patterns for any changes which occurred as a result of Pass 2.

In the case of a shared lane the above is slightly different. In Passes 2 and 3, the through link saturation flow rate is calculated as in Equation 19, and the left turn link's saturation flow rate is calculated according to Equation 18. During Pass 2, the latter's maximum flow rate is the lesser of the value calculated by Equation 18 or the rate calculated step by step by the permitted movement model.

During any simulation, a check is first made to see if the shared lane is acting as an exclusive left turn lane. If this is the case, the lane is treated as a permissive left and the approach through maximum flow rate,  $TMFR$ , reverts to

$$TMFR = LSATF_i * (N - 1)/N \quad (20)$$

where  $LSATF_i$  is link saturation flow (total for all lanes), vphg; and  $N$  is number of approach lanes as defined previously.

If, however, the lane is shared, the respective saturation flows are as follows:

$$S'_s = (LSATF_i/N) * (V'_s/V_s(ETV)') \quad (21)$$

and

$$S'_l = (LSATF_i/N) * (V'_l/V_s(ETV)') \quad (22)$$

where

$S'_s$  = modified through saturation flow from a shaded lane, vphg;

$S'_l$  = modified left turn saturation flow from a shared lane, vphg;

$V'_s$  = modified through volume in the shared lane, vphg;

$V_s(ETV)'$  = modified volume in the shared lane, ETV/hr;

$V'_l$  = left turn volume, vph; and  $LSATF_i$  and  $N$  are as previously defined.

Both saturation flows are assumed to be constant over their effective greens. In reality both flows vary over time as opposing flow varies; however, to model this would require a recursive approach, which was beyond the scope of this enhancement. The adopted procedure emulates traffic on a cycle-by-cycle rather than a step-by-step basis.

## CONCLUSIONS AND RECOMMENDATIONS

The procedure divided the available shared lane capacity among its tenants on the basis of ETVs. This is an approximation approaching the actual saturation flows as the lane nears capacity. In fact, capacity represents the limit when the analytical and actual saturation flows meet.

Shared lanes can now be modeled explicitly by using the existing or only slightly modified inputs. Inclusion of an explicit treatment into TRANSYT represents a significant enhancement, filling a previous void in the package. This procedure may be considered for adoption in other traffic analysis models as well.

Field validation of the adopted shared lane saturation flow models should, however, be undertaken. Only through such testing can it be assured that they are reproducing "real world" results.

An iterative procedure for shared lane analysis should be investigated if the current model proves inadequate, but such a procedure would require a major revision of TRANSYT, possibly a rewrite of the entire program. Such an undertaking would be resource consuming; therefore, the benefits would need to be weighed against this cost.

The enhanced version of TRANSYT-7F is referred to as Release 5, made available in the fall, 1987. The mainframe version is available from FHWA, HTO-23, 400 Seventh Street, S.W., Washington, D.C. 20590. The microcomputer version is available from the McTrans Center, 512 Weil Hall, Gainesville, Florida 32611.

## ACKNOWLEDGMENT

The authors gratefully acknowledge FHWA's sponsorship of this project.

## REFERENCES

1. D. I. Robertson. *TRANSYT: A Traffic Network Study Tool*. Road Research Laboratory LR 253, Ministry of Transport.
2. C. E. Wallace, K. G. Courage, D. P. Reaves, G. W. Schoene, G. W. Euler, and A. Wilbur. *TRANSYT-7F User's Manual*. rev. ed. FHWA, U.S. Department of Transportation, 1987.
3. C. E. Wallace, F. J. White, and A. Wilbur. A Permitted Movement Model for TRANSYT-7F. Presented at 66th Annual Meeting of the Transportation Research Board, Washington, D.C., 1987.
4. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
5. R. Akcelik. *Timing Signals: Capacity and Timing Analysis*. Research Report ARR No. 123, Australian Road Research Board, 1981.
6. R. Akcelik. *SIDRA Version 2.2 Input and Output*. Technical Manual ATM No. 19, Australian Road Research Board, 1986.
7. B. E. Peterson, A. Hanson, and K. L. Bang. Swedish Capacity Manual. In *Transportation Research Record 667*, TRB, National Research Council, Washington, D.C., 1978, pp. 1-27.
8. *Saturation Flow Manual*. City of Edmonton Traffic Operations Section; University of Alberta Transportation Engineering Group, April 1980.
9. H. J. Lin, R. B. Machemehl, C. E. Lee, and R. Herman. *Guidelines for Use of Left-Turn Lanes and Signal Phases*. Center for Transportation Research Report CTR 3-18-80-258-1. University of Texas at Austin, Jan. 1984.
10. C. E. Lee, G. E. Grayson, C. R. Copeland, J. W. Miller, T. W. Rioux, and V. S. Savur. *The TEXAS Model for Intersection*

- Traffic—User's Guide*. Research Report 184-3. Center for Highway Research, The University of Texas at Austin, July 1977.
11. D. B. Fambro, C. J. Messer, and D. A. Andersen. Estimation of Unprotected Left-Turn Capacity at Signalized Intersections. In *Transportation Research Record 644*, TRB, National Research Council, Washington, D.C., 1977, pp. 113-119.
  12. C. J. Messer and D. B. Fambro. Critical Lane Analysis for Inter-

- section Design. In *Transportation Research Record 644*, TRB, National Research Council, Washington, D.C., 1977, pp. 26-35.
13. D. R. Drew. *Traffic Flow Theory and Control*. McGraw-Hill, New York, 1968.

---

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

# Critical Movement Analysis for Shared Left Turn Lanes

HERBERT S. LEVINSON

This paper analyzes critical movements for left turn lanes that are shared by through traffic. It shows that the critical conflict volume along a given approach is the sum of the left turns, the opposing traffic, and the blocked proportion of the through traffic in the shared lane. A set of impedance or blockage factors is derived as a function of the left turns per cycle. When two left turns per cycle occur, 60 percent of the through vehicles in the lane are blocked. When five left turns per cycle exist, 80 percent of the through vehicles in the shared lane are blocked. Thus, for all practical purposes, five or more left turns per cycle will pre-empt the shared lane. From this it follows that short signal cycles are desirable where shared left turn lanes predominate. The paper addresses shared lanes in both single and multilane approaches. It contains guidelines for computation of critical movement volumes in practice.

The left turn problem at signalized intersections has been a persistent and pervasive one for more than three decades (1, 2). A consensus exists that adjustments are needed to compensate for the longer headways (time spacings) by left turns and to reflect the additional delays or time requirements resulting from conflicts with through traffic. In addition, the impeditive effect (blockage) of through vehicles in shared left turn lanes is increasingly recognized.

This paper shows how left turns, through vehicles, and opposing traffic interact. It analyzes critical movements for a single "shared lane" approach and extends the result to the multiple lane approach case. It derives adjustment factors for use in critical lane analysis and contains illustrative examples of how the proposed procedures can be applied. In many respects it represents an extension of the planning analyses set forth in NCHRP Report 212 and the 1985 *Highway Capacity Manual* (3, 4).

## GENERAL CONCEPT

The critical lane volumes across a conflict point represent the sum of the conflicting movements along the artery and cross street. The critical lane (or conflict) volumes along the artery represent the sum of the left turn movement, the opposing through movement, and the through movements that share the lane with left turns and are blocked by them. The proportion of through movement that follows left turns is susceptible to delay by them and depends on the number of left turns in the shared lane.

This basic model assumes that the opposing through traffic

moves first and the left turns and blocked through traffic then move.

## Basic Relationships

These interrelationships best can be understood by considering only the artery volumes along a two-lane road with left turns in one direction only. Accordingly, Figure 1 shows conflict volumes for such a two-lane shared artery. The conflict volumes for a two-lane road with separate left turn lanes are shown for comparative purposes. All volumes are shown in passenger car units. The through volumes are assumed to include right turns, even though under special cases (i.e., wide cross street or right turn island) they could be deducted from the through traffic.

The critical conflict volumes along a shared two-lane, with left turns in one direction, represent the greater of the two volumes obtained from the following formulas:

$$\text{Critical lane volume} = L_1 + V_o + K_i \quad (1a)$$

$$\text{Critical lane volume} = L_1 + t_1 \quad (1b)$$

where

$t_1$  = through volume in lane shared by left turns;

$V_o$  = opposing volume; and

$K$  = impedance factor that reflects the proportion of through vehicles that follow, and are blocked by the left turns.

These definitions assume that the right-turning traffic is included in the through traffic and in the opposing traffic flow. They apply throughout the paper.

The limiting case is a three-phase operation in which the critical lane volume equals  $L_1 + t_1 + V_o$ . Here the impedance factor,  $K$ , equals one. Conversely, if a left turn lane is provided,  $K = 0$ .

Thus, the left turn impedance or blockage factor,  $K$ , is important in computing critical lane volumes. It varies from 0, where no through vehicles are delayed, to 1.0 where all through vehicles follow left turns and are delayed. Figure 2 illustrates the positions of through and left turns in a shared lane for the cases where  $K = 0.0, 0.5, \text{ and } 1.0$ .

## Estimating $K$

For formulas 1, 2, and others like them to have meaning in practice, it is necessary to determine how the impedance fac-

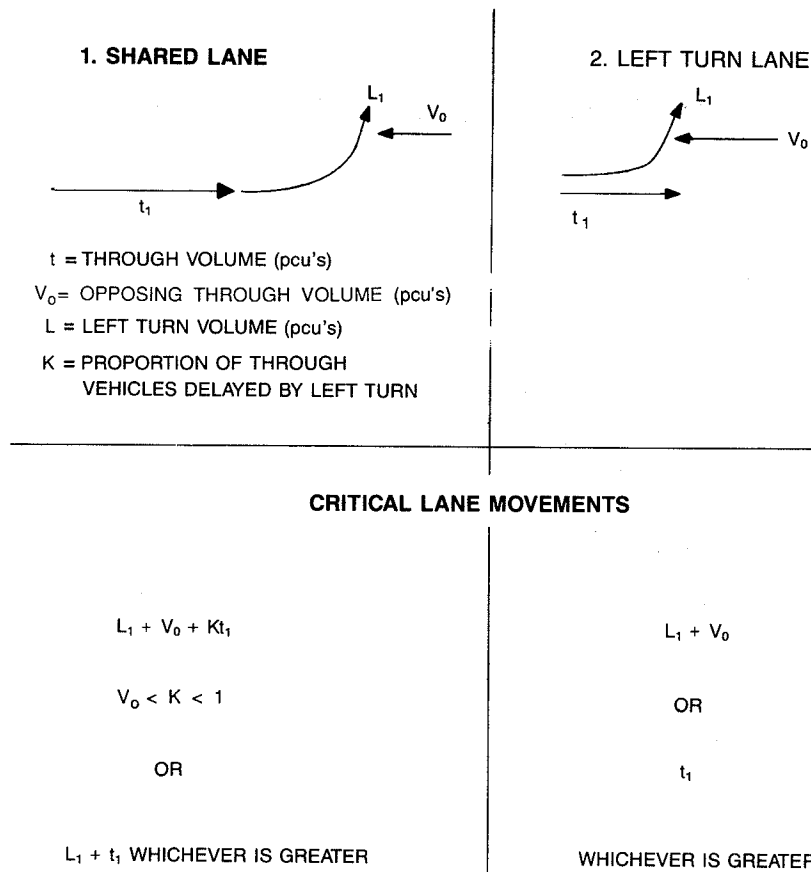


FIGURE 1 Critical movements—single lane.

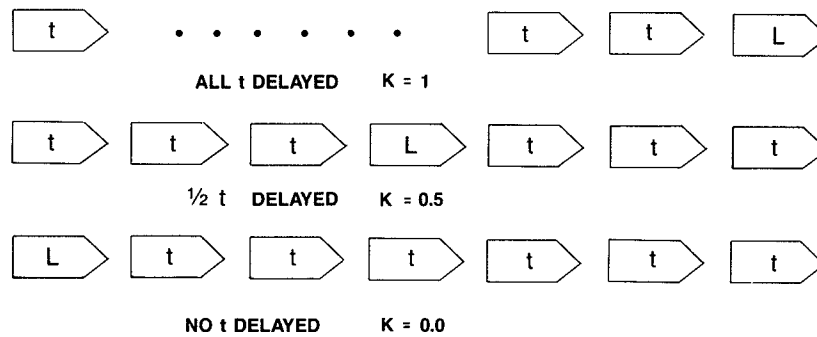


FIGURE 2 Examples of  $K$ .

tor  $K$  varies as a function of the left turns,  $L_1$ . Accordingly, values for  $K$  were obtained by two methods: (a) simulation by random numbers tables; and (b) computations based on positional probabilities assuming sampling with replacement.

*Simulation*

Random number tables were used to generate the positions of left turns and through vehicles in 10-car platoons. Fifty platoons were analyzed for each of the four cases where left turns represented 10, 20, 30, and 40 percent of the total traffic in the platoon, respectively. The average number of through vehicles and the proportion of through vehicles delayed were then computed.

*Positional Probabilities*

Probability theory was used to estimate the likelihood of the first, second, third, and  $i$ th vehicles in line being left turns. Conditional probabilities were computed, assuming that the first left turn is the  $i$ th car in line based on sampling with replacement. The probability that the first  $i - 1$  cars are through vehicles and the  $i$ th vehicle is a left turn,  $p_i$ , is given by the formula:

$$p_i = \left\{ \frac{t_1}{L_1 + t_1} \right\}^{i-1} \cdot \left\{ \frac{L_1}{L_1 + t_1} \right\} \tag{2}$$

where

$$L_1 / (L_1 + t_1) = \text{left turns as proportion of total traffic;}$$

TABLE 1 COMPUTED VALUES FOR K

LEFT TURNS AS PERCENT OF TOTAL TRAFFIC IN SHARED LANE	LENGTH OF PLATOON (Veh/Cycle)			
	5	10	15	20
1	.029	.053	.076	.094
10	.247	[.322]	.495	.576
20	.390	[.607]	.690	.757
30	.527	[.647]	.783	.834
40	.608	[.747]	.834	.874
50	.670 <sup>a</sup>	.806	.866	.900
60	.720	.838	.889	.916
70	.770	.860 <sup>a</sup>	.905	.929
80	.800	.880	.917	.938
90	---	.900	.929 <sup>a</sup>	.945
95	---	---	---	.950

[SIMULATION]

<sup>a</sup>COMPUTED VALUE ADJUSTED SLIGHTLY.

$t_1/(L_1 + t_1)$  = through vehicles as proportion of total traffic;

$L_1$  = left turns; and

$t_1$  = through vehicles.

Computed values of  $K$  are shown in Table 1 and Figure 3 for 5-, 10-, 15-, and 20-car platoons. These exhibits show how  $K$  varies as a function of platoon length and left turns as a proportion of the total approach traffic. They indicate that simulation and random sampling with replacement give similar results. Salient findings are as follows:

1. The  $K$ -values increase with increasing proportions of left turns ( $p$ ) in traffic, but they never reach 1.00. For 5-vehicle groups, the maximum is 0.80 when  $p = 0.80$ ; for 10-vehicle groups, 0.90; and for 20-vehicle groups, 0.95.
2. The values of  $K$  increase faster than the proportions of

left turns in the shared lane. For example, when the left turns account for half of the vehicles in the shared lane, then from 67 percent to 90 percent of the through vehicles in that lane would be delayed, depending on the platoon (queue) length.

3. The values of  $K$  increase with queue length for any given proportion of left turns in the lane. This suggests that a shorter traffic signal cycle length would reduce delays whenever left turns share a lane with through vehicles. For example, a lane that carries 480 through vehicles and 120 left turns in an hour ( $p = 20$  percent) would result in a  $K$ -value of .584 for a 60-second cycle (10 vehicles per cycle) and .690 for a 120-second cycle (20 vehicles per cycle). The number of through vehicles that would be delayed would be 280 and 331, respectively.

4. The curves follow a formula of the form  $K = [A(p)]$ ; pilot analysis suggests that  $K = p^{(.55 - 0.020)}$  where  $Q$  = estimated vehicles/lane/cycle in the queue and  $p$  = proportion of left turns in the shared lane.

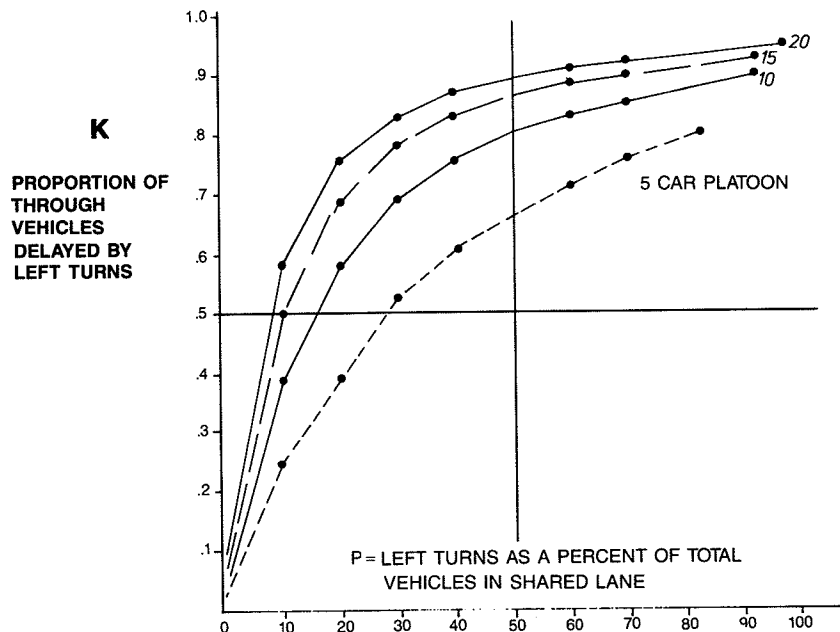


FIGURE 3  $K$  as a function of  $p$ .

TABLE 2 SUGGESTED  $K$ -VALUES FOR APPLICATIONS

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

Source: Computed

 **$K$ -Values for Application**

Table 1 shows that the  $K$ -values can be represented by a single set of numbers that are a function of the number of left turns per signal cycle or platoon. The suggested  $K$ -values are shown in Table 2 and Figure 4. They provide a reasonable approximation of left turn blocking for most practical conditions, and they significantly simplify the computational procedures, especially for multilane approaches.

The curve shown in Figure 4 can be approximated by the formula:  $K = 1 - e^{-.75 \sqrt{L_1}}$ . It also is interesting to note that  $K \leq L_1 / (L_1 + 1)$ . In both formulas,  $L$  represents the left

turns per cycle. The values of  $K$  increase rapidly at first and then begin to taper off in the following situations:

- When one left turn per cycle occurs, approximately 40 percent of the through vehicles in the shared left turn lane would be blocked.
- When three left turns per cycle occur, approximately 70 percent of the through vehicles in the shared left turn lane would be blocked.
- When five left turns per cycle occur, approximately 80 percent of the through vehicles in the shared left turn would be blocked.

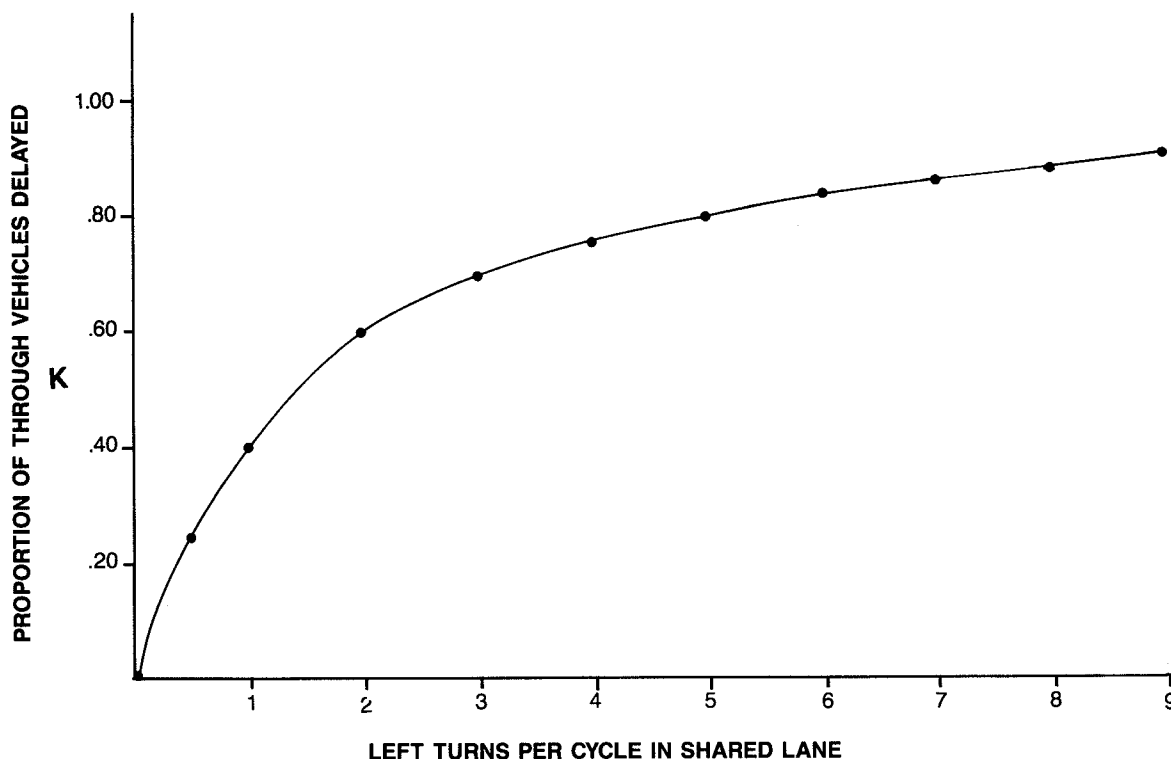
The  $K$ -values should be applied directly to the through traffic in the shared lane, and the product should be added to the opposing through volume and left turn volume to obtain the critical lane volumes for the artery. They also should be used for the cross street; the total critical lane movements at an intersection would represent the sum of the critical artery and cross street movements.

The application of these  $K$ -values is straightforward. If there are 12 through vehicles per lane per cycle, three left turns, and eight opposing vehicles, the critical movement would be estimated as follows:

$$\begin{aligned} \text{Critical movement} &= 3 \text{ (left)} \\ &+ 8 \text{ (opposing)} + 12 \text{ (through)} \cdot (K) \end{aligned}$$

Since  $K = .70$ , the critical movement would be  $11 + 8.4$  or 19.4 vehicles per cycle.

This analysis does not take into account the added time required by each left turn. Thus, if the computed critical lane flows (for the artery) are  $L + V_o + K_{l1}$ ,  $L$  should be increased by a factor of  $F$  to account for the increased headway.

FIGURE 4 Left-turn impedance factor ( $K$ ).





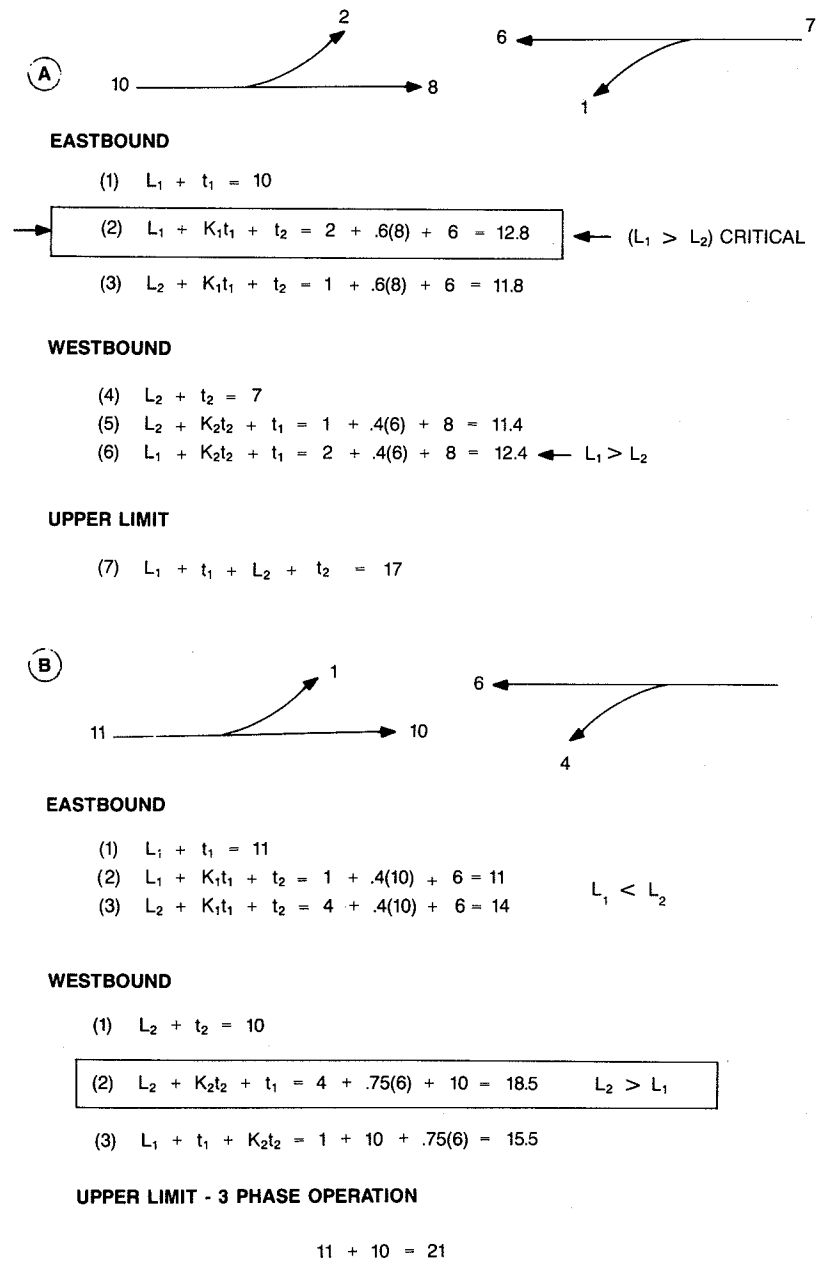


FIGURE 7 Example 2—single shared lanes, left turns on both approaches.

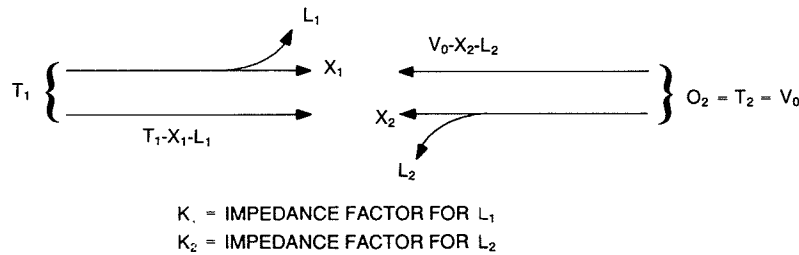
various distributions of through and turning vehicles for each approach on a four-lane arterial with left turns in both directions. The results of this simulation, shown in Table 3, provide a basis for the guidelines that follow.

In practice, the following steps will prove useful in dealing with two-lane approaches when left turns occur from each approach:

1. Divide the volumes on each approach equally by lane.
2. Compute the critical lane volumes on each approach based on the assumed lane distribution. The computations are as follows: Left turns + opposing through traffic in the shared lane (or + the opposing outside lane volume) +  $K$  (same direction through traffic in the shared lane). This will yield a critical conflict volume for each of the two approaches.

3. Select appropriate critical lane volumes based on the following criteria:

- |   |   |  |
|---|---|--|
| Case 1: Equal volumes on both approaches (including equal left turns)   | } | Use critical lane volume for either approach         |
| Case 2: Unequal through volumes but equal left turns on both approaches |   | Use heavier critical lane volume                     |
| Case 3: Heavy total volumes and heavy left turns on one approach        | } | Use average critical lane volume for both approaches |
| Case 4: Equal total volumes on both approaches with unequal left turns  |   |  |
| Case 5: Light total volumes with heavy left turns on both approaches    |   |  |



'EQUILIBRIUM SOLUTION'

$$(1) \quad \frac{(\text{LEFT TURN})_1 + \text{OPPOSING} + \text{DELAYED THROUGH}}{L_1 + V_0 - X_2 - L_2} + K_1 X_1 = L_2 + (T_1 - X_1 - L_1) + K_2 X_2$$

SOLVING FOR  $X_2$

$$X_2 = \frac{2(L_1 - L_2) + (V_0 - T_1) + X_1(1 + K_1)}{1 + K_2}$$

CONSTRAINT -  $X_2 \geq 0$

IF  $T_1 \geq T_2$ , IT CAN BE SHOWN EXPERIMENTALLY

THAT  $X_1 = \frac{T_1 - L_1}{2}$ . THIS MAKES IT POSSIBLE TO SOLVE FOR  $X_2$ .

FIGURE 8 General formula—two-lane approach with left turns.

Figure 9 illustrates critical lane computations for two-lane approaches with shared left turn lanes (Example 3). Critical movements are also computed based on a signal operation that has each direction move on a separate phase; this represents the worst case or upper limit of the critical lane movement.

This example represents a Case 3 condition. The critical lane volume of 990 vph compares with a worst case condition of 1,080 vph.

**Multilane Approaches—Left Turns from One Direction Only**

These cases can be solved directly by formula assuming equal queues in each lane on the multiple lane approach (Figure 10). Formulas are given in Table 4 for estimating the through vehicles in the shared lane. The through vehicles in the other lanes can be computed; these flows will equal the critical lane volume. Figure 11 illustrates this procedure (Example 4).

**CONCLUSIONS AND EXTENSION**

This paper analyzes the impact of left turns on through traffic where they share a common lane. It applies an impedance or blockage factor to estimate the proportion of through vehicles

that would be delayed and, therefore, must be considered in the critical lane computations. In this respect it represents a logical extension of earlier capacity analysis, and it is consistent with ongoing approaches.

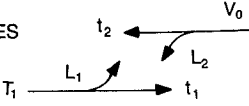
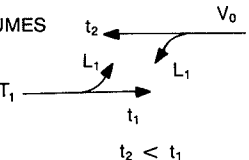
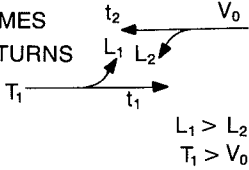
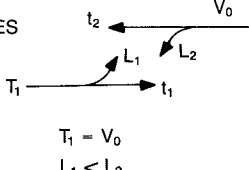
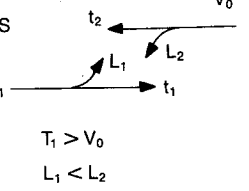
The left turn blockage factor ( $K$  factor) eliminates the need for left turn equivalent factors. In other respects, the procedures are generally similar to the critical lane computational procedures set forth in the 1985 *Highway Capacity Manual*. Right turns, for example, might be excluded from critical movement analyses under certain geometric conditions.

Several findings are significant: (a) Left turns in a shared lane block through vehicles in that lane; (b) When there are more than five or six left turns per cycle, for all practical purposes, they pre-empt the shared lane; and (e) Short traffic signal cycles are desirable where shared left turn lanes predominate.

The left turn impedance or blockage factor,  $K$ , provides an important input into deriving an intersection capacity formula that directly reflects the capacity losses due to blocked vehicles. Such a formula has been developed including an approximation for practical application.

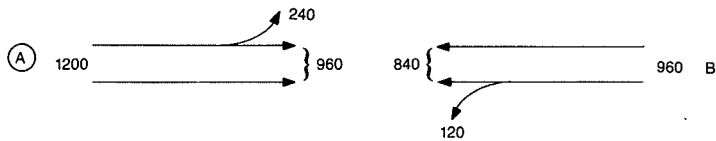
The impedance factors are based on probability and simulation analysis. They can be modified based on local experience or field tests without invalidating the methods. Additional field studies are desirable to verify the analyses and to adjust them as needed. Further refinements for right turns and for advance or trailing greens also are desirable.

**TABLE 3 GUIDELINES FOR ALLOCATING THROUGH TRAFFIC TO SHARED LEFT-TURN LANE—TWO-LANE APPROACH**

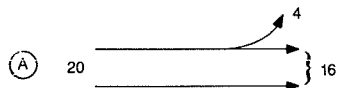
<b>CASE 1</b>	EQUAL VOLUMES EQUAL LEFT TURNS		1. DIVIDE APPROACH VOLUMES EQUALLY BY LANE 2. COMPUTE CRITICAL VOLUME ON EACH APPROACH
<b>CASE 2</b>	UNEQUAL VOLUMES EQUAL LEFT TURNS		1. DIVIDE APPROACH VOLUMES EQUALLY BY LANE 2. COMPUTE CRITICAL VOLUMES ON EACH APPROACH 3. USE HEAVIER VALUE (1-5% OVERSTATEMENT)
<b>CASE 3</b>	HEAVY VOLUMES HEAVY LEFT TURNS		1. DIVIDE APPROACH VOLUMES EQUALLY BY LANE 2. COMPUTE CRITICAL VOLUMES ON EACH APPROACH 3. USE HEAVIER VALUE (1-3% OVERSTATEMENT)
<b>CASE 4</b>	EQUAL VOLUMES UNEQUAL LEFT TURNS		1. DIVIDE APPROACH VOLUMES EQUALLY BY LANE 2. COMPUTE CRITICAL VOLUMES ON EACH APPROACH 3. USE AVERAGE VALUE (1-2% UNDERSTATEMENT)
<b>CASE 5</b>	LIGHT VOLUMES HEAVY LEFT TURNS		1. DIVIDE APPROACH VOLUMES EQUALLY BY LANE 2. COMPUTE CRITICAL VOLUME ON EACH APPROACH 3. USE AVERAGE VALUE (2-5% UNDERSTATEMENT)

NOTE: IN SOME CASES HEAVY  
 LT WILL PRE-EMPT 1 LANE

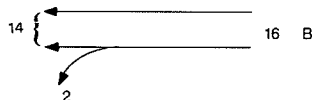
1. HOURLY FLOW RATES



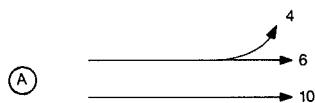
2. VEHICLES PER CYCLE:



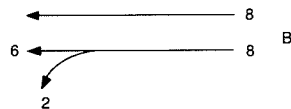
60-SEC CYCLE



3. CRITICAL LANE VOLUMES -



EQUAL LANE USE



(K)

$$(A) \quad 4 + 8 + .75(6) = 16.5$$

$$(B) \quad 2 + 10 + .60(6) = 15.6$$

4. SELECT CRITICAL LANE VALUE TO BE USED: SINCE BOTH LEFT TURNS AND TOTAL TRAFFIC IS HEAVIEST ON APPROACH A, THIS IS CASE 3 THEREFORE USE HEAVIEST VOLUMES , 16.5

5. COMPUTE HOURLY VOLUME

$$16.5 \cdot 60 = 990 \text{ vph}$$

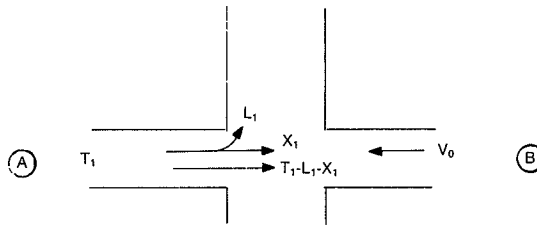
6. CHECK "WORST CASE"

THIS ASSUMES THAT EACH DIRECTION OPERATES ON A SEPARATE PHASE:

$$\frac{20}{2} + \frac{16}{2} = 18.0 / \text{CYCLE}$$

$$18.0 \times 60 = 1080 / \text{HOUR}$$

FIGURE 9 Example 3—two-lane approach.



LET:

$T_1$  = TOTAL TRAFFIC (ON APPROACH A )

$L_1$  = LEFT TURNS (ON APPROACH(A))

$X_1$  = THROUGH VEHICLES IN SHARED LANE (APPROACH A )

$V_0$  = OPPOSING VOLUME (APPROACH B )

$K_1$  = LEFT TURN IMPEDANCE FACTOR

CONSIDERING THE TWO LANE CASE:

THE CRITICAL LANE VOLUME IS EITHER

$$T_1 - L_1 - X_1 \quad (1)$$

OR

$$L_1 + \frac{V_0}{2} + K_1 X_1 \quad (2)$$

FOR THE ASSUPTION OF EQUAL QUEUES,  
THESE VALUES ARE SET EQUAL

$$\text{i.e. } T_1 - L_1 - X_1 = L_1 + \frac{V_0}{2} + K_1 X_1$$

SOLVING FOR  $X_1$ , THIS YIELDS

$$X_1 = \frac{T_1 - 2L_1 - V_0}{1 + K_1}$$

PROVIDED THAT  $X_1 \geq 0$

IN THE THREE LANE CASE:

$$\frac{T_1 - L_1 - X_1}{2} \text{ IS SET EQUAL TO}$$

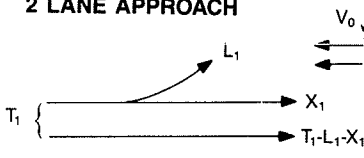
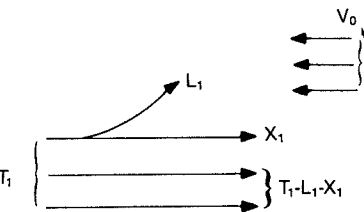
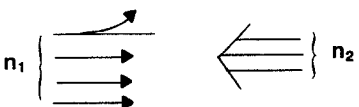
$$L_1 + \frac{V_0}{3} + K_1 X_1$$

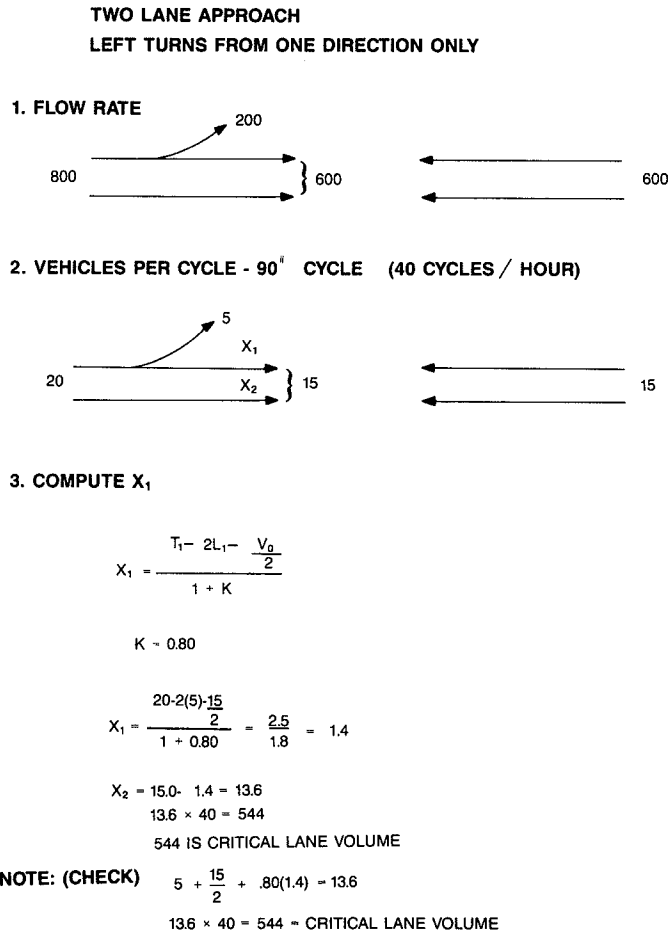
WHERE THERE ARE  $n_1$  LANES ON APPROACH  
A, AND  $n_2$  LANES ON APPROACH B,

$$\frac{T_1 - L_1 - X_1}{n_1 - 1} \text{ IS SET EQUAL TO } L_1 + \frac{V_0}{n_2} + K_1 X_1$$

**FIGURE 10** Derivation of formulas for shared left-turn lanes on multilane approach with left turns in only one direction.

TABLE 4 FORMULAS FOR MULTILANE ROADS (APPROACHES) WITH LEFT TURNS FROM ONE APPROACH ONLY—LANE DISTRIBUTION

<p><b>2 LANE APPROACH</b></p> 	$X_1 = \frac{T_1 - 2L_1 - V_0/2}{1 + K_1}$ <p>(A)</p> $X_1 \geq 0$
<p><b>3 LANE APPROACH</b></p> 	$X_1 = \frac{T_1 - 3L_1 - 2V_0/3}{1 + 2K_1}$ <p>(B)</p> $X_1 \geq 0$
<p><b>MULTI-LANE APPROACH</b> <b>GENERAL FORMULA</b></p> 	$X_1 = \frac{T_1 - n_1 L_1 - \frac{n_1 - 1}{n_2} V_0}{1 + (n_1 - 1)K_1}$ <p>(C)</p> $X_1 \geq 0$ <p><math>n_1 =</math> NO. OF LANES ON APPROACH 1  <math>n_2 =</math> NO. OF LANES ON APPROACH 2  <math>K_1 =</math> LEFT TURN IMPEDANCE FACTOR (DIRECTION 1)</p>



**FIGURE 11** Example 4—two-lane approach, left turns from one direction only.

## REFERENCES

1. B. D. Greenshields, D. Schapiro, and E. L. Erickson. *Traffic Performance at Urban Street Intersections*. Technical Report No. 1. Yale Bureau of Highway Traffic, 1946.
2. *Highway Capacity Manual*. HRB, National Research Council, Washington, D.C., 1965.
3. *Transportation Research Circular 212: Interim Capacity Materials*. TRB, National Research Council, Washington, D.C., 1980.
4. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.

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



# Computation of Signalized Intersection Service Volumes Using the 1985 Highway Capacity Manual

KENNETH G. COURAGE AND JOHN ZEN-YOUNG LUH

The 1985 Highway Capacity Manual (HCM) signalized intersection analysis method has provided a powerful technique for determining how well an intersection will operate given the traffic volumes, intersection configuration, and signal timing. It has been less successful in producing a practical technique for determining the service volume for a specified level of service. The main problem is the irreversibility of the delay equation upon which the level of service is based. This paper explores a number of numerical analysis techniques for solving the delay equation in reverse. Specific computational procedures are demonstrated for determining the traffic volume which will produce a given delay. Both manual methods, using a worksheet, and computerized solutions are presented. The main problem with determining service volumes results from the interactions among the independent variables of the delay equation, not from the need for a mathematical solution to the equation itself. Because of the complexities of the HCM technique, there are discontinuities in the volume-delay relationship. There are situations which require the simultaneous solution of two equations, and there are situations in which a unique solution does not exist. Several simple computational modules were developed to deal with specific situations covered in this paper. These are demonstrated using several sample calculations from Chapter 9 of the HCM.

The 1985 Highway Capacity Manual (HCM) (1) has introduced a completely new analysis technique for signalized intersections. There are two major differences between the 1985 method and its predecessors. The first is a substantially more complex analysis procedure, especially for dealing with opposed left turns. The second is a redefinition of the level-of-service (LOS) criterion from a demand/capacity base to a vehicular delay base.

For most purposes, this has been a step forward. It has defined the quality of traffic movement in the same way that the motorist perceives it (*i.e.*, delay). Previous definitions were essentially mathematical indexes from the motorist's point of view. It has also improved the accuracy and reliability of the analysis model. The overall methodology has become widely accepted throughout the United States.

Since 1965, the HCM has suggested that the method may be used to assess the level of service of a facility with known design parameters and traffic volumes or to determine the design parameters or traffic volumes which would produce a given level of service. The concept of "service volume" was

introduced in the 1965 HCM (2) as a more flexible alternative to the simple notion of "capacity." The service volume of a particular facility reflects the volume of traffic which can be accommodated at a specified level of service. It clearly and concisely expresses the adverse effect of introducing additional traffic onto the facility.

Service volume has been recognized as a useful concept and, as such, has found its way into the official analysis procedures of governmental agencies. For example, in Florida, the criterion of 10 percent of the service volume for LOS D is applied to determine the status of a proposed traffic generator with respect to the assessment of impact fees (3). The difficulty of computing service volume by the 1985 HCM has led many agencies in Florida to resort to a curious blend of 1985 and pre-1985 analysis techniques. The need for procedures which use the 1985 methodology for computing service volume is well recognized in Florida and throughout the country. That is the subject of this paper.

In the following discussion, the problem will be identified and explored. A basic method will be suggested for dealing with the simple case. The complexities which surround the simple case will be explained. Finally, specific computational procedures will be proposed, both for worksheet and computerized solutions.

## THE PROBLEM

The computational model used in the 1985 HCM is illustrated in Figure 1. The essential ingredients of this model are four inputs:

- Cycle length,  $C$ ;
- Green time,  $g$ ;
- Capacity,  $c$ ; and
- Volume to capacity ratio,  $X$ ;

and one output, delay. The delay is an explicit but formidable function of the four input variables which appears in the HCM as equation 9-18.

The complexity of the function poses no problem for computing delay from known values of the four inputs. The equation, while time consuming in manual applications, lends itself well to computerization, and an ample choice of software is available for this purpose.

Determining the service volume (*i.e.*, the traffic volume required to produce a specified delay) is not an easy task. To

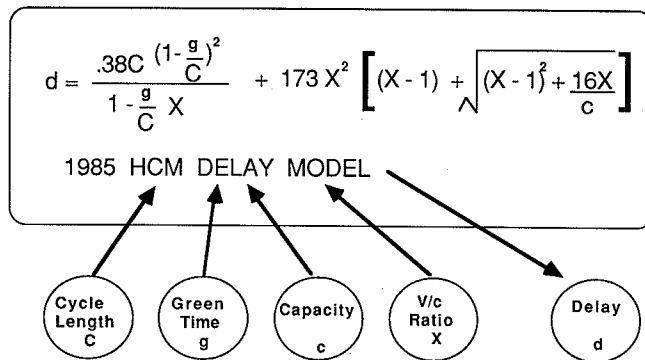


FIGURE 1 1985 *Highway Capacity Manual* equation for computation of delay at signalized intersections.

begin with, volume itself is not a specific input to the delay equation. It may, however, be derived directly from two of the inputs, capacity and volume to capacity ( $v/c$ ) ratio. So the problem is to find, using the delay equation, a  $v/c$  ratio which will produce a specific delay given values for cycle length, green time, and capacity.

#### Analytical Solution

The most desirable solution would be an analytical one in which the terms of the delay equation were algebraically manipulated to yield the value of  $X$  as an explicit function of the other terms. There is no record in the literature of this feat having been accomplished, and the complexity of the equation suggests that the analytical solution be left as a challenge to the academic world.

#### The Exhaustive Search Solution

The exhaustive search, or "brute force" technique, is illustrated in HCM sample calculation 6, in which a complete table of values is developed for all levels of service and all arrival types for a single lane group involving only through movements. The HCM points out that the example is presented for purposes of illustration only, and does not actually recommend the technique for addressing complex situations involving more than one lane group. The method of presentation of this example in the HCM demonstrates very clearly the concept of solving the delay equation implicitly for one of the input variables.

#### The Trial-and-Error Solution

The trial-and-error solution is always a legitimate one for working a problem in reverse. With sufficient time and effort, a satisfactory combination of inputs should be found which will produce a specified delay. It is certainly possible to use one of several microcomputer programs which perform highway capacity calculations in a trial-and-error mode. This would, however, be very time consuming, especially when several approaches must be treated simultaneously. In this paper, a more efficient and productive method will be sought.

#### The Nomographic Solution

A nomographic procedure has been implemented for the 1965 method (4) and updated for the 1985 method (5). The 1965 procedure was widely used for determining service volumes for a single approach. The 1985 HCM method deals concurrently with all approaches to an intersection. This is a major departure from the 1965 method, which treated each approach independently.

The nomographs provide an excellent graphic method which is reversible for the "simple case" (which will be treated later in this discussion). However, it will also be demonstrated that there are several practical complexities for which more automated computational techniques provide a more feasible method of solution.

#### The Numerical Solution

A solution using numerical methods offers a more scientific approach to the problem. Numerical methods involve an iterative procedure in which a solution is sought by multiple trials according to a well-defined set of rules. In this paper, three numerical solutions will be introduced. The first is a primitive method which is suited to manual implementation using a worksheet. The second two offer more elegant techniques which are too complex for manual implementation but which are well-suited to the microcomputer.

#### A BASIC PROCEDURE FOR A SIMPLE CASE

The discussion to this point has centered on the complexities of the delay equation itself. It will become apparent, however, as the solution develops that the real complexities are found in the interdependence among what appear to be the independent variables of the delay equation. Before addressing these practical complications, let's deal with the simple case. The "simple case" is defined as one in which the four "independent" variables are truly independent. By this definition, we must restrict the analysis to single lane groups with Type 3 arrivals and no permitted left turn movements.

#### Data Requirements

The following data are required for the basic numerical solution:

- Cycle length,  $C$ ;
- Effective green time,  $g$ ;
- Adjusted saturation flow rate,  $s$ ;
- Capacity,  $c$ ; and
- Target delay,  $D_T$ .

The cycle length and effective green time are input data and must be determined prior to the analysis. The adjusted saturation flow rate may be determined either from the worksheet given in the HCM for this purpose or by one of several available software choices which performs HCM computations. The lane group capacity is computed as:

$$c = s \times \frac{g}{C}$$

The target delay would normally be the upper threshold for the desired level of service. Before specifying a particular value of delay, it is necessary to verify that the value is a feasible solution to the delay equation for the 0-100 percent range saturation. This is accomplished by evaluating the equation separately for  $X = 0$  and  $X = 1$ , from which the lowest feasible delay is computed as:

$$d = 0.38/C (1 - g/C)^2$$

and the highest feasible delay becomes:

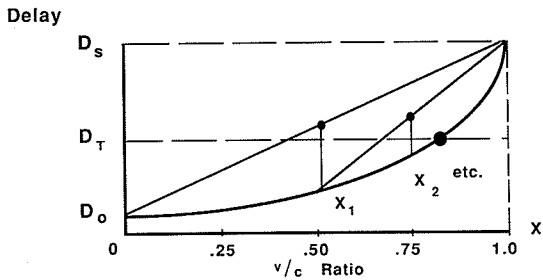
$$d = 0.38C (1 - g/C) + 173 \sqrt{\frac{16}{c}}$$

This raises an important point: A service volume does not always exist for all levels of service.

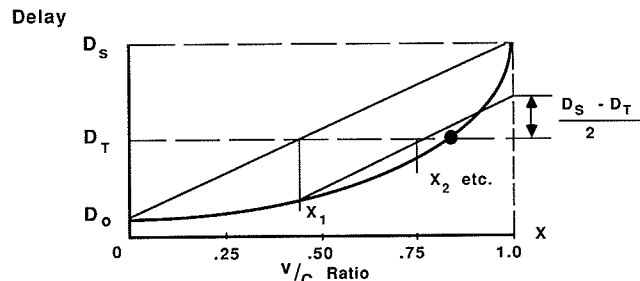
**Service Volume Computation**

The first step in the service volume computation is to find the value of  $X$  which will produce the target delay. This requires a choice of numerical analysis techniques. The Bolzano, or bisection, method (6) will be used initially to illustrate the process. This method uses the following steps, which are illustrated in Figure 2:

1. Determine the upper and lower limits of  $X$  for the solution (initially 0 and 1). The delays at these end points are shown in Figure 2 as  $D_0$ , for  $X = 0$  and  $D_s$ , for  $X = 1$  (i.e., saturated operation).
2. Evaluate the delay at a point halfway between the upper and lower limits. This point is shown on Figure 2 as  $X_1$  for the first iteration and  $X_2$  for the second.



The Bisection Method for Worksheet Solutions



The Modified Linear Interpolation Method for Computer Solutions

3. If the delay at the midpoint is lower than the target delay  $D_T$ , then the midpoint becomes the new lower limit. If not, then the midpoint becomes the new upper limit.

4. Repeat steps 1 to 3 until the delay of the midpoint is within the required degree of accuracy with respect to the target delay.

When a suitable value of  $X$  has been found, the next step is to compute the service volume (i.e., multiply  $X$  by the lane group capacity).

**Computational Considerations**

The maximum number of iterations to achieve a solution will depend on the required accuracy of the results. The maximum error as a percentage of the desired delay will be:

$$\frac{100}{2^N}$$

where  $N$  is the number of iterations. So, for example, to ensure a maximum error of one percent as many as seven iterations could be required.

A worksheet for the bisection method is shown in Figure 3. This is a relatively primitive method, and its main advantage is conceptual simplicity. Another numerical analysis technique, known as the modified linear interpolation method (7), offers faster convergence (i.e., fewer iterations) but is too involved for manual worksheet computations. The modified linear interpolation method is also illustrated in Figure 2, which shows the similarities and differences with respect to the bisection method. The main advantage of the modified linear interpolation method is that the  $X$ -value chosen for a

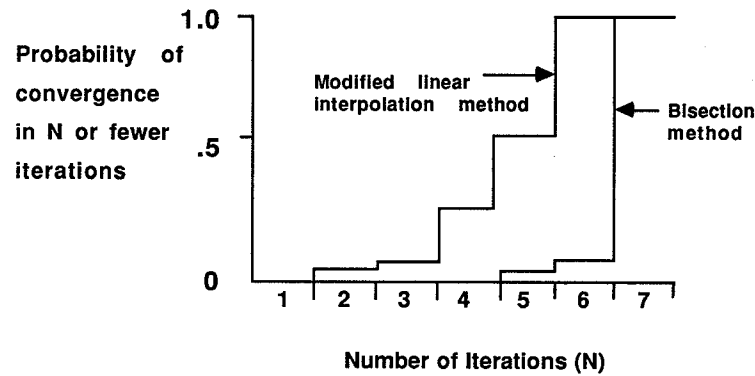
Service Volume Worksheet :						① Target Delay _____	
Bisection Method						② Maximum Error _____	
Iteration Number	③	④	⑤	⑥	⑦	⑧	⑨
	Lower Limit	Upper Limit	Mid Point	Error			
	X	Delay	X	Delay	X	Delay	
1							
2							
3							
4							
5							
6							
7							

⑩ Final X .....	_____
⑪ Capacity .....	_____
⑫ Service Flow Rate .....	(⑩ x ⑪)
⑬ PHF .....	_____
⑭ Lane Utilization Factor .....	_____
⑮ Service Volume .....	(⑫ x ⑬) / ⑭

FIGURE 3 Worksheet for computation of service volume by the bisection method.

FIGURE 2 Numerical methods for solving the delay equation to obtain the  $v/c$  ratio which will produce a specified delay.



**Note: The delay equation was solved 10,000 times by each method**

**FIGURE 4** Number of iterations required to solve the delay equation by two different numerical analysis methods.

particular iteration depends on the relative distance of the target delay from the two end points in the previous iteration. In the bisection method, the  $X$ -value is always chosen halfway between these two end points, regardless of their relative distance from the target delay. In the modified linear interpolation method, the  $X$ -value for the second iteration is placed at a distance from the end point, which is computed by:

$$X_2 = 1 - \frac{D_s - D_T}{2} \times \frac{1 - X_1}{\frac{D_s - D_T}{2} - (D_1 - D_T)}$$

where  $D_s$  and  $D_1$  are delays for upper limit 1.0 and lower limit  $X_1$ , respectively.

Values for subsequent iterations are computed in the same manner. While the manual implementation of this method would be very tedious, the rapid convergence makes it ideally suited to computerized solutions. A comparison of the convergence characteristics of these two methods is shown in Figure 4.

To produce this comparison, the delay equation was solved 10,000 times by both methods to determine the  $v/c$  ratio which would produce a randomly generated value of the target delay. Figure 4 shows the distribution of the number of iterations required to reach a solution. It is clear from this figure that the modified linear interpolation method is the more efficient of the two.

### PRACTICAL COMPLICATIONS

The methodology discussed to this point assumes that the four inputs are independent of each other and of the traffic volume on the lane group. These simplifying assumptions are not valid in situations with permitted left turns or platoon arrival characteristics which are affected by progression. Further complications arise when two lane groups with opposing left turns must be considered at the same time.

There are four areas of interaction between the independent variables in the delay equation and the traffic volume on a particular lane group:

- Interaction within a lane group,
- Interaction between two lane groups on one approach,
- Interaction between two opposing approaches, and
- Interaction among all approaches to an intersection.

There is no interaction within a lane group except when the arrivals are affected by progression. This is evident in table 9-13 of the HCM, which illustrates that the progression adjustment factor,  $PF$ , is a function of the  $v/c$  ratio for all arrival types other than Type 3. The result is a discontinuous delay function which requires a piecewise solution.

The main interaction between two lane groups on one approach occurs when one lane is shared by left turns and through movements. When heavy left turn movements monopolize a shared lane to the point where it becomes a de facto left turn lane, it is necessary to break the approach into two separate lane groups. If all other factors remain constant, the break point will occur at a specific value of  $X$ . This introduces another discontinuity into the delay equation.

The interaction between two opposing approaches is more complicated. In addition to the de facto left turn lane problem, there are two elements in the HCM methodology which combine to make the saturation flow adjustment factor for one left turn a function of the traffic volume in the opposing direction. First, the left turn equivalency is a function of the volume of the opposing through movement. This means that a given left turning volume will place a higher demand on its own approach as the volume increases on the opposing approach. Second, the green time available for the left turn movement will be reduced as the opposing volume increases, because more time will be required to service the queue of opposing vehicles which have arrived on the red signal. This will have the same effect on the left turn demand.

One final interaction occurs among all approaches as traffic volumes vary. That, of course, is the signal timing itself. Variations in traffic volume will result in variations in the appropriate cycle lengths, green times, and even offsets. Signal timing changes will have a significant effect on all of the inputs to the delay equation.

Because of all of the interdependencies of the supposedly independent variables in the delay equation, and because of

the dependency of all of them on traffic volume, it is very difficult to prescribe a straightforward mathematical process by which service volume may be computed "in reverse" as a function of delay. There are, however, some practical techniques which may be used to illustrate some of the complexities and to simplify others. The balance of this discussion will explore those techniques.

Before proceeding with the discussion, it is necessary to define precisely what is meant by the 1985 HCM technique. The technique is considered in this paper to consist of three elements. The first is the narrative description given in the body of Chapter 9 of the HCM. The second is the interpretation of that narrative description contained in the six sample calculations presented at the end of the chapter. The third is the Highway Capacity Software (HCS) developed by the Federal Highway Administration as a faithful implementation of the HCM method. The HCS will serve as a reference for verifying the accuracy of computational techniques to be introduced in this paper.

### DEALING WITH SIGNAL PROGRESSION

Uncoordinated signals exhibit random arrivals which are defined by the HCM as "Type 3." Coordinated signals usually produce a more favorable distribution of arrivals throughout the cycle (Type 4 or 5) on links which are given a high design priority and a less favorable distribution (Type 1 or 2) if the design priority is low. The HCM assigns a progression adjustment factor, which modifies the delay computed by the delay equation. The PF depends on the arrival type, the type of signal control equipment, and the  $v/c$  ratio. The dependency on the  $v/c$  ratio is of greatest interest to this discussion because it introduces a relationship between  $v/c$  and delay which is not treated by the delay equation. Therefore, even the most sophisticated method of solving the delay equation backwards will not produce a service volume, except for Type 3 arrivals, in which the progression adjustment factor is independent of the  $v/c$  ratio.

When arrivals are affected by progression, the relationship between  $v/c$  and delay will be discontinuous as illustrated in Figure 5. This figure shows an example of the delay as a function of  $v/c$  for each of the five arrival types. The example assumed the following values of the input data:

Cycle length, $C$	= 90 sec,
Green time, $g$	= 45 sec,
Saturation flow, $s$	= 3,200 vphg, and
Capacity, $c$	= 1,600 vph.

Notice that there are two types of discontinuities; those with no solution and those with multiple solutions. Arrival Types 1 and 2 produce gaps with multiple solutions while arrival Types 4 and 5 produce gaps with no solution. This is evident from the structure of HCM table 9-13 and holds true for all cases.

The discontinuities shown in Figure 5 are based on the assumption that the progression factors are determined as threshold values from HCM table 9-13. The discontinuities could be eliminated by interpolation, however, interpolation is discouraged by the HCM in the presentation of sample calculation number 1, and the HCS method is based strictly on thresholds.

So, before applying the numerical techniques described earlier to arrivals affected by progression, it is necessary to determine the  $v/c$  range in which the solution will be found. If the desired value of delay falls in only one  $v/c$  range, then the appropriate numerical method may be applied using the end points of the  $v/c$  range as the initial upper and lower limits. If the desired value of delay falls into a gap between  $v/c$  ranges, then it may be argued that the  $v/c$  transition point (either .6 or .8) is, in itself, the solution. Finally, if two  $v/c$  ranges contain the desired delay, then there is no unique service volume which can be defended as the solution, which results from a strict application of the 1985 HCM, especially as it is implemented in the HCS.

### DEALING WITH PERMITTED LEFT TURNS

Permitted left turns, according to the HCM definition, are given a green signal simultaneously with the opposing through traffic. When the signal turns green, the queue of opposing through vehicles must be serviced before the left turns may proceed. When the queue of opposing through vehicles has dissipated, the remainder of the green interval is available to service the opposing through vehicles, at their arrival rate, and the left turns at a rate which depends on the volume of opposing through traffic. A mathematical model of this process is incorporated in the HCM as a combination of table 9-12 plus a supplemental worksheet. The result is a left turn adjustment factor,  $f_{LT}$ , which is used in the computation of saturation flow rate for the lane group. In cases where a heavy left turn is opposed by a heavy through volume, the left turn adjustment factor will have a substantial impact on the saturation flow rate.

The problem, then, is that the saturation flow rate, which is one of the inputs to the delay equation, is influenced by the  $v/c$  ratio, which is for our purposes, the output. If each approach is examined separately (*i.e.*, the conditions on all other approaches are held constant), the complication is not as severe as the more practical case in which the opposing approaches are treated simultaneously. The problem with treating the approaches separately is that the service volumes computed for any one approach would only be valid for one set of volumes on the opposing approach. If the objective is to find the service volumes for both approaches, then a simultaneous solution of two equations must be pursued.

The solution to this problem for any pair of opposing approaches (say northbound and southbound) will be of the form:

$$D_N = f_1(X_N, X_S)$$

$$D_S = f_2(X_N, X_S)$$

for given values of all of the other inputs to the delay equation where  $X_N$  and  $X_S$  are the  $v/c$  ratios for the northbound and southbound approaches;  $D_N$  and  $D_S$  are the delay values for the northbound and southbound approaches. This yields a set of two equations in two unknowns, more commonly written in the form:

$$f_1(X_N, X_S) - D_N = 0$$

$$f_2(X_N, X_S) - D_S = 0$$

These two equations may be solved using Newton's method

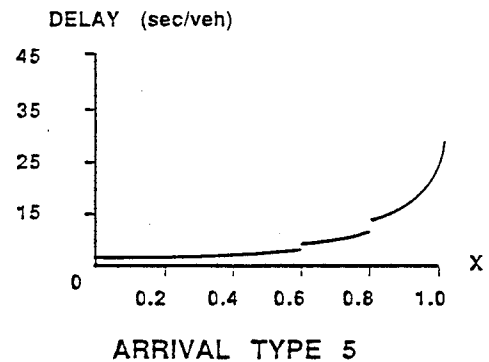
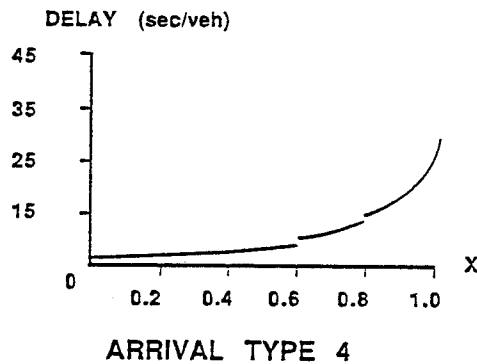
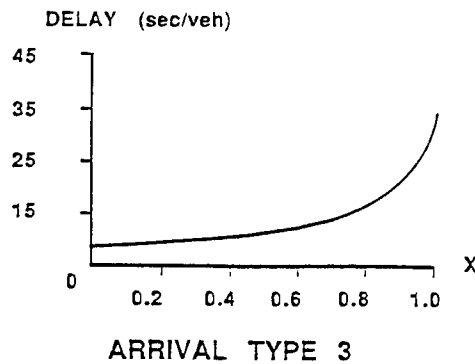
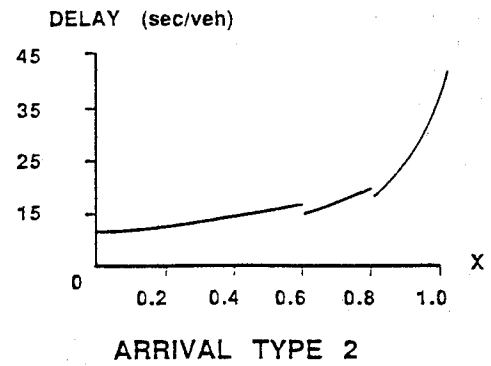
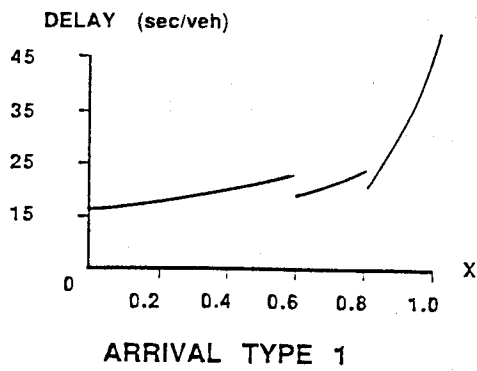


FIGURE 5 Discontinuities in the relationship between delay and  $v/c$  ratio ( $X$ ) produced by different arrival types.

for the simultaneous solution of two equations (7). This numerical technique is easily programmed on a microcomputer, but, because of its complexity, the textbook description will not be repeated here. An example of the solution will be presented later.

#### DEALING WITH PROTECTED PLUS PERMITTED LEFT TURNS

Left turns which proceed on a protected phase (green arrow) and a permitted phase (solid green) present the most complicated analysis of all because of the need to apportion the

volume between the two phases based on judgment. There are three cases to consider. The first, in which all left turns are assigned to the protected phase, is really no more complicated than the protected-only treatment from a computational point of view. The only real problem is that, as pointed out by Bonneson and McCoy (8), the HCM method handles the green split computations for this situation in a questionable manner. The second case, in which the volume apportionment is constant between the two phases, poses no real computational difficulty except that it requires separate computations for each phase. The third case, in which the volume apportionment between the two phases is a function of the total approach volume, cannot be handled by the numerical

techniques described in this paper. In this case, trial and error methods must be applied as a last resort.

## COMBINATIONS OF LEFT TURN TREATMENTS

It is quite possible that two opposing approaches which must be treated simultaneously will have different configurations with respect to left turn treatment and lane group formation. In many cases, the computations may be performed independently, solving a single equation. The simultaneous solution of two equations is required only when the volumes on two different lane groups exert a mutual effect on each other's saturation flow rates. The most common example of this requirement is a shared lane (through vehicles plus left turns) in two opposing directions. If one of the directions has an exclusive left turn lane and the other has a shared lane, then the computations will be independent as long as the left turns from the exclusive left turn lane are not considered to be in conflict with the left turns from the shared lane in the opposite direction.

## SAMPLE CALCULATIONS

The numerical analysis techniques described in this paper have been programmed into a number of specialized computational modules, each of which can determine the traffic volume which will produce a specified delay on one approach or a pair of conflicting approaches. A comprehensive software package, which will solve the general case, does not exist at this time. Such a package would be substantially more difficult to develop than the existing packages which determine delay and level of service. It would also require that some liberties be taken with the HCM methods to produce approximate solutions where strict applications of the methods do not produce unique answers.

The sample calculations shown at the end of Chapter 9 of the HCM were mentioned earlier. These calculations illustrate the application of the techniques described in the body of the chapter. Selected examples have been chosen to demonstrate the "reverse" solutions described in this paper. A maximum error of 1 percent between the calculated delay and the target delay has been used in the analysis.

### The Simple Case

Let's begin with the simple case, defined previously as a single lane group with Type 3 arrivals and no permitted left turn movements. The northbound approach for sample calculation 2 is a perfect choice for this purpose. It is a two-lane approach on a one way street with Type 3 arrivals.

The results of the computations are shown in Figure 6. The relationship between delay and service volume is illustrated, and the service volume for each level of service is indicated. Note that there is no service volume for LOS A because the delay, even at  $X = 0$ , is beyond the 5-second upper limit for LOS A. Similarly, there is no service volume for LOS D because the capacity of the approach is reached prior to the threshold for LOS D.

## Permitted Left Turns

Sample calculation 1 presented in the HCM provides an excellent example of permitted left turns with shared lanes. The configuration is illustrated in Figure 7. The northbound and southbound movements take place from single lane approaches shared by through movements, left turns, and right turns. The eastbound and westbound movements both have two-lane approaches with the left lane shared by the through movements and left turns.

Figure 7 shows the computed volumes on the four approaches which would produce the delays which are shown in the HCM as the solutions to the problem. Note that these values agree within the specified 1 percent of the volumes originally given in the sample calculation. The error could, of course, be reduced by increasing the number of iterations. This demonstrates that the "reverse" solution to the equation was satisfactory. Figure 7 also shows the computed service volumes for each level of service.

Note that, for the same reason as the previous computation shown in Figure 6, there are no service volumes for LOS A or LOS E. There was also no solution to the equation for LOS B on the eastbound approach. This was not because a service volume didn't exist, but because of the discontinuity in the delay equation caused by the progression factor. This problem could have been solved either by trial and error or by interpolation of the progression factor.

The service volumes shown on Figure 7 are valid only when the same level of service applies to all approaches of the intersection. The method described in this paper could be extended to create a matrix of solutions with different levels of service on each approach.

## CONCLUSIONS AND RECOMMENDATIONS

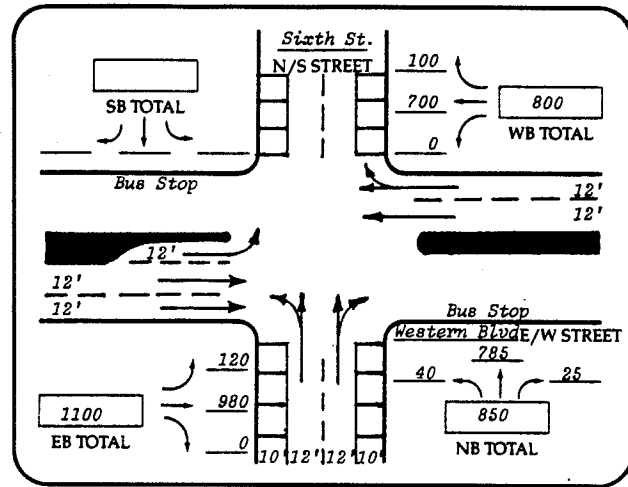
The study described in this paper has demonstrated that it is feasible to solve the 1985 HCM delay equation in reverse, using numerical analysis methods, to determine the traffic volume which will produce a specified delay for each approach to a signalized intersection. In some cases, it is necessary to solve two simultaneous equations. A worksheet is presented for dealing manually with simple cases, but computer methods should be considered for all service volume computations.

The mathematical solution to the equation is not difficult, but major problems are presented by interaction between the traffic volumes and the other inputs to the delay equation. These interactions produce discontinuities which require some trial and error to arrive at a solution. Furthermore, a unique solution which can be defended as the result of a strict application of the HCM method does not always exist. Finally, the areas requiring judgment in the determination of level of service (e.g., allocation of left turn volume between a permitted and protected phase) destroy the reversibility of the computational process. These complications all support the conclusion that the development of software to solve the general case would be a formidable task which would, as a minimum, require some official clarifications on the HCM procedures. Ideally, some minor modifications to the procedures would improve the reversibility of the computations.

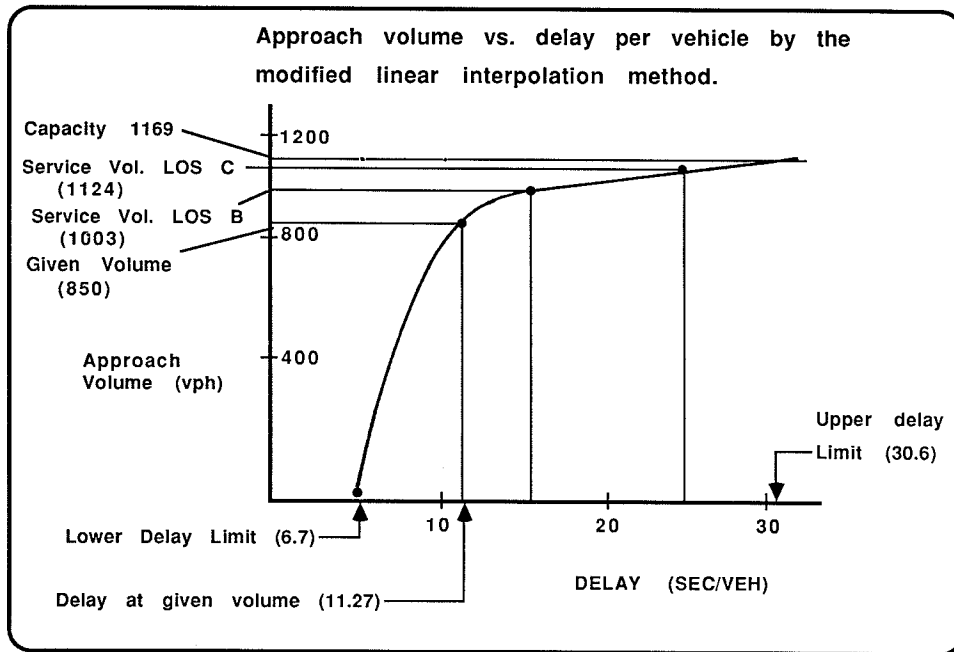
Probably the best example of a clarification or modification

**Northbound Movement**

- 50 second cycle
- 20.4 seconds green
- PHF = .95
- Arrival Type 3
- Volume = 850 vph.
- Vol. adj. factor 1.1
- Capacity 1169 vph
- Delay 11.27 sec/veh



**Intersection Configuration and Operating Parameters**



**FIGURE 6** Computational example for through volumes and unopposed left turns (HCM sample calculation 2).

is found in the progression adjustment factor. The discontinuity in the present tabular function makes the computational process hard to reverse. There is no reason to believe that the effect of progression on delay is discontinuous. A stronger emphasis on interpolation in the manual should provide a mandate for modification of the HCS to include an interpolation feature.

The subject of the progression adjustment factor is currently under review in connection with an NCHRP research project. The research team should be encouraged to consider continuity and reversibility of the model in developing recommendations for revisions to this section of the manual.

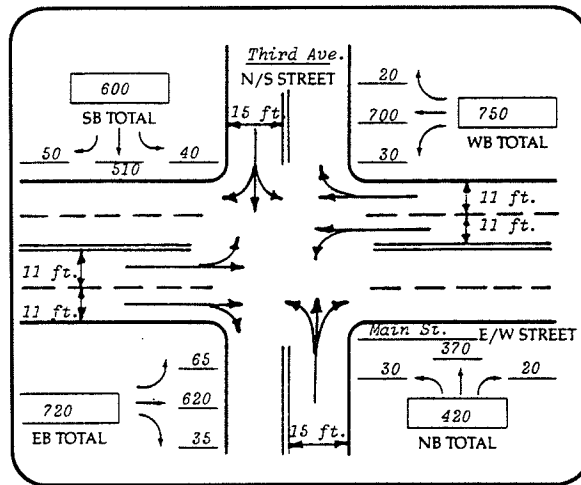
Some consideration should also be given to the develop-

ment of stronger guidelines for apportionment of left turning volumes between the protected and permitted phases. The judgment required at this point makes the delay computations irreversible.

The HCM signalized intersection analysis method has provided a powerful technique for determining how well an intersection will operate given the traffic volumes, intersection configuration, and signal operation. It has been less successful in producing a practical method of determining the service volume for a specified level of service. The computational techniques presented in this paper may be used to solve specific cases but, at this point, a comprehensive software package for computing service volumes does not exist.



2 phase operation  
 50 second cycle  
 PHF = .90  
 Arrival type: EB 4  
                   WB 2  
                   NB 3  
                   SB 3



Intersection Configuration and Operating Parameters

Comparison of computed volume vs. given volume				
	Target Delay	Given Volume	Computed Volume	Percent Error
Eastbound	27.84	720	716	.55
Westbound	25.58	750	748	.27
Northbound	10.96	420	420	0
Southbound	18.59	600	600	0

COMPUTED SERVICE VOLUMES for LOS A-D				
	LOS A	LOS B	LOS C	LOS D
Eastbound	---	No Solution	698	715
Westbound	---	202	751	821
Northbound	---	527	581	612
Southbound	---	527	581	612

FIGURE 7 Computational example for permitted left turns (HCM sample calculation 1).

## DISCUSSION

EDMOND C.-P. CHANG

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

The signalized intersection analysis method in the 1985 Highway Capacity Manual has provided very powerful techniques for evaluating the intersection operations under given conditions. Essentially, computing the predicted signal delay-based evaluation model in the 1985 HCM involves the calculations of four input variables, *i.e.*, the cycle length, green time, signal capacity, and volume-to-signal capacity ratio. The delay equation was formulated from theoretical considerations as well as from results of empirical experiments. However, since the 1985 HCM does not explicitly allow for system evaluations

for planning analyses, it has not been successful in providing easy methods to determine the minimum required service volumes for a specific level-of-service in the transportation planning process. This paper explores the possibilities of applying different numerical techniques for calculating the needed traffic volume levels from the 1985 HCM delay equation under a given level-of-service. The study was performed by solving the delay equation in reverse order from the sample problems in the 1985 HCM. The techniques being discussed in this paper include the analytical solution, exhaustive search solution, trial-and-error solution, nomograph approximation solution, and iterative numerical solution. Both manual methods, worksheet calculations, and computerized solutions were performed and presented. In addition, specific study procedures were applied for determining the traffic volumes to

generate a certain delay value as defined by delay-level-of-service criteria.

The study indicated that it is very feasible to estimate the proper traffic volumes for producing the specified delay and providing the precise estimation of the level of service of an approach to a signalized intersection. The results from this investigation can later serve as the "Threshold Values" or the "Design Criteria" for use in the planning analysis through quick response-type hand calculations. However, the study found that the major difficulties in determining the proper service volumes are not due to the mathematical characteristics of different components in the HCM delay equation, but rather result from the interdependency among the study variables and discontinuities in the arterial volume-density relationships. It should be noted that the four basic elements in the 1985 HCM analysis are not totally independent, rather, they are influenced by their corresponding values. Besides the interactions among each of the independent variables in the delay estimation equation, the traffic volume on a particular lane group may also be affected by the interactions within, between two lane groups, between two opposing approaches, and among all the approaches to the intersection. These potential interactions may result from the arterial progression, shared-lane operations, de facto left turn lane, permitted signal operations, and the random traffic volume variations. However, some of the unsatisfactory research from this investigation may have resulted from three operational problems:

1. Since there are heavy interdependencies among the major study variables, the interactions are causing barriers to solving the precise computation and determining the minimum traffic volume requirements for the given intersection level of service values.
2. Discontinuities do exist within the existing numerical values for estimating the progression adjustment factor, permitted left turn adjustment factor, and the level of service definitions.
3. There are needs for improving the interpretation of the current 1985 HCM and providing better interpolation of calculated values in the numerical calculations for implementing different HCS software.

Therefore, other possible approaches can be considered by limiting the solution ranges or providing a different representation to improve the research techniques under a given intersection level of service. The "Localized Solution" may be obtained by limiting the ranges of the study variables needing to be solved. The "Variable Specific Solution" may be made by confining the solutions of the traffic volumes in certain operating ranges to obtain the feasible numerical solutions from simultaneous equations. In addition, the improvements on the values of the existing adjustment factors in the 1985 HCM, such as the progression adjustment factors, left turn adjustment factors, and saturation flow rates, into the continuous equation can resolve most of the operational problems. The continuous relationships can improve the implementation of highway capacity methodologies for specific planning objectives and provide the needed estimation for the particular levels of service analysis.

In addition to the numerical approaches being investigated, the Artificial Intelligence techniques (AI/ES), being intro-

duced successfully in many disciplinary areas, may also offer an alternative to solve this problem. AI languages have the advantages to allow the heuristic human knowledge to be considered in the problem-solving process. Representing the knowledge in the AI programs means choosing a set of representative conventions and suitable structures for describing the objects, relations, and analysis processes. The design can begin by choosing a conceptual framework to represent the problem, either symbolically or numerically. Then, the conventions can be chosen through the given languages for implementing the design. Often, defining the conceptual analysis framework is more difficult than implementing the system with specific programming languages. Nevertheless, the proper definitions and translations of the system operational characteristics into the production rules offer another opportunity for allowing the users to reexamine the analysis from different search directions.

In particular, an alternative "Semantic Network" or "Frame-Based" representation technique, as proposed in 1987 by Hendrickson *et al.*, can also be used to represent the capacity performance analysis through the applications of semantic network representation techniques as developed in the Artificial Intelligence research area. Essentially, the fundamental relationships, as defined in the 1985 HCM delay equation, can be summarized into four major components, *i.e.*, service flow rates, geometric design characteristics, signalization conditions, and resultant intersection level of service analysis. As indicated in Figure 8, the analytical interactions in the 1985 HCM delay model can be represented through a series of elements in the Semantic Network. The elements include the "Nodes (Variables)," "Links (Relationships)," and "Frames (Combinations of Nodes and Links)." All the interactive relationships can then be quantified into one of the five basic relations as shown in Figure 9. For example, the calculation of the saturation flow rate can be considered as one of the four basic "Frames" which are constructed from the relations of each adjustment factor in the HCM. Once the basic relationships have been defined within the submodel, each submodel will inherit those characteristics and transfer the needed relations upward without having to requantify the interactions individually.

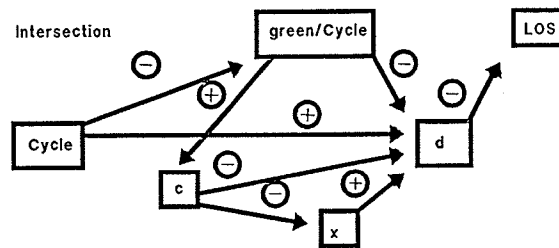
Figure 8, as described, illustrates two semantic network examples of this specific problem. Figure 8(a) presents the relationships among the different variables and corresponding semantic networks being used to represent the lane group analysis in the 1985 HCM delay equation. In this figure, the boundaries are used to indicate the different variables at different aggregation levels or the relative interactions among those variables. Figure 8(b) illustrates the basic relations and the design elements in the capacity equation. As indicated, four basic relations are used to represent the relative influences and functional cause and effects among each of the elements in the model. For example, the Signal Capacity ( $C$ ) and the Saturation Flow ( $S$ ) are both at the lane group level, while the fraction of the green time ( $g/C$ ) is at the alternative phase of aggregation. On the other hand, the determination of the Saturation Flow rate ( $s$ ) involves the interactions among the different adjustment factors. Again, in turn, the Saturation Flow ( $s$ ) will become the element that may affect the calculation of the estimated signalized intersection delay. The Saturation Flow ( $s$ ) therefore becomes one of the factors that will change the estimation of the volume levels under the given

EQUATION

$$d = 0.38 \frac{\text{Cycle} (1 - \text{green}/\text{Cycle})^2}{(1 - (\text{green}/\text{cycle}) X)}$$

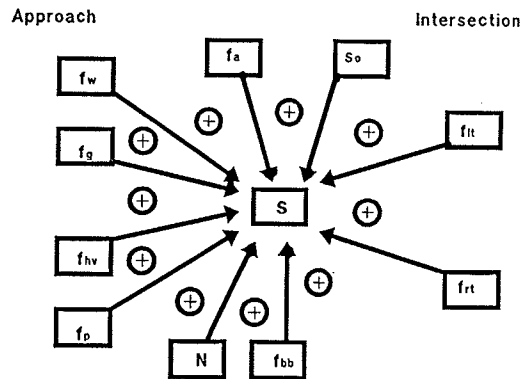
$$+ 17.3 X^2 [(x-1) + \sqrt{(x-1)^2 + (16 X/c)}]$$

SEMANTIC NET



(a) Elements in the 1985 HCM Delay Equation.

$$s = s_o N f_w f_g f_{bb} f_p f_a f_{rt} f_{it} f_{hv}$$



(b) Components in the Calculation of the Saturation Flow Rate.

FIGURE 8 Functional relationships in the 1985 HCM lane group analysis through the semantic network representation scheme.

NODE TO NODE RELATION	TYPE OF RELATION
	V <sub>1</sub> INCREASE, V <sub>2</sub> INCREASE V <sub>2</sub> CALCULATE AFTER V <sub>1</sub>
	V <sub>1</sub> INCREASE, V <sub>2</sub> DECREASE V <sub>2</sub> CALCULATE AFTER V <sub>1</sub>
	V <sub>2</sub> = SUM OF V <sub>i</sub> V <sub>2</sub> HIGHER THAN V <sub>1</sub>
	V <sub>1</sub> IS GREATER THAN OR EQUAL TO 100.

FIGURE 9 Basic relations used in the semantic network representation.

level of service. On the other hand, it has also inherited the basic characteristics of the individual factors without having to define them again in this semantic network scheme.

By representing this particular capacity problem and relating the interrelated variables symbolically, the semantic network representation scheme can provide a more flexible definition to the whole analysis problem. Regardless of the terminologies and variables being used, the network representation scheme can be used to represent the relationships and delay equations analytically. The representation of this analysis procedure based on the "Semantic Networks" and "Frames" concept can offer better illustrations to the relative impacts among different variables. It will also allow the easier implementation of computerized search, facilitate the analysis interpretation, and provide a comprehensive check of the solution scheme. These representation techniques may later be adapted to the Expert Systems solution framework for allowing object-oriented problem solving. By localizing the analysis knowledge in each frame or submodel, the network

representation can further support the parallel processing of the different approaches needed for various highway capacity analyses.

### AUTHORS' CLOSURE

The reviewer's comments reinforce the original conclusions of this paper. The 1985 *Highway Capacity Manual* analyzed intersection methodology is not amenable to "reverse" solutions. Substantial additional research would be required to develop an explicit technique for determining service volumes, and further clarification would be required on the interpretation of the current Highway Capacity Manual procedure. The reviewer has suggested two possible approaches. Undoubtedly there are several others.

### REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985, 516 pp.
2. *Special Report 87: Highway Capacity Manual*. HRB, National Research Council, Washington, D.C., 1965, 411 pp.
3. Florida Department of Community Affairs, Rule 95-2.0255, Page 5.
4. J. E. Leisch. Capacity Analysis Techniques for Design of Signalized Intersections. *Public Roads*, Vol. 34, Nos. 9 and 10, 1967.
5. J. E. Leisch. *Capacity Analysis Techniques for Design of Signalized Intersections*. Institute of Transportation Engineers, 1986.
6. J. M. McCormick and M. G. Salvadori. *Numerical Methods in FORTRAN*. Prentice-Hall, Englewood, 1964.
7. C. F. Gerald. *Applied Numerical Analysis*, 2nd ed. Addison-Wesley, Reading, Mass., 1978.
8. J. A. Bonneson and P. T. McCoy. Operational Analysis of Exclusive Left-Turn Lanes with Protected/Permitted Phasing. In *Transportation Research Record 1114*, TRB, National Research Council, Washington, D.C., 1987, pp.74-86.

---

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

# Some New Data That Challenge Some Old Ideas About Speed-Flow Relationships

B. N. PERSAUD AND V. F. HURDLE

Several issues related to the upper branch of the speed-flow curve are addressed using data gathered at a freeway bottleneck in Toronto. Some important findings evolve, some of which challenge conventional beliefs. At low to moderate flows, speed on the upper branch is found to be insensitive to flow, a notion that is becoming increasingly accepted; at higher flows, the data suggest that speed decreases with increasing flow, but the fall-off is not nearly so precipitous as is commonly thought. The data further suggest that the presence or absence of an upstream queue is a more important variable than flow for predicting freeway speeds. Perhaps the most important finding is that belief in a precipitous speed drop may very well have resulted from a misinterpretation of data that arises because the speed of vehicles discharged from a queue varies with location in the bottleneck.

In this paper, questions relating to the upper, high-speed branch of speed-flow diagrams will be examined in some depth through the use of data collected on a freeway in Toronto, Canada. To an informed reader, it might be puzzling that such fundamental questions should still be of interest after so many years of research, but the authors believe they are not only deserving of study, but qualify as truly neglected areas. Furthermore, that neglect may very well lead to unjustified decisions regarding freeway construction or control, hence have serious implications for transport policy.

Most of the literature seems to take it for granted that there is a functional relationship between flow and speed and that this relationship can be represented by a curve with a well known and clearly defined shape. This paper also assumes that the relationship exists and can provide useful information for at least some purposes, but raises questions about whether the curve's shape is what most books would lead one to believe. The current state of the art in North America, as reflected in the speed-flow curves for freeways in the Transportation Research Board's 1985 *Highway Capacity Manual* (1), has evolved from numerous empirical studies, so one might suppose that little is to be gained by yet more empirical studies. However, the fact that these curves have gone through an evolutionary process and are quite different from previously used curves—those in earlier editions of the *Highway Capacity Manual*, for example—does not necessarily imply that they are correct, but should, instead, be taken as justification for further research that explores whether or not this process of evolution is complete. Further, in this research area, what constitutes current lore is often based on consensus of empirical research findings, so any contribution toward the for-

mation of a consensus can be seen as worthwhile. Thus, even if the result of this type of research is merely a confirmation of current beliefs, a contribution will have been made in that one can have increased confidence in current beliefs. In the case of this paper, the results will tend to confirm some current beliefs, but give cause for doubting that others are correct.

Once one makes the basic assumption that it makes sense to talk about speed-flow curves at all, the questions about the upper branch that appear to be unresolved fall into two areas and can be illustrated using the curves from the *Highway Capacity Manual* (1) shown in figure 1. The first question concerns the part of the curve at low to moderate flows. Recent literature—Roess, McShane, and Pignataro (2), Hurdle and Datta (3), and Allen, Hall, and Gunter (4), for example—suggests that the upper branch is quite flat at these flows, at least for North American conditions where speed limits are set and enforced in such a way that rather few vehicles travel a great deal faster than the average speed. Indeed, the new *Highway Capacity Manual* states that "There is a substantial range of flow over which speed is relatively insensitive to flow; this range extends to fairly high flow rates." (1, page 3-5). However, despite this trend in current thinking, the bulk of the available literature suggests that speed drops even at quite low flows, so it seems likely that there are still some nonbelievers who need to be further convinced.

Furthermore, there is a real question as to just how insensitive to flow speeds are within the range of zero to perhaps 1,500 vehicles per hour per lane. Figure 1 shows a drop of 6 to 8 mph over this range, but the authors can see no evidence of such a drop in the data presented in the three works mentioned above (2, 3, 4). A comparison of the data points in

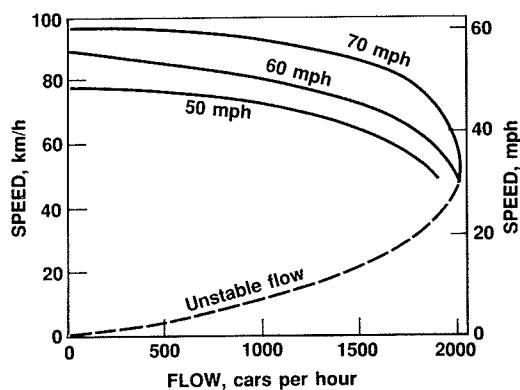


FIGURE 1 Speed-flow curves from the *Highway Capacity Manual* (1).

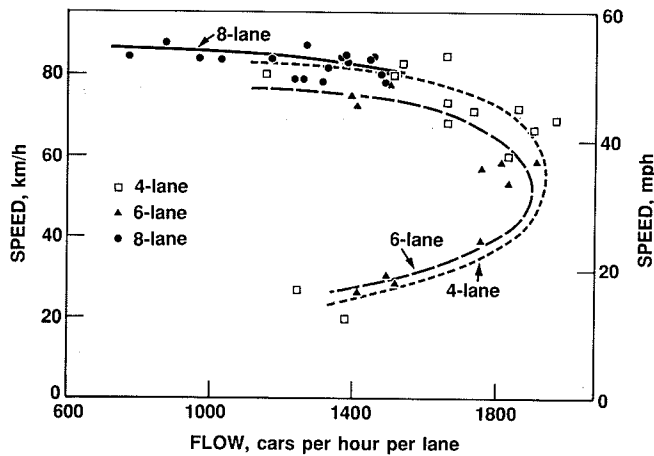


FIGURE 2 Speed-flow data and curves proposed by Roess et al. (2).

figure 2, taken from Roess, McShane, and Pignataro (2), and figure 3, taken from Hurdle and Datta (3) is particularly interesting in this regard. This is because the two figures also appear in the *Highway Capacity Manual* (1) and because there was a 55 mph speed limit on the New York area parkways where the figure 2 data were gathered, but a limit of 100 km/h (62 mph) at the Canadian location of the figure 3 study. This difference is clearly reflected in the two figures, but the authors can see no evidence in either of them that the average speed dropped as the flow increased from zero to 1500 vehicles per hour per lane. Certainly, it did not drop anything like the 6 to 8 mph indicated in figure 1.

The second major question area concerns the high-flow portion of the upper branch. The question here has three parts: At high flows, is speed no longer insensitive to flow? If it is not, where is the "break" point in flow, and what is the shape of the diagram at higher flows? Figure 1 suggests that the answer to this set of questions is that there is no real break, but that the slope of the curve changes very gradually at first, then increasingly rapidly. In the case of the 70 mph design speed, this slope change continues until the curve becomes vertical; in the words of a *Highway Capacity Manual*

summary prepared for personnel of the Federal Highway Administration (5), "... speed precipitously declines as flow approaches capacity." While this answer appears to reflect common belief, some doubt lingers. Hurdle and Datta (3), for example, suggest that speeds on the entire upper branch are not a function of flows at all, but of whether or not vehicles have been in an upstream queue. The results of the current study do not support quite such an extreme view, but neither do they support the idea that there is a precipitous drop in speed as the roadway's capacity is approached. Furthermore, they indicate that the presence or absence of an upstream queue is a more important variable than flow if one wants to predict freeway speeds and that when there is a queue upstream, the speed is primarily a function of the distance from the observation point to the head of the queue.

Before presenting those results, however, it is instructive to review some of the existing literature in hopes of discovering the foundations of current thinking about the upper branch. This examination will provide a basis for judging current thinking and a backdrop against which results from this study can be presented. In reviewing previous empirical work, two bundles of issues—categorized according to whether they are conceptual or analytical—appear to be of primary interest. Along the way, some possible pitfalls will be discovered; it seems quite possible that some current beliefs about the upper branch may have come into being because of improper handling of these issues. In the current study, great care was taken to avoid these pitfalls, but the extent to which earlier studies avoided them is not clear from the published literature.

### CONCEPTUAL AND ANALYTICAL ISSUES

The first issue to be examined arises from the authors' belief that the way one draws the upper branch is dictated by how one conceptualizes the entire speed-flow diagram. Probably the most common concept is that the upper branch represents free-flow conditions, the lower branch represents unsteady operation characteristic of conditions in a queue, and capacity occurs where the two branches meet. The authors have no quarrel with this concept, but believe that a problem arises

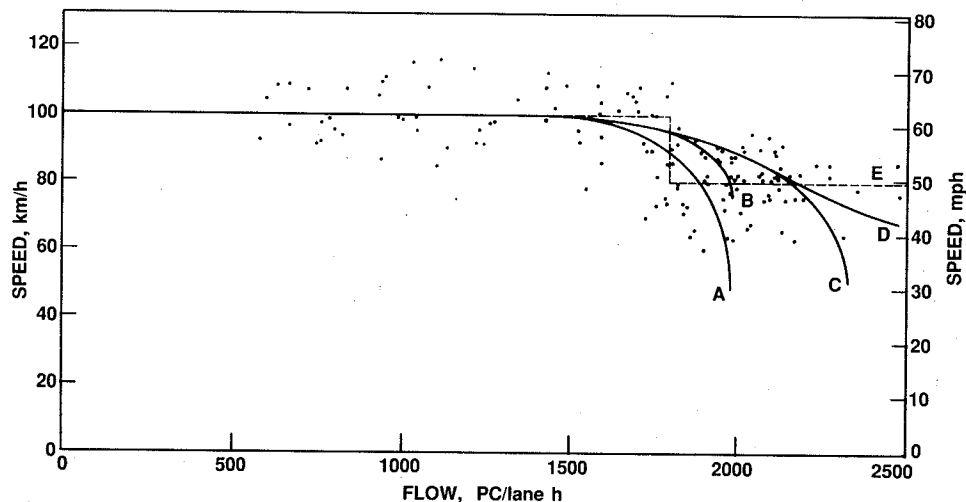


FIGURE 3 Speed-flow data and some possible curve shapes discussed by Hurdle and Datta (3).

because there has been little convincing evidence to indicate how the two branches meet, or if they meet at all.

It is very common to assume that the two branches form a single, smooth, continuously differentiable curve, but we can see no logical reason why this should be true necessarily. It is easy to see, however, that the shape of the upper branch one draws is very dependent on whether one assumes that the entire speed-flow curve is continuous, continuously differentiable, or neither. In particular, if one assumes continuous differentiability, then the curve must be vertical at the right end, as is the 70-mph curve in figure 1. Such curves are very common, but the authors suspect that most of them become vertical, not because of anything in the data on which they are supposedly based, but because some researcher had an a priori notion that the speed-flow curve—or, more likely, the speed-density curve—must be continuously differentiable. Probably the best-known empirical study addressing these issues is that of Drake, Schofer, and May (6), who fitted several of the well-known hypotheses—Greenburg, Underwood, Edie, Greenshields, and so on, to find out which one best fit their data. Some of these hypotheses imply a continuous speed-flow curve, some a continuously differentiable one, while some suggest a discontinuous curve of two or three regimes. After applying sophisticated statistical tests, the authors concluded that “the various hypotheses endured these tests with little differentiation.” Unfortunately, this conclusion is probably primarily a result of the fact that statistical testing is a very blunt tool for the purpose. As discussed later in this section, Duncan (7), who relied more on subjective, visual methods and less on statistical procedures, instead concluded that some possible curves—in particular, those that are continuously differentiable—were not compatible with his data.

This paper’s description of statistical testing as a blunt tool can be appreciated if one considers the problem of deciding which of the five speed flow curves in figure 3 best fits the data. There are two problems with using statistical testing procedures to solve this problem. The first is that the vertical scatter of the data points is so great that one would need a very large data set to show that the fit of one curve was significantly better than another, even when the curves differ as radically as those in figure 3. This difficulty is aggravated by the fact that data for high flow, uncongested conditions are difficult to obtain because such conditions ordinarily last for such a short time that only a few data points can be obtained from each day of observation. The second difficulty is one of definition: how does one statistically compare curves A and B, which become vertical at about 2,000 passenger car units per hour per lane, with curves D and E, which extend to the highest flows observed, but not to low speeds?

A third, closely related analytical issue has to do with curve fitting. If one wants to fit a curve to data by statistical methods, one must first specify an algebraic form for the curve, a parabola, for example, or some sort of logarithmic equation. The choice of algebraic form, however, can be more important in determining the shape of the curve than the data. Suppose, for example, that two researchers try to fit a curve to data similar to that in figure 2 by least squares procedures. Both decide to fit a curve of form

$$y = a + b_1x^2 + b_2x^2 + \dots,$$

but researcher A treats speed as the dependent variable ( $y$ ) and excludes the observations at less than 30 mph as obviously

not upper branch data, while researcher B treats flow as the dependent variable ( $y$ ) and uses all of the data. Neither procedure can be criticized as obviously unreasonable, but for any data even vaguely resembling that in figure 2, researcher B will obtain a curve that becomes vertical at the right end: a result that researcher A cannot possibly obtain.

In presenting their data, the authors shall try to avoid this problem by not doing any numeric curve fitting, yet discussing the data as though a curve were being fitted. In essence, each reader is asked to think of fitting a curve to the data either by eye or by some marvelous, as yet uninvented, analytical method that can yield a curve which truly fits the data without the researcher having to specify a form for the equation. In doing so, any ability to produce numeric results in favor of a greater freedom to discuss the shapes of curves that exist only in the readers’ minds is sacrificed. Naturally, the authors hope these curves will resemble those in their minds, but they have deliberately refrained from drawing any curves, preferring to rely on gentle persuasion rather than visual suggestion.

A second decision the authors made is not to show data points obtained when the study section was congested (*i.e.*, lower branch data). In part, this simply reflects the fact that their study section—described later in this paper—is a bottleneck, so normally causes congestion upstream rather than becoming congested itself. However, it does occasionally become congested, so there is a very limited amount of lower branch data. That it is not shown reflects the authors’ doubts that the two branches of the speed-flow relationship form a single, smooth curve, but readers are asked to keep an open mind on this issue. This paper will also, initially, omit observations made while a queue existed upstream from the study section. That these observations do not constitute legitimate upper branch data is one of the paper’s main points, but readers will eventually be shown the data and asked to judge for themselves.

The final issue in this bundle relates to the question of whether one examines the speed-flow relationship directly or indirectly. In tracing the evolution of speed-flow diagrams, one has to suspect that popular ideas about their shape arose from the work of investigators who first explored the speed density relationship, then inferred a speed-flow relationship from the expression flow = speed  $\times$  density. Such an approach may well have seemed appropriate because speed and density were “natural” variables in car-following theory and because speed-density data has considerably less scatter than speed-flow data. Duncan (7), however, struck a telling blow against this philosophy in a landmark paper that, unfortunately, seems not to be as well known as it deserves to be. He first fitted two plausible relationships to some speed-density data, one continuous and one with a discontinuity, then calculated and plotted the corresponding speed-flow relationship for each. Next, he transformed the data to speed-flow form and, using some sound intuitive arguments, fitted new curves to this transformed data. Again, what was produced was a variety of shapes for the upper branch, but with the interesting feature that the shape of the curves based on the transformed data was radically different at the high flow end than the shape inferred from the fitted speed-density curves.

This issue can be further illustrated by examining it in the context of a data set and some curves presented by Leutzbach (8). Figure 4, prepared from figures in Leutzbach’s paper, shows 1-minute speed-density data gathered at four locations

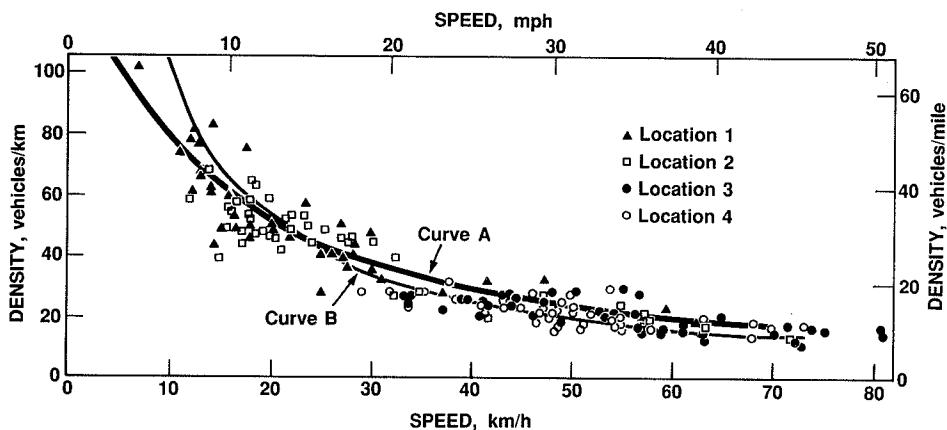


FIGURE 4 Speed-density curves prepared from figures in Leutzbach (8).

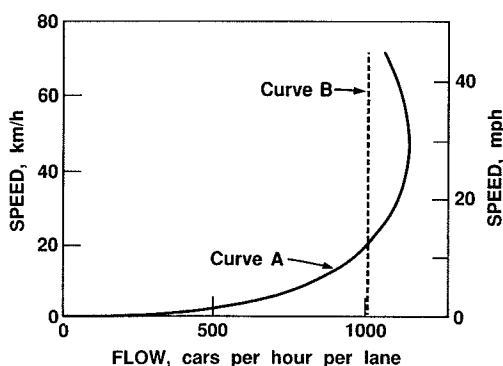


FIGURE 5 Speed-flow relationships implied by the curves in Figure 4.

300 meters apart at the beginning of a one-lane bottleneck while there was a queue at the entrance. Leutzbach fit curve A by eye, and it appears to fit the data quite well; it translates into the speed flow relationship indicated by curve A of figure 5. Curve B of figure 4, also taken from a diagram in Leutzbach's paper, is a fitted speed-density relationship of the form  $\text{speed} = \text{constant}/\text{density}$ , implying, as one might expect of vehicles being discharged from a queue, that flow is reasonably constant in the bottleneck: the speed-flow relationship is merely a vertical line in figure 5. Thus, even though there is very little difference between the two speed-density curves in the region where data was available, the implied speed-flow relationships are radically different. From this illustration, there is once again a clear message: one must be cautious when transforming relationships from speed-density form to speed-flow form or vice versa. Specifically, one must be very cautious about using this type of transformation to form ideas about the shape of the speed-flow diagram.

## EMPIRICAL RESULTS

The study (9) on which this paper is based was carried out in the vicinity of a bottleneck on the Gardiner Expressway in Toronto, Canada. As shown in figure 6, the Spadina Avenue entrance ramp joins the freeway and forms a fourth lane which is dropped within a 700 m radius curve after about 1.2 km to

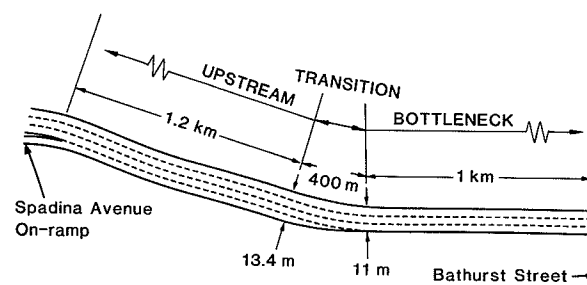


FIGURE 6 Plan view of study area (not to scale).

form a three-lane bottleneck. The speed limit is 90 km/h (56 mph) and the freeway has many restrictive design features, but speeds in excess of 100 km/h are common in the study area. Trucks are prohibited in the median lane, so flows expressed in vehicles per hour in this lane are, in effect, equivalent to flows in passenger car units per hour. Since this eliminates the complication of determining and applying passenger car equivalency conversions, most of the results presented in this paper will be for the median lane only.

Traffic leaving the downtown area during the afternoon rush period was observed by taking pictures with a 16 mm time-lapse camera mounted approximately 360 m above the ground on a tower located just off the left edge of the figure. The low to moderate flows in the opposite (inbound) direction were also captured on the film. For the outbound, high-flow direction, there were about 36 minutes of data over three days that showed free-flow conditions just before the queue formed upstream of the bottleneck and substantially more data after the queue had formed. These three days yielded about three hours of low to moderate inbound flow data as well. The outbound freeway segment was divided into sections as indicated in figure 6 and a 2-minute averaging interval was used for speed-flow measurements. With the data reduction method used (9, 10), average flows and densities were, in effect, obtained directly, while average speeds were computed from the expression  $\text{flow} = \text{speed} \times \text{density}$ . A discussion of the accuracy of the method of computation and a comparison of calculated speeds with values obtained by direct measurement of individual cars' travel times is included in the authors' reference (10).



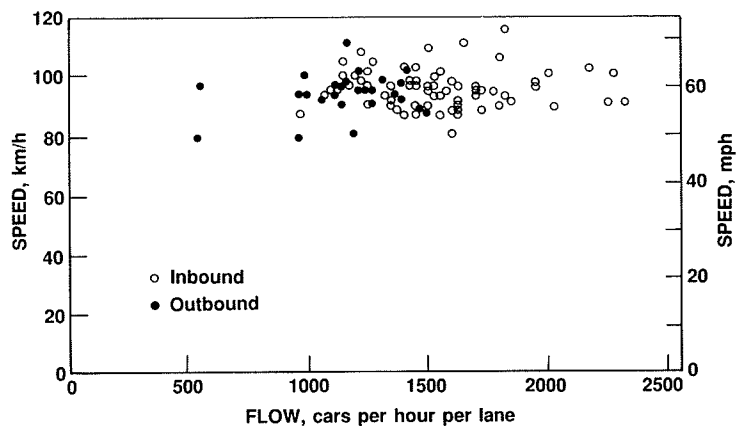


FIGURE 7 Speed-flow plot for inbound and outbound off-peak flows.

**SHAPE OF THE UPPER BRANCH AT LOW TO MODERATE FLOWS**

As indicated above, low to moderate flows prevailed in the inbound direction, so it was decided to examine the shape indicated by the inbound speed-flow data. The entire freeway bends in the area of the outbound lane drop and the inbound median lane segment for which data was extracted is just downstream of that curve. In figure 7, the speed-flow observations for this segment are indicated by open circles. To check whether the two directions have similar characteristics and to supplement the inbound data, some outbound median lane data were gathered during an off-peak period. These observations, which are denoted on figure 7 by the solid circles, indicate that speeds at moderate flows are about the same in both directions. The most striking aspect of figure 7, however, is that, for the flows of interest in this part of the paper—those up to about 1,800 vehicles per hour—speed in both directions fluctuates around an average of about 95 km/h at all flow levels, with no indication of a decrease in speed as the flow increases.

The final thing to note about figure 7 is that there are eleven inbound data points at flows larger than 1,800 vehicles per hour, all at speeds greater than 90 km/h, apparently indicating that the insensitivity of speed to flow applies to all flows. However, since these data points are few and scattered, they can be more properly discussed in the context of the relatively

large number of high-flow outbound observations. This is done in the next two sections of this paper.

**SHAPE OF THE UPPER BRANCH AT HIGH FLOWS**

This section will explore the shape of the high-flow portion of the upper branch of the speed-flow curve. To do so, it is useful to first look at the outbound data on those days when there were observations without a queue present. As indicated earlier, three days of filming captured short periods of time when the outbound flows in the median lane were in excess of 1,500 vehicles per hour, and there was no noticeable queue. The three days' speed flow observations in the median lane during this period are plotted as x's in figure 8 and the outbound off-peak data points introduced in figure 7 as solid circles.

On each day, a queue formed upstream, so one can be sure that the capacity of the lane was reached. The x's include all 2-minute observations made before this happened, but none made after it occurred. Thus, whether they include conditions at capacity is perhaps questionable, but they certainly include conditions approaching capacity. (Readers are cautioned not to infer from figure 8 that the capacity of this lane is more than 2,400 vehicles per hour. The flows shown are based on counts only 2 minutes long, so the amount of random fluctuation is considerable. The highest flows observed would

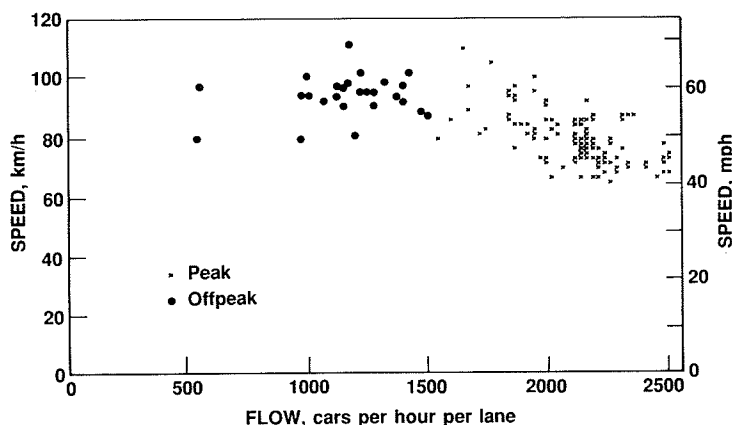


FIGURE 8 Speed-flow plot for the median lane, outbound.

undoubtedly be exceeded if one were to watch for a larger number of days, but they are never sustained for long periods of time. It would be unreasonable to say that just because they happened to occur, the capacity must be still higher.)

Visual inspection of figure 8 strongly suggests that, for flows exceeding 1,800 vehicles per hour, speeds do fall off gradually with increasing flows—quite a contrast to the pattern for lower flows. It is also apparent that the fall-off is not nearly so precipitous as the curves in figure 1 suggest. In fact, if a curve were to be fit by eye to the data for flows greater than 1800 vehicles per hour, a straight line would seem to appear. Even a curve with a small, but positive, second derivative seems compatible with the data, but the authors do not wish to suggest anything so radical. What they do want to suggest, however, is that a curve with a large negative second derivative such as the 70 mph curves in figure 1 seems incompatible with the data: if the curve had the shape shown in figure 1, at least some of the points at the right end of figure 8 would be expected to have lower speeds.

It would be tempting to try to estimate the shape of the speed drop and perhaps learn some more from the data by curve fitting or parameter estimation. As discussed earlier, however, the nature of the data provides a major stumbling block: the scatter in speeds is too large compared to the apparent change in the mean speed for these traditional techniques to be very useful without an extremely large data set. In addition, since data for many sections are combined, every vehicle is likely to be included in several observations; therefore, the data points are not all independent and statistical tests and the estimation of confidence limits on the parameters would be difficult, if not impossible, to carry out.

Figure 8 also suggests that the lowest speeds occurring at flows similar to, or in excess of, the flows normally quoted as capacity are of the order of 65–70 km/h (40–43 mph)—substantially higher than the “capacity” speed of 30 mph (48 km/h) indicated in figure 1. However, since figure 1 is based on average speed and flow per lane, the question arises: Do this and other findings based on median lane data apply to data averaged over all lanes? Because of the difficulty in accurately aiming the time-lapse camera, less data was available for the other lanes than for the median lane, but what is available (figure 9) clearly supports all of the conclusions so far about the shape of the speed flow curve. All that appears to be different is that average speeds and flows per lane are

somewhat lower than in the median lane. The highest flow is now about 2,075 vehicles per hour per lane at speeds, as before, of 65–75 km/h (40–47 mph).

The finding that the fall-off in speed might not be so sharp as is commonly believed has some important implications. The first is that level-of-service criteria may need to be revised. Page 3-5 of the *Highway Capacity Manual (1)* states on the basis of the 70 mph curve in figure 1 that, “As capacity is approached, small changes in volume or rate of flow will produce extremely large changes in operating conditions, *i.e.*, speed and density. Level-of-service criteria for freeways reflect this, with the poorer levels defined for reasonably large ranges in speed and density, while the corresponding range in flow rates is rather small.” If the fall-off in speed is neither as large nor as sudden as that shown in figure 1, then it is easy to see that predictions of the level of service to be expected at some given flow are likely to be overly pessimistic. The second implication is, in a sense, related to the first. Many believe that it is sound economic policy to prevent flows from reaching the levels at which the precipitous drop in speed occurs. Naturally, if the drop is not precipitous, such a policy would require rethinking.

Before concluding this section, it is worthwhile to give some special consideration to the interesting nature of the inbound flows in figures 7 and 9. Figure 10, which is a merger of the inbound and outbound observations from figures 7 and 8, clearly shows that, while for flows less than 1,800 vehicles per hour the two directions are visually indistinguishable, at larger flows there is little overlap. It is, therefore, tempting to suggest that the two directions behave differently at high flows, but with only eleven high-flow inbound data points, such a suggestion would require further support. However, one could speculate that under stable operating conditions a bunch of “brave” drivers can produce a high flow at high speeds in the median lane, but—as evidenced by the absence of high-flow, inbound data in figure 9—it is unlikely that such drivers would be found in all lanes during the same time interval.

#### POSSIBLE EXPLANATION FOR THE BELIEF IN A PRECIPITOUS SPEED DROP

Because the finding that there might not be a precipitous speed drop is so contrary to conventional wisdom and because

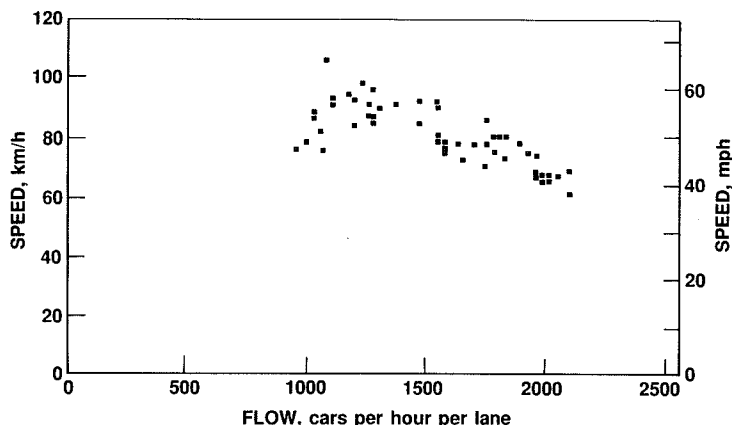


FIGURE 9 Speed-flow plot (3-lane average)—upper branch.

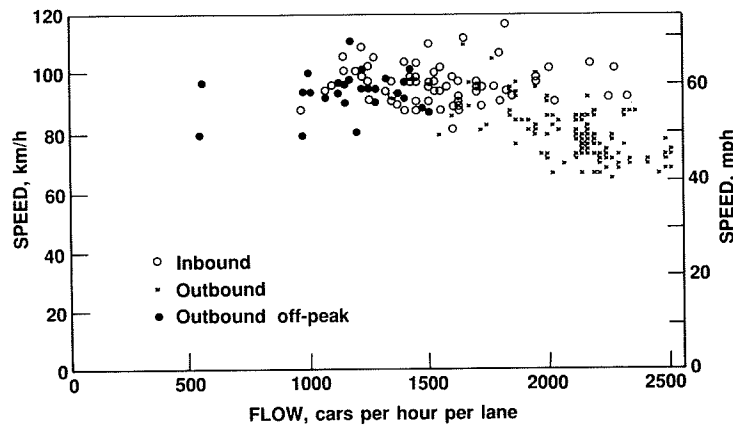


FIGURE 10 Speed-flow plot showing data from Figures 7 and 8 combined.

of the implications, it is natural to question whether such a finding from one empirical study can have wide applicability. If so, this would imply that studies on which current beliefs are based have erred in their conclusions. This is a difficult issue to address since it is not clear to what extent previous studies have sought to avoid the pitfalls hinted at earlier. It is possible, however, to point to at least one way in which one can erroneously arrive at a conclusion that there is a precipitous drop in speed at high flows.

The data points in figure 11, which is plotted with a different speed scale than the previous figures, are one day's speed-flow data observations averaged over all three lanes of the bottleneck while a queue was present upstream. Different symbols have been used to identify speed-flow data obtained in 110-meters long sections of the bottleneck; section 1 is located just past the lane drop in figure 6 and section 8 begins 768 m (not quite half a mile) farther downstream. For each section, the points appear to form a small cluster indicating reasonably steady speeds and flows. Average speed increases while average flow remains constant as one progresses into the bottleneck. The reason for this is that the vehicles upstream from the bottleneck are waiting in queue at either very low speeds or in the familiar stop and go fashion, so cannot be moving very fast in section 1 at the very upstream end of the bottleneck: simply because instantaneous speed change is not physically possible. Within the bottleneck, however, the vehicles do accelerate, so the speed gradually increases as one moves downstream. The only surprising thing about this is that the acceleration is so small and continues over such a long distance.

While the speed change data is of considerable interest in itself, the main reason for presenting it here is to point out that if one did not recognize the data points as "queue discharge" observations, they might easily be construed as supporting belief in a precipitous drop in speeds at high flows. It is easy to see this by combining the legitimate upper branch data in figure 10 with the "false" upper branch data in figure 11. The resulting plot, figure 12, clearly indicates how one can be led astray. This situation is not at all far-fetched; it is quite possible for upper branch speed-flow observations to "accidentally" become polluted by data such as that in figure 11, since data is usually gathered in such a way that it is difficult to tell if and when there is a queue upstream. In fact, very few published studies even say anything about whether there was a queue upstream. If data from more than one location are mixed together, as the authors suspect is often the case, the likelihood of misinterpretation becomes even greater since the different clusters of points, if all plotted with the same symbol, are likely to look exactly like the data one would expect to see if speeds dropped precipitously as the flow approached capacity.

SUMMARY

In this paper, several issues related to the upper branch of the speed-flow curve have been addressed using data gathered in the vicinity of a freeway bottleneck. Some important findings have resulted. There was confirmation of current belief that at low to moderate flows speed is insensitive to flow,

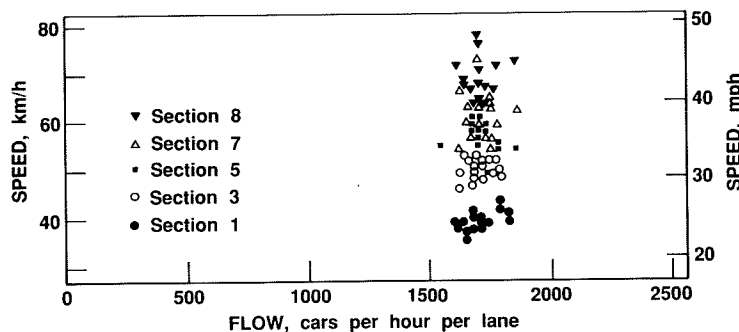


FIGURE 11 Speed-flow observations (3-lane average) at different locations in the bottleneck with a queue upstream.

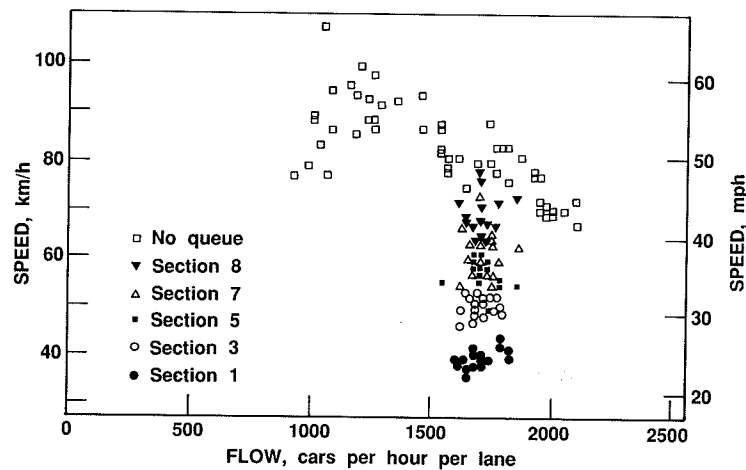


FIGURE 12 Merger of Figures 9 and 11 (at the scale of Figure 11) to illustrate “false” precipitous speed-drop.

though these results are in closer agreement with speed-flow curves in several of the references (2, 3, 4) than with the one in the *Highway Capacity Manual* (figure 1). At higher flows the study's data suggest that speed does decrease with increasing flow, but that the fall-off may not be nearly so precipitous as is commonly believed. Perhaps the most striking finding is that belief in a precipitous drop in speed at high flow may very well have resulted from misinterpretations of data that arose because the speed of vehicles discharged from a queue varies with location in the bottleneck, but the flow does not.

#### ACKNOWLEDGMENT

This research was supported by the National Science and Engineering Research Council of Canada.

#### REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985, 516 pp.
2. R. P. Roess, W. R. McShane, and L. J. Pignataro. Freeway Level of Service: A Revised Approach. In *Transportation Research Record 699*, TRB, National Research Council, Washington, D.C., 1979, pp. 7-16.
3. V. F. Hurdle and P. K. Datta. Speeds and Flows on an Urban Freeway: Some Measurements and a Hypothesis. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 127-137.
4. B. L. Allen, F. L. Hall, and M. A. Gunter. Another Look at Identifying Speed-flow Relationships on Freeways. In *Transportation Research Record 1005*, TRB, National Research Council, Washington, D.C., 1985, pp. 54-64.
5. *The 1985 Highway Capacity Manual—A Summary*. Office of Traffic Operations, Federal Highway Administration, U.S. Department of Transportation, 1986.
6. J. S. Drake, J. L. Schofer, and A. D. May. A Statistical Analysis of Speed Density Hypotheses. In *Highway Research Record 154*, HRB, National Research Council, Washington, D.C., 1967, pp. 53-87.
7. N. C. Duncan. A Further Look at Speed/Flow/Concentration. *Traffic Engineering and Control*, Vol. 20, 1979, pp. 482-483.
8. W. Leutzbach. Testing the Applicability of the Theory of Continuity on Traffic Flow at Bottlenecks. *Proc.*, 3rd International Symposium on the Theory of Traffic Flow, New York, 1965 (L. C. Edie, R. Herman, and R. Rothery, eds.), New York, 1967.
9. B. N. Persaud. *Study of a Freeway Bottleneck to Explore Some Unresolved Traffic Flow Issues*. Ph.D. dissertation. University of Toronto, Toronto, Canada, 1986.
10. B. N. Persaud and V. F. Hurdle. Freeway Data from Time-lapse Film—an Alternative Approach. *Transportation Science*, in press.

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

# Capacity Analysis of Two-Lane Highways

DAVID L. GUELL AND MARK R. VIRKLER

The general terrain procedure for two-lane highways in the Highway Capacity Manual contains two flaws which could be significantly improved by changing the definition for when a vehicle is delayed. The first problem is that, even under the best roadway and traffic conditions, any two-way flow rate greater than 43 percent of capacity falls in levels of service D or E. Furthermore, levels of service D and E are too broad to provide definitive information about the flows within these levels. By changing the definition for when a vehicle is being delayed from a headway of 5 seconds, as given by the Manual, to a headway of 3.5 to 4.0 seconds, more useful level-of-service categories result. The second problem is that the general terrain procedure does not yield results compatible with the specific grade procedure. Given otherwise identical traffic and roadway conditions, a two-lane highway will often be categorized as having a better level of service on a specific grade than on a level or rolling terrain segment. This is opposite of what one would expect. By reducing the delay definition to 3.5 or 4.0 seconds, this inconsistency is not completely eliminated but is greatly reduced. Cases could be made for defining the delayed headway as any value from 2.0 to 6.0 seconds. For the purposes of the Highway Capacity Manual, a value between 3.5 and 4.0 seconds would provide more reasonable and consistent results.

The 1985 edition of the Highway Capacity Manual (1) has introduced a major revision to the procedures for analyzing uninterrupted flow on two-lane highways. New criteria are given for establishing the levels of service, and separate criteria are used for general terrain segments and specific grades.

For general terrain segments, any two-way flow rate greater than 43 percent of capacity is in level of service D or E, even under the very best roadway and traffic conditions. By comparison, for similar best conditions on a freeway it takes 77 percent of one-way capacity to be in level of service D; on other multilane highways it takes 71 percent.

Because different procedures are used for defining the levels of service for general terrain segments and specific grades, a serious inconsistency in the level of service can arise between a general terrain segment and a specific grade. Under identical traffic conditions, the level of service on a specific grade can often be better than the level of service on flat terrain.

The purpose of this paper is to examine the level-of-service criteria for general terrain segments and to offer possible alternatives. These alternative criteria would increase the boundary between levels of service C and D from the current 43 percent of capacity to between 53 and 60 percent under the best conditions. It will also be shown that the proposed alternatives reduce the inconsistency in the level of service between a general terrain segment and a specific grade.

## GENERAL TERRAIN SEGMENTS

The Highway Capacity Manual (1) uses "percent time delay" as the primary measure of level of service for general terrain segments on two-lane highways. Average travel speed and capacity utilization are secondary measures. Percent time delay is defined as the average percent of time that all vehicles are delayed while traveling in platoons due to the inability to pass. Motorists are defined to be delayed when traveling behind a platoon leader at speeds less than their desired speed and at headways less than 5 seconds. As a surrogate for percent time delay, the percent of vehicles traveling at headways less than 5 seconds can be used. This surrogate is more easily measured in a field study than the percent of the time that all vehicles are delayed. The cut-off values for levels of service A, B, C, and D are respectively defined by the values 30, 45, 60, and 75 for percent time delay. Table 1 [table 8-1 from the Manual (1)] shows that for level terrain and 0 percent no passing zones the volume to capacity ( $v/c$ ) ratio associated with level of service C is 0.43. As terrain becomes more severe and the restrictions on passing become greater, the  $v/c$  ratio for each level of service decreases.

Neither the Manual (1) nor the research document (2) which formed the basis for the chapter on two-lane highways offer an explanation as to why the cut-off values between the levels of service were selected as 30, 45, 60, and 75 percent time delay. Also, strong reasons are not given as to why 5 seconds was selected as the definition for being delayed. Reference is made in the research document (2) to another study (3) in which 6 seconds was suggested as the definition for being delayed. However, in a recent paper covering a study in the Netherlands, Botma (4) found that a preliminary analysis of the data determined that no preference could be deduced from using 4, 5, or 6 seconds. Because 5 seconds appears to have been arbitrarily selected, it may be that a smaller value could be selected as long as it leads to reasonable  $v/c$  ratios for the various levels of service and as long as safety is not sacrificed.

It is the purpose here to offer two alternatives to the  $v/c$  ratio values given in table 1 for the level-of-service criteria for general terrain segments. The alternatives are based on the same percent time delay values used in the Manual (1) and on two different definitions for being delayed: 4 and 3.5 seconds.

First an argument will be offered as to why the 5-second headway may be too conservative a value by which to define a vehicle as being delayed and why 4 or 3.5 seconds might be more practical. It should be noted that at 60 mph the head-to-head spacing between two vehicles traveling at 5-seconds

TABLE 1 LEVEL-OF-SERVICE CRITERIA FOR GENERAL TWO-LANE HIGHWAY SEGMENTS (1, Table 8-1)

LOS	PERCENT TIME DELAY	v/c RATIO <sup>a</sup>																				
		LEVEL TERRAIN						ROLLING TERRAIN						MOUNTAINOUS TERRAIN								
		AVG <sup>b</sup> SPEED	PERCENT NO PASSING ZONES						AVG <sup>b</sup> SPEED	PERCENT NO PASSING ZONES						AVG <sup>b</sup> SPEED	PERCENT NO PASSING ZONES					
			0	20	40	60	80	100		0	20	40	60	80	100		0	20	40	60	80	100
A	≤ 30	≥ 58	0.15	0.12	0.09	0.07	0.05	0.04	≥ 57	0.15	0.10	0.07	0.05	0.04	0.03	≥ 56	0.14	0.09	0.07	0.04	0.02	0.01
B	≤ 45	≥ 55	0.27	0.24	0.21	0.19	0.17	0.16	≥ 54	0.26	0.23	0.19	0.17	0.15	0.13	≥ 54	0.25	0.20	0.16	0.13	0.12	0.10
C	≤ 60	≥ 52	0.43	0.39	0.36	0.34	0.33	0.32	≥ 51	0.42	0.39	0.35	0.32	0.30	0.28	≥ 49	0.39	0.33	0.28	0.23	0.20	0.16
D	≤ 75	≥ 50	0.64	0.62	0.60	0.59	0.58	0.57	≥ 49	0.62	0.57	0.52	0.48	0.46	0.43	≥ 45	0.58	0.50	0.45	0.40	0.37	0.33
E	> 75	≥ 45	1.00	1.00	1.00	1.00	1.00	1.00	≥ 40	0.97	0.94	0.92	0.91	0.90	0.90	≥ 35	0.91	0.87	0.84	0.82	0.80	0.78
F	100	< 45	—	—	—	—	—	—	< 40	—	—	—	—	—	< 35	—	—	—	—	—	—	

<sup>a</sup> Ratio of flow rate to an ideal capacity of 2,800 pcph in both directions.

<sup>b</sup> Average travel speed of all vehicles (in mph) for highways with design speed ≥ 60 mph; for highways with lower design speeds, reduce speed by 4 mph for each 10-mph reduction in design speed below 60 mph; assumes that speed is not restricted to lower values by regulation.

headway is 440 feet. This represents a per lane density of 12 vehicles per mile, the value associated with level of service A on freeways. At 4 and 3.5 seconds, the spacings are 352 and 308 feet, and the per lane densities are 15 and 17 vehicles per mile. Both of these conditions are associated with level of service B on a freeway.

From an operational standpoint, a vehicle being delayed by another slower moving vehicle follows at some distance until an opportunity to pass becomes available. A typical following headway just before the beginning of a pass could not be verified from the literature, but the research document (2) states that some headways as small as three-quarters of a second were observed during a two-lane highway study in Canada. An analysis of following in which the reaction time for the following driver is 1 second, the braking rates of the leader and follower are equal, and the "safety gap" after stopping is 5 feet, leads to a required following headway of less than 1.5 seconds for all initial speeds over 40 mph if both vehicles are cars. If the lead vehicle is a WB-50 truck, the required following headway is less than 2.2 seconds.

For purposes of this argument, a more conservative following headway of 2.5 seconds will be used to establish a definition for being delayed. It is assumed that the following vehicle has a desired speed of 60 mph and that it will always have a headway greater than 2.5 seconds behind the slower leading vehicle. It is further assumed that the following vehicle does not use the brakes to maintain this headway but that braking is achieved only by removing the driver's foot from the gas pedal. The definition for being delayed is, therefore, established to be the action of removing the foot from the gas pedal to maintain a headway of at least 2.5 seconds. Table

2 gives the headways associated with having to remove the foot from the gas pedal for various speeds of the leading vehicle and for various rates of deceleration in order that the following vehicle always trails by at least 2.5 seconds.

Consider as an example the case of the lead vehicle traveling at 45 mph, the following vehicle at 60 mph, and a deceleration rate of 1.5 ft/sec/sec. Based on the 45 mph speed and the minimum headway of 2.5 seconds, the following vehicle should have a head-to-head spacing of 165 feet. To achieve the speed change from 60 to 45 mph at a braking rate of 1.5 ft/sec/sec, the following vehicle decelerates for 14.67 seconds during which time the spacing between the two vehicles decreases by 61 feet. Therefore, the original spacing when the foot must be removed from the gas pedal must be 326 feet. This spacing is associated with a headway of 3.7 seconds based on the speed of 60 mph.

Table 2 shows that for the range of speeds and braking rates given the worst condition occurs when the lead vehicle has a speed of 45 mph and the deceleration rate is 1.5 ft/sec/sec. At speeds in the 40 to 60 mph range, the deceleration due to removing the foot from the gas is generally higher and in the 2.5 to 3.0 ft/sec/sec range (5, p. 24). Therefore, it appears that 3.5 seconds or a more conservative 4 seconds may be reasonable definitions for being delayed as alternatives to the 5 seconds used in the Manual (1).

To obtain v/c ratios for the 4- and 3.5-second definitions for being delayed, the procedures used to develop the ratios given in table 1 for the 5-second definition were examined. The research document (2) describes how a simulation model was used to establish the flow rates and, hence, the v/c ratios. The distribution of vehicle headways used as input to the

TABLE 2 HEADWAY TO CONSTITUTE DELAY

SPEED OF LEADING VEHICLE (mph)	HEADWAY AT WHICH DECELERATION MUST START (sec) at DECELERATION RATE (ft/sec/sec)				
	1.5	1.8	2.0	2.5	3.0
	55	2.5	2.5	2.5	2.5
50	2.9	2.8	2.7	2.6	2.5
45	3.7	3.4	3.3	3.0	2.8

Desired speed of following vehicle = 60 (mph)



TABLE 4 LEVEL-OF-SERVICE CRITERIA FOR GENERAL TWO-LANE HIGHWAY SEGMENTS—ROLLING TERRAIN

LOS	DEFINITION OF DELAY (sec)	v/c RATIO					
		Percent No Passing Zones					
		0	20	40	60	80	100
A	5	0.15	0.10	0.07	0.05	0.04	0.03
	4	0.19	0.14	0.10	0.08	0.06	0.05
	3.5	0.22	0.16	0.12	0.09	0.07	0.06
B	5	0.26	0.23	0.19	0.17	0.15	0.13
	4	0.33	0.28	0.24	0.20	0.18	0.17
	3.5	0.38	0.32	0.27	0.23	0.21	0.19
C	5	0.42	0.39	0.35	0.32	0.30	0.28
	4	0.51	0.46	0.41	0.37	0.35	0.33
	3.5	0.59	0.53	0.47	0.43	0.40	0.37
D	5	0.62	0.57	0.52	0.48	0.46	0.43
	4	0.79	0.73	0.67	0.62	0.59	0.56
	3.5	0.90	0.83	0.77	0.71	0.67	0.64
E	ALL	0.97	0.94	0.92	0.91	0.90	0.90

TABLE 5 LEVEL-OF-SERVICE CRITERIA FOR GENERAL TWO-LANE HIGHWAY SEGMENTS—MOUNTAINOUS TERRAIN

LOS	DEFINITION OF DELAY (sec)	v/c RATIO					
		Percent No Passing Zones					
		0	20	40	60	80	100
A	5	0.14	0.09	0.07	0.04	0.02	0.01
	4	0.18	0.13	0.09	0.05	0.03	0.01
	3.5	0.20	0.15	0.10	0.06	0.03	0.02
B	5	0.25	0.20	0.16	0.13	0.12	0.10
	4	0.31	0.25	0.20	0.16	0.13	0.11
	3.5	0.35	0.28	0.23	0.18	0.15	0.12
C	5	0.39	0.33	0.28	0.23	0.20	0.16
	4	0.48	0.41	0.35	0.30	0.26	0.23
	3.5	0.54	0.46	0.40	0.34	0.30	0.26
D	5	0.58	0.50	0.45	0.40	0.37	0.33
	4	0.73	0.64	0.57	0.50	0.45	0.41
	3.5	0.83	0.73	0.65	0.57	0.52	0.47
E	ALL	0.91	0.87	0.84	0.82	0.80	0.78

ditions for being delayed. Note that on both figures the lines for 0 percent no passing fall between the two dashed lines and that the 100 percent no passing lines fall to the left of the left dashed line. This relationship repeats that noted on figure 1 associated with the 5-second definition for being delayed.

The  $v/c$  ratios (tables 3 through 5) associated with the 4- and 3.5-second definitions for being delayed are approximate alternatives for the values given in the Highway Capacity Manual (1) for the 5-second definition. They should be considered approximate because they were not obtained using the simulation model. The comparison between figure 1 and figures 2 and 3 indicates that these  $v/c$  ratios are reasonable. The results indicated that for level terrain and 0 percent no passing zones (table 3), the  $v/c$  ratios for level of service C are 0.53 and 0.60 respectively for the 4- and the 3.5-second definitions

compared to 0.43 for the 5 seconds. A more complete comparison will be discussed later in relation to an example.

### SPECIFIC GRADES

The level-of-service criteria for specific grades on two-lane highways is not based on percent time delay but on the average speed of all the vehicles in the upgrade direction. Table 6 [table 8-2 of the Manual (1)] gives the speeds associated with each level of service. Based on this, the  $v/c$  ratios for grades from 3 to 7 percent are given in table 8-7 of the Manual (1) and repeated here for 3 and 7 percent grades in table 7.

As a result of having two different level-of-service criteria, one for general terrain segments and one for specific grades,



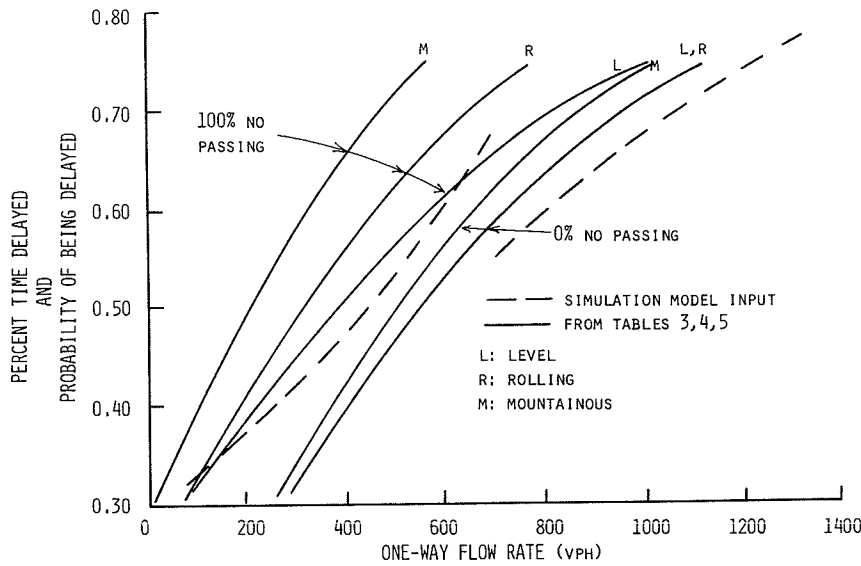


FIGURE 2 Delay-flow rate relationship for 4-second definition of delay.

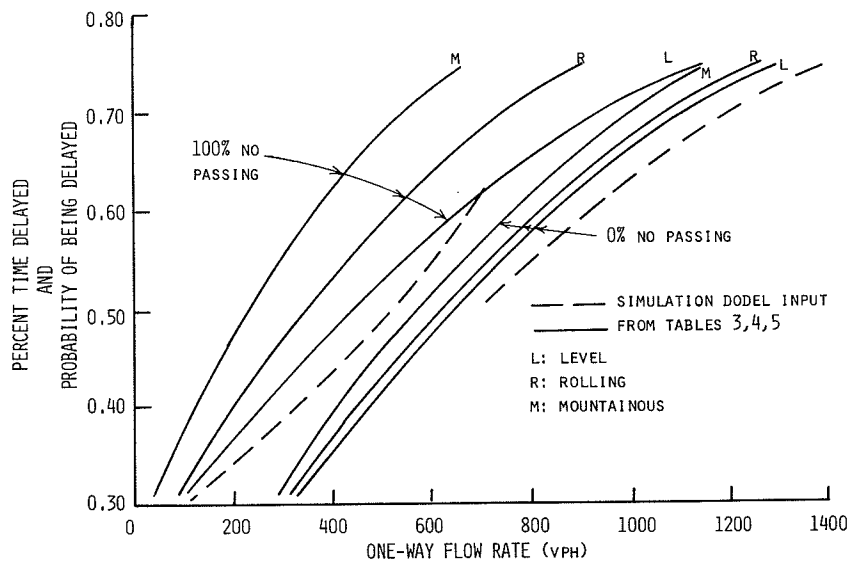


FIGURE 3 Delay-flow relationship for 3.5-second definition of delay.

it is possible for an inconsistency to occur in the level of service between level or rolling terrain segments and a moderate grade. This will be demonstrated by an example.

**COMPARATIVE EXAMPLE**

Consider a high-design, two-lane highway with design speed of 60 mph, 12 foot lanes, and usable shoulders at least 6 feet wide. The percent of no passing zones is taken as zero for level terrain, 10 percent for rolling terrain, and 20 percent for the 1-mile-long specific grades of 4 or 6 percent. The only no passing zone on the specific grades is near the top. The directional distribution of the traffic stream is 60-40 (for the spe-

cific grade, 60 percent is in the upgrade direction) and the only heavy vehicles are 12 percent trucks. The two-way service flows as computed by the procedures described in the Manual (1) and associated with each level of service are given in table 8 for the three definitions of being delayed.

The first four columns on figure 4 show the results obtained using the Manual's 5-second definition for being delayed for the general terrain segments and for the two specific grades. For the general terrain segments, levels of service A through C cover service flows up to about 990 and 720 vehicles per hour (vph), respectively, for level and rolling terrain. However, for the two specific grades, these first three levels of service cover service flows to about 1,630 and 1,100 vph for 4 and 6 percent grades, respectively. Notice that a service

TABLE 6 LEVEL-OF-SERVICE CRITERIA FOR SPECIFIC GRADES (I, Table 8-2)

LEVEL OF SERVICE	AVERAGE UPGRADE SPEED (MPH)
A	> 55
B	> 50
C	> 45
D	> 40
E	> 25 - 40 *

\* The exact speed at which capacity occurs varies with the percentage and length of grade, traffic compositions, and volume.

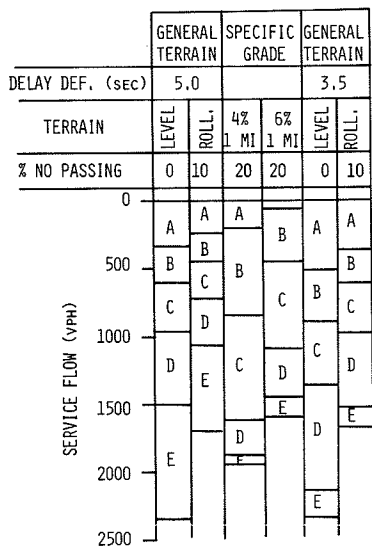
TABLE 7 VALUES OF v/c RATIOS FOR SPECIFIC GRADES (I, Table 8-7)

LOS	PERCENT GRADE	v/c RATIO					
		Percent No Passing Zones					
		0	20	40	60	80	100
A	3	0.27	0.23	0.19	0.17	0.14	0.12
	7	0.00	0.00	0.00	0.00	0.00	0.00
B	3	0.64	0.59	0.55	0.52	0.49	0.47
	7	0.34	0.27	0.22	0.18	0.15	0.12
C	3	1.00	0.95	0.91	0.88	0.86	0.84
	7	0.77	0.65	0.55	0.46	0.40	0.35
D	3	1.00	1.00	1.00	1.00	1.00	1.00
	7	0.93	0.82	0.75	0.69	0.64	0.59

TABLE 8 EXAMPLE: TWO-LANE HIGHWAY SERVICE FLOWS

LOS	DEFINITION OF DELAY (sec)	GENERAL TERRAIN LEVEL		SPECIFIC GRADE	
		ROLLING	4%-1mi	6%-1mi	
	% NO PASS. =	0%	10%	20%	20%
A	5	350	240	200	50
	4	450	320		
	3.5	520	370		
B	5	620	440	850	440
	4	780	540		
	3.5	900	620		
C	5	990	720	1630	1100
	4	1220	860		
	3.5	1380	1000		
D	5	1500	1060	1890	1460
	4	1900	1350		
	3.5	2160	1540		
E	ALL	2350	1700	1950	1600

NOTE: 12 Percent trucks



**FIGURE 4** Example comparison—5.0- and 3.5-second definitions of delay.

flow in the range 1,500 to 1,630 vph is in level of service E for the level terrain segment, but in level C for the 4 percent grade. Even the 6-percent grade has a higher service flow for level C than the level terrain segment. From an examination of figure 4, one might conclude that a two-lane highway built as a series of 1-mile grades in the range of 4 to 6 percent would provide a better level of service than a level road with no passing restriction. This conclusion is not in keeping with traditional thinking about two-lane highway grades.

The last two columns on figure 4 give the service flows for the 3.5-second definition for being delayed. It can be seen that this definition expands the range of service flows for levels of service A through D and reduces the range in level E compared to the 5-second definition. Service flows up to 1,380 vph on level terrain and 1,000 vph on rolling terrain are now in level of service C instead of near the limits of level D for the 5-second definition. In addition to expanding the range of service flows within the acceptable levels of service, the

3.5-second definition is also less inconsistent with the 4 percent specific grade than the 5-second definition. There is still an inconsistency at the boundary between levels of service C and D, but it is less severe than that associated with the 5-second definition.

**CONCLUSIONS**

The selection of 5 seconds as the definition for being delayed is not strongly supported in the 1985 edition of the Highway Capacity Manual (1), and any value between 2 and 6 seconds may be reasonable. Using 5 seconds as the definition for being delayed also produces an inconsistency in the level of service between certain general terrain segments and specific grades. The selection of 4 or 3.5 seconds as the definition for being delayed would ameliorate these problems.

A field study should be carefully designed and conducted to determine at what headway drivers begin to feel delayed by a leading vehicle. The study should determine if a driver's perception of being delayed is dependent on roadway parameters, such as design speed or posted speed, and the traffic volume conditions.

**REFERENCES**

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985, 516 pp., Chpt. 8.
2. J. Messer. Two-Lane, Two-Way Rural Highway Capacity. Final Report, NCHRP Project 3-28A. Texas Transportation Institute, College Station, Tex., February 1985.
3. J. F. Morall. Two-Lane, Two-Way Highway Capacity—Canadian Data and Input to NCHRP 3-28A. July 1981.
4. H. Botma. Traffic Operation on Busy Two-Lane Rural Roads in The Netherlands. In *Transportation Research Record 1091*. TRB, National Research Council, Washington, D.C., 1986, pp. 126-131.
5. *Transportation and Traffic Engineering Handbook*. Institute of Transportation Engineers, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1976.

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

