TRANSPORTATION RESEARCH
RECORD

No. 1494

Highway Operations, Capacity, and Traffic Control

Traffic Operations, Traffic Signal Systems, and Freeway Operations 1995

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1995

Transportation Research Record 1494

ISSN 0361-1981
ISBN 0-309-06161-X
Price: $37.00

Subscriber Category
IVA highway operations, capacity, and traffic control

Printed in the United States of America

Sponsorship of Transportation Research Record 1494

GROUP 1—TRANSPORTATION SYSTEMS PLANNING AND ADMINISTRATION
Chairman: Thomas F. Humphrey, Massachusetts Institute of Technology

Public Transportation Section
Chairman: Subhash R. Mundle, Mundle & Associates, Inc.

Committee on Light Rail Transit
Chairman: Thomas F. Larwin, San Diego Metropolitan Transportation Development Board
Secretary: John W. Schumann, LTK Engineering Services
Gregory P. Benz, Jack W. Boorse, David Calver, Douglas R. Campion, Thomas J. Carmichael, John H. Chaput, Ronald DeGraw, Donald O. Eisele, Rodney W. Kelly, S. David Phraner, Steven E. Polzin, Pilka Robinson, Peter J. Schmidt, Joseph S. Silien, John D. Wilkins, Nigel H. M. Wilson, Oliver W. Wischmeyer Jr., Richard Wolsfeld Jr.

GROUP 3—OPERATION, SAFETY, AND MAINTENANCE OF TRANSPORTATION FACILITIES
Chairman: Jerome W. Hall, University of New Mexico

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

Committee on Transportation System Management
Chairman: Wayne Berman, FHWA U.S. Department of Transportation
Secretary: Gary R. Erenrich, County of Fairfax
Guzin Akan, Thomas J. Higgins, Frannie F. Humplick, Rajendra Jain, Harvey R. Joyner, Steven Z. Levine, Jonathan L. Levy, Timothy J. Lomax, Michael D. Meyer, Susan L. Moe, Jouko A. Parviainen, George J. Scheuernstuhl, R. Sivanandan, Susanne Pelly Spitzer, Andrzej B. Tomecki, Cy Ulberg, Douglas W. Wiersig

Committee on High-Occupancy Vehicle Systems
Chairman: Donald G. Capelle, Parsons Brinckerhoff
Secretary: Dennis L. Christiansen, Texas A&M University
David E. Barnhart, John Bonsall, Donald J. Emerson, Charles Fuhs, Alan T. Gonseth, Leslie N. Jacobson, William A. Kennedy, Theodore C. Knappen, James R. Lightbody, Timothy J. Lomax, Adolf D. May Jr., Jonathan David McDade, C. J. O'Connell, R. L. Pierce, Lew W. Pratsch, Morris J. Rothenberg, Sheldon G. Strickland, Gary K. Trietsch, Katherine F. Turnbull, Carole B. Valentine, Jon Williams

Committee on Freeway Operations
Chairman: Jeffrey A. Lindley, Federal Highway Administration
Secretary: Peter M. Briglia Jr., Washington State Transportation Center
Daniel H. Baxter, Glen C. Carlson, Robert F. Dale, Walter M. Dunn Jr., Jack L. Kay, Randall A. Keir, Jim Kerr, Job J. Klijnhout, Peter R. Korpal, Robert E. Maki, Joseph M. McDermott, Nancy L. Nihan, Jeffrey E. Purdy, James R. Robinson, James F. Shea, William W. Stoeckert, Thomas Urbanik II, Thomas C. Werner, Sam Yagar

Committee on Traffic Signal Systems
Chairman: Herman E. Haenel, Advanced Traffic Engineering
Secretary: Alberto J. Santiago, Federal Highway Administration
Rahmi Akcelik, Edmond Chin-Ping Chang, David J. Clowes, Robert A. De Santo, Donald W. Dey, Gary Duncan, Raj Ghaman, Robert David Henry, Bahman Izadmehr, Leslie Kelman, Alfred H. Kosik, Feng-Bor Lin, Anson Nordby, David C. Powell, Walter T. Ragsdale, Ajay K. Rathi, Stephen Edwin Rowe, Miro Supitar, W. Scott Wainwright, Charles E. Wallace, Sam Yagar, Robert D. Yankovich

Transportation Research Board Staff
Robert E. Spicher, Director, Technical Activities
Richard A. Cunard, Engineer of Traffic and Operations
Peter L. Shaw, Public Transportation Specialist
Nancy A. Ackerman, Director, Reports and Editorial Services

Sponsorship is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 1994.

Transportation Research Record 1494

Contents

Foreword

The papers in this volume are from the 1995 Annual Meeting of the Transportation Research Board and are related by their focus on transportation systems management. They discuss intelligent transportation systems (ITS), high occupancy vehicle (HOV) systems, freeway operations, and traffic signal systems and cover a wide range of problems reflecting the concerns of theoreticians and practitioners.

These specific areas of traffic operations are receiving considerable attention because of the emphasis on ITS, provisions of the Intermodal Surface Transportation Efficiency Act (ISTEA), implications of the Clean Air Act Amendments, ever-increasing traffic congestion, and recognition of the importance of incident magagement for reducing nonrecurring traffic congestion.

Those readers with an interest in freeway operations will find papers related to freeway flows, freeway management, incident management, and ramp metering. The specific area of HOV systems has papers related to reversible lanes and evaluation of existing systems.

Traffic signal systems are examined in papers related to real-time traffic signal systems, bus priority systems, and computer models for traffic signal systems.

Analysis of Temporal and Spatial Variability of Free Speed Along a Freeway Segment

J. ALLEN STEWART, HESHAM RAKHA, AND MICHEL VAN AERDE

To determine the speed-flow relationship for a highway section, a number of parameters must be estimated. These include free speed, speed at capacity, capacity, and jam density. Because of fluctuations in demand, variations in driver behavior, and geometric and environmental conditions, these parameter values may vary both spatially for different stations and temporally for different days. To use these speed-flow relationships to estimate link travel times or diversion capacities, or for incident detection algorithms, these spatial and temporal variations in the speed-flow relationships need to be quantified so that accurate estimates of the relevant traffic parameters can be made. This work presents a statistical analysis of the variability of free speed estimates for 24 stations along a section of I-4 in Orlando, Florida during a 4-month period. This analysis is a first step in performing similar analyses of capacity, speed at capacity, and jam density. In the analysis presented in this work, it was found that free speed estimates along I-4 had a standard deviation of 4.7 km/hr and were most dependent on the location at which they were observed. This location factor explained 60 percent of the sum of squared errors. Minor variations in free speed from one day to another were overshadowed by these spatial differences and accounted for approximately 6 percent of the sum of squared errors. These two findings suggest that on this freeway section there is little loss in accuracy if many days of data are aggregated for a specific location, but a great loss in accuracy if many locations are averaged for the same day. There is also little to be gained by estimating day-of-the-week specific free speeds.

The objective of the research reported in this work was twofold. The primary purpose was to ascertain whether there were statistically significant differences in free speed estimates from one location to the next, or from one day to the next. In the absence of such variations, it would be sufficient to calibrate a speed-flow relationship for an entire highway section based on either 1 day's worth of data at a single station or a composite single data set of all stations and days combined. The second objective was to determine whether significant temporal or spatial differences existed in the estimated free speeds. For example, it is often perceived that midweek (Tuesday through Thursday) traffic behavior is different from driver behavior on Friday or Monday. If this perception is substantiated, then it would be necessary to establish different free speeds for the same section of highway to model these different types of days.

The characteristics of the study network and the data collection time frame are presented, followed by an overview of the study procedure. The details of the Analysis of Variance (ANOVA) tests are then described, followed by the conclusions of this study.

J. A. Stewart, Royal Military College, Kingston, Ontario, Canada K7K 5L0. H. Rakha and M. Van Aerde, Queen's University, Kingston, Ontario, Canada K7L 3N6.

STUDY DESCRIPTION

Network Configuration

A 16-km (10-mi) portion of the I-4 freeway in Orlando, Florida was considered in this study, modeled as part of an Intelligent Vehicle Highway System—Institute of Transportation Studies benefit assessment. I-4 is a major route that extends across the center of Florida from the southwest (Tampa) to the northeast (Daytona), passing by the Disney World complex to the west of the study area. The detectorized portion of the I-4 freeway is located near downtown Orlando, extending from 33rd Street to the southwest, and ending downstream of Maitland Boulevard to the northeast, as illustrated in Figure 1.

A total of 24 loop-detector stations were located along I-4, numbered from 1 to 25, with no data provided for Station 10. The spacing of the detector stations ranged from approximately 0.40 to 0.90 km (0.25–0.54 mi). There were no major terrain variations along the detectorized section of the I-4 freeway, as Orlando is rather flat. However, at many interchanges with arterials, the freeway was elevated. The entire detectorized section of I-4 was composed of three lanes in each direction.

Data Collection Time Frame

The analysis period included traffic data for portions of a 4-month period during the winter of 1992–1993. The data included 11 days in November 1992, 29 days in January 1993, 26 days in February 1993, and 11 days in March 1993. This data set amounted to a total of 75 days of 30-sec data, with approximately 10 different days of data available for each day of the week.

The Freeway Management Center (FMC) dual loop detectors measured and logged the flow, occupancy, and space mean speed for each of the three lanes at 30-sec intervals. These data were aggregated for this analysis into 5-min data summaries to reduce the amount of data to be handled, while still capturing most of the variability in the traffic conditions. An average lane flow, occupancy, and mean speed estimate for each station were generated from the individual loop detector measurements. In estimating the average speed at a specific station, the loop speeds were weighted by the volume on each set of loops.

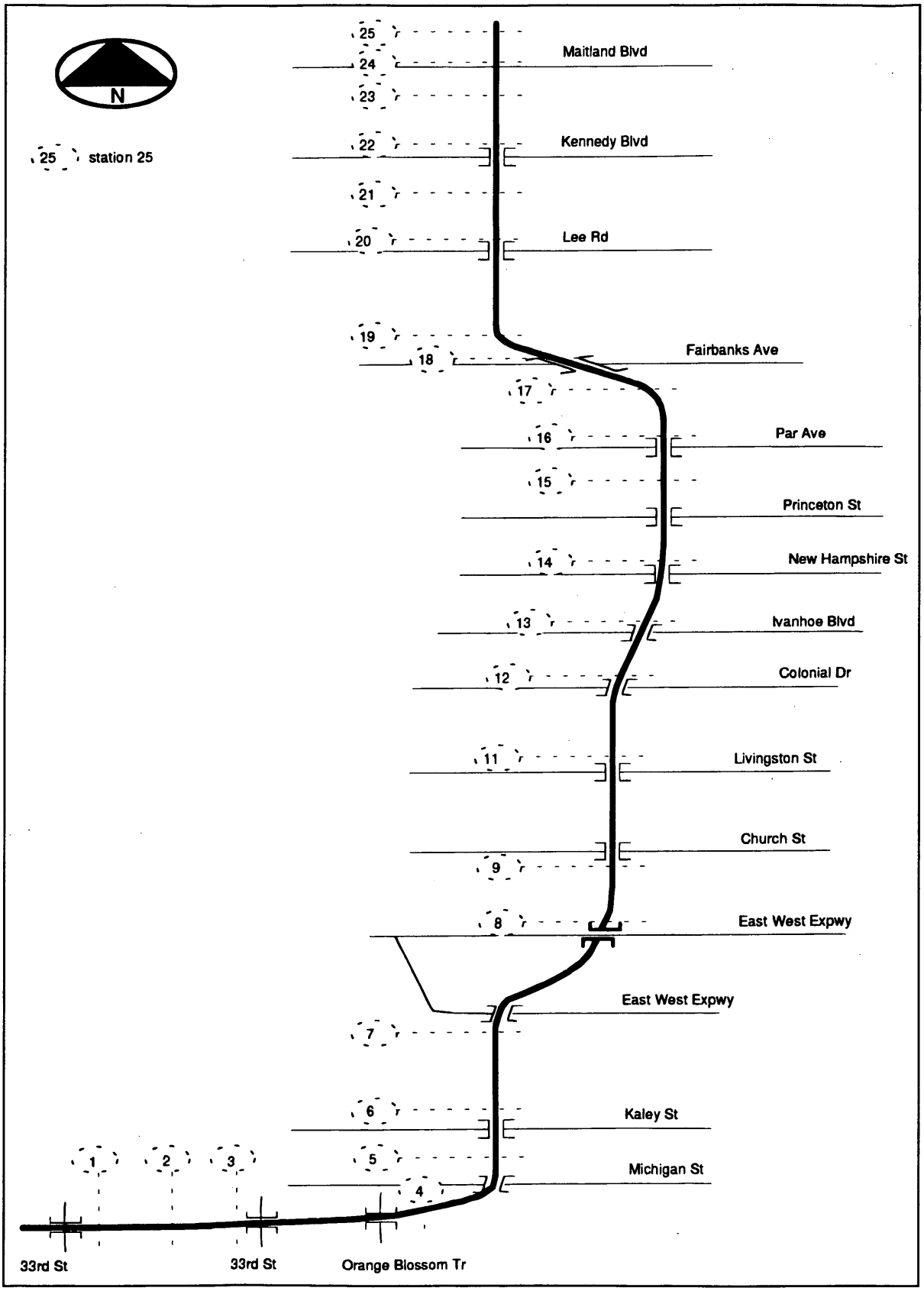


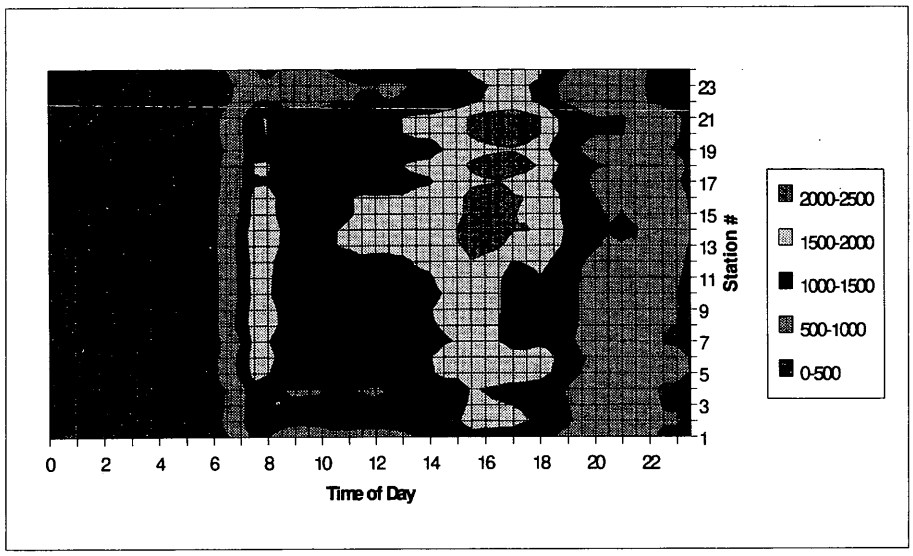
FIGURE 1 Location of FMC detector stations along I-4 freeway

OVERVIEW OF STUDY PROCEDURES

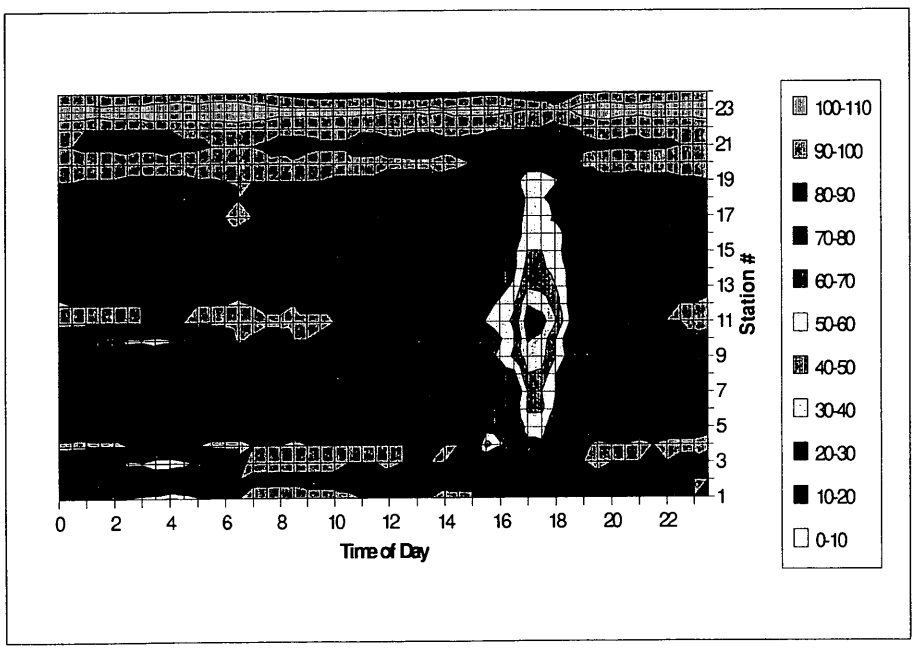
Typical Traffic Conditions Along I-4

Based on the FMC data for all of the available days within the 4-month period, it was possible to generate surfaces that represented the average speed, average flow, or average occupancy at a particular station and at a particular time of day. Figure 2(*a*) represents the resulting average flow surface in the eastbound direction along the I-4 section. The x-axis represents the time period from 0 at midnight at the start of the day, to 24 at midnight at the end of the day, and the y-axis represents the station numbers that are traversed. The eastbound flow proceeds in the upward direction from Station 1 to Station 25. For each cell combination of time of day and station, the z-axis represents the average hourly lane flow rate measured.

Figure 2(*a*) shows that the flow gradually increased at 6:00 a.m. along all eastbound stations until it reached a flow of approximately 2,000 vehicles per hour (vph) per lane at 8:00 a.m. along detector



(*a*)



(*b*)

FIGURE 2 (*a*) Temporal and spatial variation in 30-min. eastbound average lane flow (vph/lane). (*b*) Temporal and spatial variation in 30-min. eastbound average lane speed (km/hr).

Stations 5 through 18. The flow increased again during the p.m. peak from approximately 3:00 p.m. until 6:30 p.m. at Stations 12 through 22. Figure 2(*a*) shows that the flow from 5:00 to 7:00 p.m. at Stations 7 through 12 was lower than 2,000 vph/lane. However, after examining Figure 2(*b*), it appears that during this period the speeds were also low (20 to 40 km/hr). Thus, the lower flow measurements were most likely caused by the presence of congestion, and not by a reduction in demand.

Figure 2(*b*) illustrates that only Stations 5 through 20 experienced speeds near or in the congested portion of the speed-flow relationship (speeds less than 60 km/hr) during the p.m. peak in the eastbound direction. As subsequent analyses of the variability of speed at capacity, capacity, and jam density required the use of congested data, only stations with congested data were subsequently considered for the fitting of a complete speed-flow relationship.

Structure of Speed-Flow Relationships

The selection of a particular shape for a speed-flow relationship has been a topic discussed by traffic engineers for more than 50 years. May (*1*) provides an excellent discussion and comparison of the various single- and multiregime models, and describes their respective strengths and limitations in the context of producing reasonable free speed, speed at capacity, capacity, and jam density estimates. In response to these limitations, a new single regime speed-flow relationship was developed [see Figures 3(*a*) and 3(*b*)]. This relationship is described in detail by Van Aerde (*2*). The main features are the highly linear and almost horizontal behavior in the uncongested region, the speed at capacity in excess of one-half of the free speed, and the jam density value, which is higher than two times the density at capacity, yet still finite. Of particular interest to this study is the fact that the free speed, which theoretically occurs when the volume is 0, can be extrapolated reliably from the near-linear uncongested portion of the curve. An average of the speeds observed when flows are below a given maximum flow threshold (e.g., $V/C < 0.5$) would always represent an underestimate of the free speed in view of the small negative slope of the curve in this region.

Estimation of Speed-Flow Parameters

Figures 3(*a*) and 3(*b*) illustrate sample fits of Equation 1 and its density counterpart to data collected over an entire day at Detector Station 13. The discrete points represent the 5-min loop detector measurements, and the continuous curve represents the fit estimated by the curve-fitting model.

To generate the free speed estimates at each station, a heuristic curve-fitting program was developed that selects the speed-flow relationship parameters that produce the minimum normalized square error in a three-dimensional flow-speed-density data space (*3*). This curve-fitting model estimates four parameters, namely free speed (u_f), speed at capacity (u_c), capacity (q_c), and jam density (d_j). The structure of the speed-flow relationship is represented in Equation 1. Equations 2 through 5 are used to calculate the three model parameters, c_1, c_2, and c_3, in addition to the intermediate parameter k.

It appears from Figure 3(*a*) that the macroscopic relationship captures most of the deterministic variation in speed-flow while achieving a reasonable compromise estimate when stochastic variability exists. The four parameters for Station 13 selected by the model were: u_f = 87.2 km/hr, u_c = 70.6 km/hr, q_c = 1,925 vph, and d_j = 92.2 vehicles/km. In Figure 3(*a*), free speed is identified as the higher y-axis intercept, speed at capacity is the y-axis value that corresponds to the maximum flow point (nose of curve), and capacity is the maximum x-value. Jam density is the inverse of the slope of the fitted curve as it emerges at the origin to the axes in Figure 3(*a*), but it is more easily identified as the x-value at 0 speed in Figure 3(*b*).

$$q = \frac{u}{c_1 + \frac{c_2}{u_f - u} + c_3 u} \tag{1}$$

$$k = \frac{c_1}{c_2} = \frac{(2u_c - u_f)}{(u_f - u_c)^2} \tag{2}$$

$$c_2 = \frac{1}{d_j\left(k + \frac{1}{u_f}\right)} \tag{3}$$

$$c_1 = kc_2 \tag{4}$$

$$c_3 = \frac{-c_1 + \frac{u_c}{q_c} - \frac{c_2}{(u_f - u_c)}}{u_c} \tag{5}$$

where

c_1 = fixed distance headway constant (km),
c_2 = first variable distance headway constant (km²/hr),
u_f = free speed (km/hr),
u_c = speed at capacity (km/hr),
u = prevailing speed associated with headway h (km/hr),
q = flow rate of traffic traveling at speed u (km/hr) (vph),
q_c = flow at capacity (vph),
d_j = jam density (vehicle/km), and
k = dimensionless constant to set the speed at capacity relative to the free speed.

It should be noted in Figures 3(*a*) and 3(*b*) that to generate a satisfactory fit of a typical speed-flow relationship's jam density and speed at capacity at a specific location, sufficient data points in both the uncongested and the congested portion of the curve are required. This is the basis for the fact that the curve-fitting model was set to not estimate the desired four parameters if no points existed in the congested region (speeds less than 60 km/hr) of the curve.

Typical Spatial and Temporal Variation in Free Speed

Figure 4 demonstrates the temporal and spatial variation in the free speed estimates at Stations 9 to 22 over a sample 10-day period. Because of a lack of points in the congested portion of the speed-flow relationship at Stations 1 to 8 and Stations 23 to 25, the curve fits and therefore the free speed estimation was performed only for the stations located in the downtown area (Stations 4 to 22). The surface plot shows that the free speed ranged from 80 to 110 km/hr. It appears that the free speeds were relatively constant during the 10-day period, as indicated by the minor variations in the y-axis direction. However, the speeds varied to a greater extent for the different locations along the x-axis. The variation in free speed was in the range of approximately ñ15 percent of the average free speed

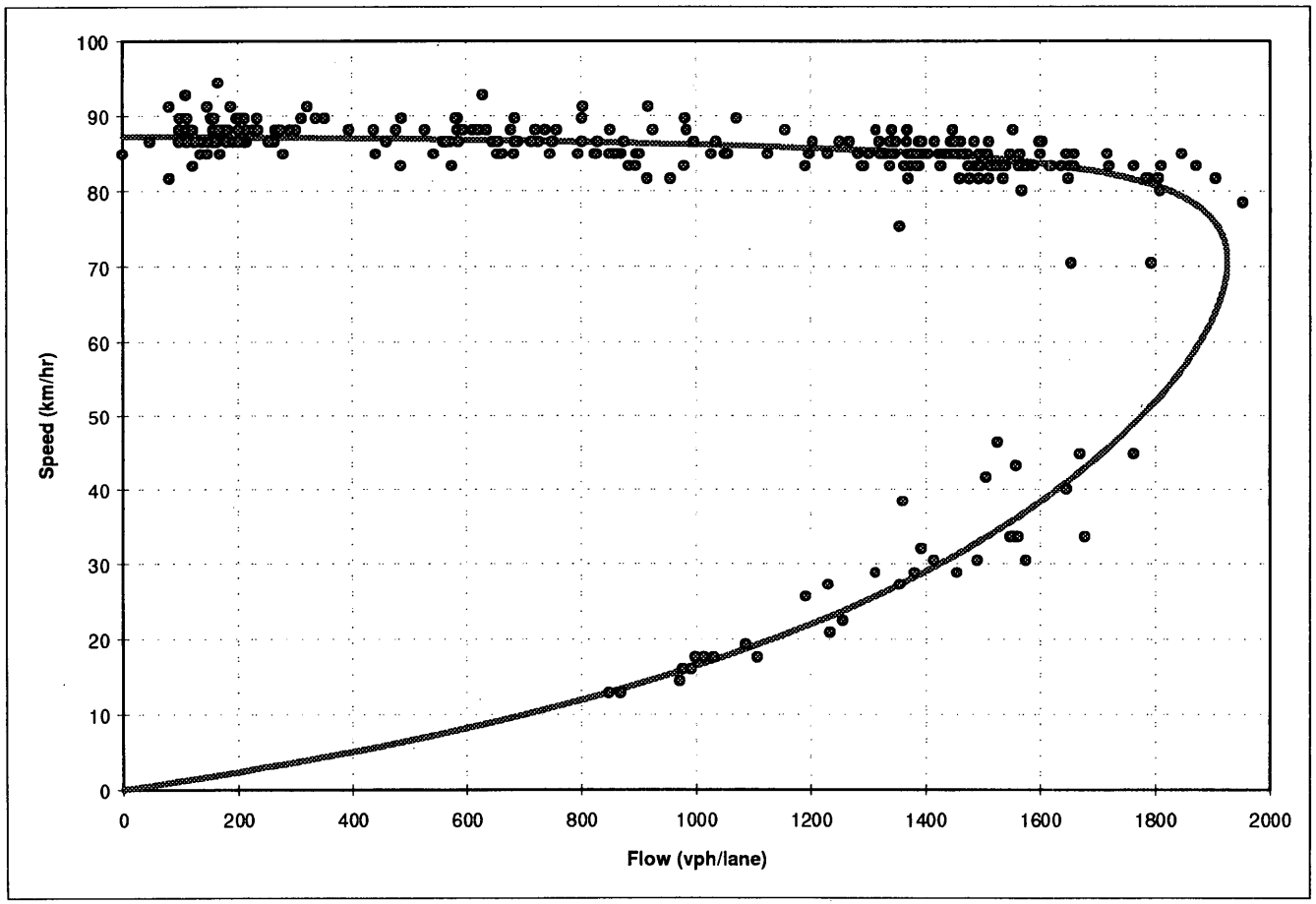


FIGURE 3 (*a*)A typical speed-flow fit to I-4 data (u_f = 87.2 km/hr, u_c = 70.6 km/hr, q_c = 1,925 vph, d_j = 92.2 veh/km).

(continued on next page)

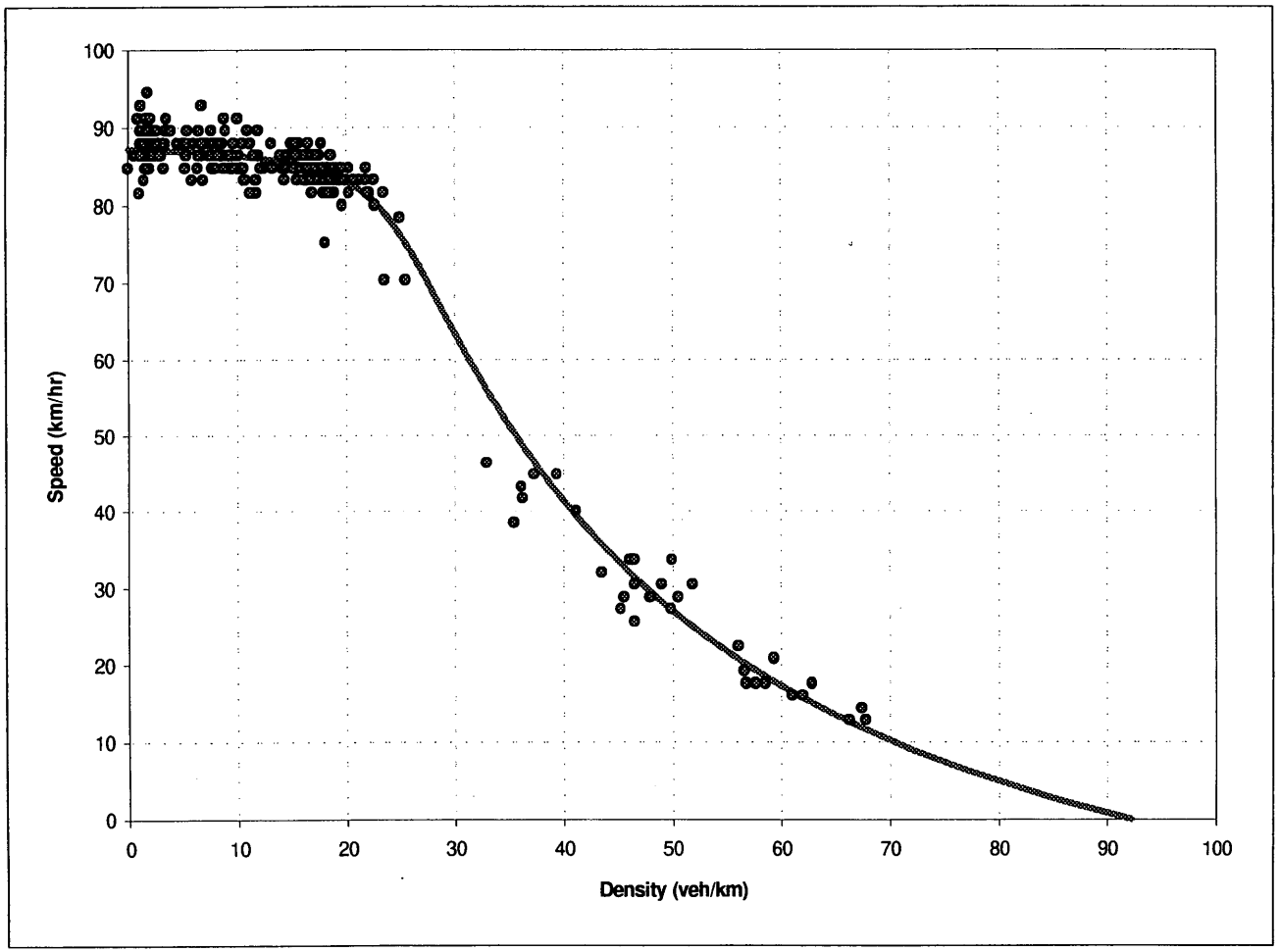


FIGURE 3 (*b*) A typical speed-density fit to I-4 data (u_f = 87.2 km/hr, u_c = 70.6 km/hr, q_c = 1,925 vph, d_j = 92.2 veh/km.

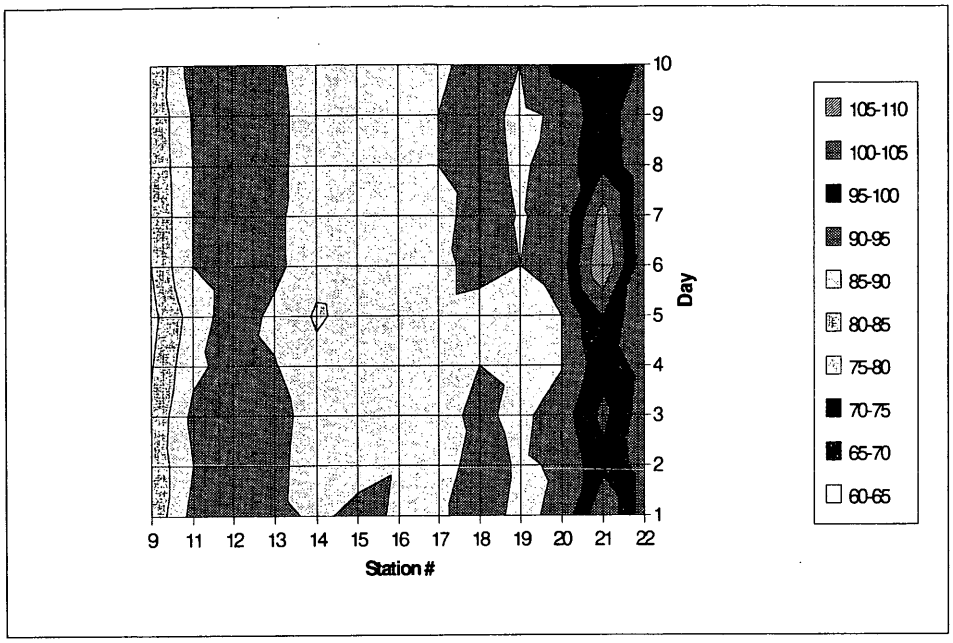


FIGURE 4 Temporal and spatial variation in free speed along I-4 (km/hr).

and had a standard deviation (SD) of 4.7 km/hr. A more detailed analysis of the free speed variation follows.

INTRODUCTION TO ANALYSIS OF VARIANCE

An examination of the speed contours in Figure 4 suggests that the free speed is much more spatially dependent than temporally dependent. This qualitative assessment prompted a statistical ANOVA of the free speed data to ascertain whether different days of the week or different station locations, or both, affected the value of the free speed in a statistically significant fashion. To complete this analysis, a data set of free speeds as a function of the day of the measurement (75 different days) and the location (24 station locations) was produced for subsequent analysis using SYSTAT (*4*).

Screening of Data

For this data set, data for the eastbound direction at Stations 1, 2, 3, 10, 23, and 24, and for the westbound direction at Stations 1, 2, 3, 4, 5, 10, 13, and 21 were removed because of a lack of congested data. Data from most Saturdays and Sundays were also removed. After removing several other days for the same reason, approximately 31 days of acceptable data remained for the westbound direction and 33 days of acceptable data remained for the eastbound direction.

Analysis Scenarios

After the data set had been conditioned, a series of ANOVAs was carried out (*5*, *6*). A brief introduction to the procedure follows.

The data set was split first into two main sets, one for the eastbound and one for the westbound direction. These data sets were treated separately for the rest of the analyses. Three different ANOVA models were fit for the aggregated data set. A one-way ANOVA was performed on calendar date (Analysis 1*a*) and the location factor was analyzed (Analysis 1*b*). In the third analysis, a two-way ANOVA was conducted because the date factor and location factor could be significant (Analysis 1*c*). Because the data set contained free speeds for each location for several weeks, it was possible to group the data by the day of the week rather than by the calendar date. This grouping permitted a one-way ANOVA with replication of measurement (Analysis 2).

To explore the effect of location and date within a single week of data, the eastbound direction for Monday, January 25, through Friday, January 29, 1993, was analyzed. For the westbound direction, screening of the data set made it impossible to find a continuous Monday through Friday period. Hence, the period from Friday, January 22, to Thursday, January 28, (excluding the weekend) was analyzed. As with the entire data set, three ANOVAs were fit, along with two one-way ANOVAs (grouped either by day or location) and one two-way ANOVA (Analyses 3*a*, 3*b*, and 3*c*, respectively).

An additional set of ANOVAs was fit to explore the premise that traffic behavior during the core midweek period (Tuesday, Wednesday, and Thursday) is different from Monday or Friday. For this reason, a two-way ANOVA with replication, similar to Analysis 2, was performed on a data set of Tuesday, Wednesday, and Thursday data (Analysis 4).

A total of 13 different analyses of variance models was fit for each direction. Typical data sets used in these analyses are shown in Tables 1 and 2. Tables 3 and 4 summarize the results of the most important ANOVAs performed. The following sections discuss each series of ANOVAs.

Table 2 shows that the mean free speed changed significantly along the route in the eastbound direction. The westbound direction experienced a similar change in free speed; however, because of space limitations, the results are not presented in this work. The large drop in free speed at Station 9 in the eastbound direction and

TABLE 1 Data Set for Eastbound Free-Flow Speed km/hr (Mondays)

Station	02-Nov-92 MON	09-Nov-92 MON	25-Jan-93 MON	01-Feb-93 MON	22-Feb-93 MON	29-Mar-93 MON	MEAN	STD
4	92.2	90.9	90.0	93.1	97.5	92.5	92.70	2.38
5	87.5	85.3	82.5	87.2	89.1	86.9	86.42	2.07
6	87.5	89.4	84.7	88.1	84.4	84.1	86.37	2.05
7	87.5	90.0	85.3	86.3	89.4	86.9	87.57	1.66
8	90 .0	90.9	83.1	87.5	86.9	87.2	87.60	2.50
9	85 .0	82.5	78.1	80.0	79.7	77.8	80.52	2.52
11	91.3	91.3	88.8	90.0	89.4	88.8	89.93	1.05
12	95.6	95.0	92.5	92.5	90.6	91.3	92.92	1.82
13	91.3	92.5	88.8	91.3	96.6	90.3	91.80	2.42
14	88.4	87.5	83.8	86.6	89.4	86.6	87.05	1.76
15	90.6	89.4	85.0	86.9	88.4	90.0	88.38	1.92
16	88.8	95.6	87.2	88.4	92.5	87.5	90.00	3.05
17	86.6	91.6	89.1	88.8	86.3	88.8	88.53	1.76
18	89.4	93.8	90.3	91.6	90.0	90.6	90.95	1.44
19	88.8	90.9	88.1	90.0	93.8	89.4	90.17	1.85
20	91.3	92.2	90.6	92.5	100.6	92.5	93.28	3.34
21	97.5	100.0	103.1	110.0	102.5	105.9	103.17	4.02
22	91.3	93.1	90.0	91.3	95.6	91.3	92.10	1.81
MEAN	**90.03**	**91.22**	**87.83**	**90.12**	**91.26**	**89.36**	**89.97**	**4.41**
STD	**2.99**	**3.78**	**5.10**	**5.70**	**5.56**	**5.24**	**1.17**	

TABLE 2 Data Set for Eastbound Free-Flow Speed km/hr (January 25-29, 1993)

Station	25-Jan-93 MON	26-Jan-93 TUE	27-Jan-93 WED	28-Jan-93 THUR	29-Jan-93 FRI	MEAN	STD
4	90	91.9	92.2	95	91.3	92.08	1.64
5	82.5	82.5	85.3	86.9	88.1	85.06	2.27
6	84.7	86.3	90	88.8	89.1	87.78	1.97
7	85.3	84.4	86.9	87.5	86.9	86.20	1.16
8	83.1	83.1	85.6	85.9	85	84.54	1.21
9	78.1	77.5	80	79.4	80	79.00	1.02
11	88.8	85.3	90	91.3	90.3	89.10	2.08
12	92.5	89.4	91.6	93.8	93.8	92.22	1.64
13	88.8	87.5	89.7	90.9	90.6	89.50	1.24
14	83.8	82.5	85.3	85.3	85.6	84.50	1.18
15	85	83.8	86.3	89.4	90.3	86.96	2.50
16	87.2	85	87.5	88.8	88.8	87.46	1.39
17	89.1	87.5	88.1	88.1	87.5	88.06	0.59
18	90.3	87.5	91.9	92.5	90.6	90.56	1.73
19	88.1	86.3	89.4	88.8	89.4	88.40	1.15
20	90.6	90	90.6	91.9	90.6	90.74	0.62
21	103.1	93.8	100	99.4	97.5	98.76	3.07
22	90	88.8	91.3	91.3	91.3	90.54	1.01
MEAN	**87.83**	**86.28**	**88.98**	**89.72**	**89.26**	**88.42**	**4.01**
STD	**4.96**	**3.63**	**3.92**	**4.08**	**3.49**	**1.24**	

TABLE 3 Summary of ANOVA for Eastbound Free-Flow Speed Along I-4 Freeway in Orlando, Florida

ANALYSIS TYPE	SAMPLE	MEAN SUM OF SQUARES						F		F_{crit}	
	SIZE	Station	(%)	Date	(%)	Error	(%)	Station	Date	Station	Date
Analysis 1 (without replication)	594	622.74	96%	23.43	4%	3.08	0%	202.36	7.61	1.64	1.47
Analysis 2 (with replication)	450	467.77	97%	9.45	2%	6.58	1%	94.04	1.90	1.65	2.40
Analysis 3 (25 Jan-29 Jan 93)	90	85.25	70%	34.34	28%	1.60	1%	53.31	21.47	1.78	2.51
Analysis 4 (midweek with repl.)	330	294.57	97%	3.46	1%	5.38	2%	65.00	0.47	1.67	3.04

TABLE 4 Summary of ANOVA for Westbound Free-Flow Speed Along I-4 Freeway in Orlando, Florida

ANALYSIS TYPE	SAMPLE	MEAN SUM OF SQUARES						F		F_{crit}	
	SIZE	Station	(%)	Date	(%)	Error	(%)	Station	Date	Station	Date
Analysis 1 (without replication)	496	310.09	92%	21.86	6%	6.04	2%	51.32	3.62	1.69	1.48
Analysis 2 (with replication)	400	244.03	86%	29.32	10%	9.51	3%	48.87	5.87	1.70	2.40
Analysis 3 (25 Jan-29 Jan 93)	80	48.98	72%	16.44	24%	2.16	3%	22.67	7.61	1.84	2.53
Analysis 4 (midweek with repl.)	240	128.93	70%	45.64	25%	10.85	6%	22.99	8.13	1.72	3.04

a corresponding increase in free speed in the westbound direction are most likely caused by an uphill grade at Station 9 in the eastbound direction. The large increase in free-flow speed at Station 20, however (in the east- and westbound directions), is most likely due to a change in free speed limit from 88 km/hr (55 mph) to 104 km/hr (65 mph).

Analysis 1: ANOVA of Entire Data Set (Monday–Friday)

The data were first grouped by calendar date to test for the significance of the calendar date factor on the free-flow speed for both the east- and westbound directions (Analysis 1*a*). The one-way ANOVA results indicated that the free-flow speed was not significantly different, at the 95 percent confidence level, from one day to the next. When these data were grouped by the location (Analysis 1*b*), the one-way ANOVA revealed that the free speed varied significantly from one location to the next. Finally, when both variables were included in a two-way ANOVA without replication (Analysis 1*c*), the results indicated that both the calendar date factor and location factor were statistically significant.

The summary results of the latter two-way ANOVA analysis are given in Tables 3 and 4, which show that the largest amount of variation (as indicated by the mean sum of squares is accounted for by the station factor. In the eastbound direction 96 percent, and in the westbound direction 92 percent of the variation in the data was due to the location factor. Four and 6 percent of the error in the respective directions was explained by the factor that accounts for the calendar day on which the data were collected. These percentages are based on mean square ratios.

For the total sum of squared errors, the error explained by the station factor was approximately 60 percent. Consequently, when specifying speed-flow relationships for this highway it is more important that a different relationship be developed for each location along the route than for each separate day. The observed minor differences from day to day prompted the analyses to determine whether the differences were systematic or random.

Analysis 2: Two-Way ANOVA with Replication (Monday–Friday)

Analysis 1*c* indicated that the free speed at a specific location did vary to some extent with the day on which the data were measured; therefore, an ANOVA was carried out to determine whether a day-of-the-week factor was a systematic source of these differences. In other words, the analysis was done to learn whether traffic behavior varies in a consistent fashion from Monday to Tuesday or Thursday to Friday. The results of these analyses are referred to as Analysis 2 and are indicated in Line 2 of Tables 3 and 4. The mean sum of squares shows that for the eastbound direction very little, if any, differences occurred between the different days of the week, as the F statistic indicated that the day-of-the-week factor is not significant at the 95 percent level of confidence ($1.90 < 2.40$). However, in the westbound direction the location factor is still the most important source of variation ($48.87 > 1.70$); there is a statistically significant difference between each day of the week ($5.87 > 2.40$). At this stage of the research, the reason the east- and westbound directions produce different results remains unclear.

Analysis 3: One Week of Data

The next series of analyses, referred to as Analysis 3, examined a continuous period of 5 weekdays. The purpose of this analysis was to determine whether a week of data would be sufficient to determine an average free-flow speed at a specific location. The results of this analysis are given in Line 3 of Tables 3 and 4. As with the entire data set for Analysis 1, three different ANOVAs were performed for each direction, two one-way analyses of the calendar

date (Analysis 3*a*) and location (Analysis 3*b*), and one two-way analysis without replication (Analysis 3*c*) using both factors. In both directions it was found that although the location factor was still the most significant factor, the day-of-the-week factor was also statistically significant. In the eastbound direction 25 percent and in the westbound direction 24 percent of the variation in free speed was due to the day-of-the-week factor. This would suggest that it is not possible to obtain a representative estimate of the free speed at a specific location by gathering data on only one day of the week.

Analysis 4: Midweek Only

Analysis 3 indicates that there were differences between the free speed obtained from one day to the next. It is often hypothesized that midweek period behavior is different from Friday or Monday behavior. As such, it might be possible to calculate two different estimates of free speed, one for each portion of the week. An analysis of the midweek data was performed to ascertain whether these temporal differences in Analyses 1 to 3 could be adequately explained by simply having midweek and Monday through Friday data grouped together. The results of this analysis, referred to as Analysis 4, are given on Line 4 of Tables 3 and 4. The values for the Mean Sum of Squares and the F statistic indicate that in the eastbound direction there is no statistically significant difference in the free speed from one day to the next during the midweek period. Virtually all of the variation in the data is explained by the location factor. Therefore, it is possible to obtain a location-specific measure of free speed for the midweek period. However, this is not the case in the westbound direction. In fact, 24 percent of the variation is due to the day on which the data were measured. This finding is consistent with the results obtained during the analysis of the entire data set using replication (Analysis 2).

CONCLUSIONS AND RECOMMENDATIONS

Several conclusions can be drawn from the analyses presented in this work. Although these conclusions are based on the specific I-4 data, the authors believe that the trends in the I-4 freeway behavior are representative of many typical freeways in North America and that the analysis used is applicable elsewhere.

First, free speeds along I-4 depend most strongly on the location where they are observed. Changes in geometry, ramp location or configuration, and speed limit may all be responsible for the observed significant differences in free speed as a function of the location factor. Second, minor variations in free speed from one day to another are due to differences between midweek versus weekend characteristics.

It is therefore recommended that when analyzing freeways such as I-4, location-specific free speeds be estimated first. Subsequently, day-of-the-week specific adjustments may be made, but these will have a less significant effect. However, even when these factors have been accounted for, some residual day-to-day variations will remain.

REFERENCES

1. May, A. D. *Traffic Flow Fundamentals*. Prentice Hall, NJ, 1990.
2. Van Aerde, M. *A Single Regime Speed-Flow Density Relationship for Congested and Uncongested Highways*. TRB, National Research Council, Washington, D.C., Jan. 1995.
3. Van Aerde, M., and H. Rakha. *Multivariate Calibration of Single Regime Speed-Flow-Density Relationships*. Queen's University, Ontario, Canada, in press.
4. *SYSTAT for Windows: Statistics, Version 5 Edition*. SYSTAT, Inc., Evanston, Ill., 1992.
5. Crow, E. L., F. A. Davis, and M. W. Maxfield. *Statistics Manual*. Book 0-486-60599-X. Dover Publications, Inc., 1960.
6. Draper, N., and H. Smith. *Applied Regression Analysis*, 2nd ed. John Wiley and Sons, Inc., New York, 1981.

Publication of this paper sponsored by Committee on Freeway Operations.

A Case for Freeway Mainline Metering

KEVIN A. HABOIAN

In this work, the merits of freeway mainline metering as a means of better managing freeway traffic congestion are explored. Freeway mainline metering involves controlling the amount of traffic entering a freeway segment to provide improved travel downstream of the control area. To date, mainline metering has not been applied to a typical urban freeway system, although the concept has been applied successfully to bridges and tunnels. Experiences at these bridges and tunnels indicate that in the presence of a bottleneck, regulating the number of vehicles through the bottleneck will result in improved freeway operations. This study investigates whether regulating mainline vehicle movements can also improve freeway operations without the presence of a bottleneck. Also addressed in this work is whether mainline metering can provide additional benefits over and above typical ramp metering. To evaluate the above hypotheses, the INTRAS simulation model was used to replicate freeway traffic operations. The mainline, metering evaluation was based on a variety of mainline volume and on-ramp control conditions. The results indicate that mainline metering can improve freeway operations downstream of the mainline meter. Most importantly, this can be accomplished without increasing the overall delay for vehicles originating upstream of the metering location. In addition, vehicles accessing the freeway from metered on-ramps downstream of the mainline meter are no longer entering a congested freeway mainline, thus reducing overall travel time. These findings appear to indicate that mainline metering is an appropriate freeway management tool.

Freeway mainline metering, which involves controlling the amount of traffic entering a freeway segment to provide improved travel downstream of the control area, is a concept that has previously invoked fear in many transportation engineers and public representatives. This fear stems from perceptions of queues, travel delays, and accidents that would increase congestion and travel times beyond those being experienced without the mainline metering control strategy. As a result of this fear, there has been very limited application of mainline metering and very little empirical data from which to confirm or challenge the above perception.

Consequently, transportation professionals have employed more accepted control strategies as a means of managing congestion. Such strategies have included ramp metering, high occupancy vehicle (HOV) lanes, variable message signs, highway advisory radio, incident response teams, etc. Though these strategies have improved operations on many freeways and highways, even the most sophisticated combination of these strategies has failed to manage congestion to ensure optimal traffic operations. By the year 2005, delay caused by congestion is projected to be five times what it was in 1984 (*1*).

The essential problem is that all inputs (defined in this paper as vehicles entering a section of the freeway via either on-ramps or the upstream mainline) to the freeway are not properly managed. Until they are, freeways will continue to be susceptible to gridlock conditions. It is appropriate to consider traffic management strategies that can better manage freeway operations, including strategies that manage or control the freeway mainline. In this work, the merits of freeway mainline metering to better manage freeway traffic congestion are explored in greater detail.

Parsons Brinckerhoff Quade & Douglas, Inc., 505 South Main Street, Suite 900, Orange, Calif. 92668.

PREVIOUS MAINLINE METERING EXPERIENCE

To date there has been no application of mainline metering on an urban freeway system. However, this concept has been applied successfully to bridges and tunnels and, to a limited extent, freeway-to-freeway connector movements in California, Minnesota, and Washington (*2*). Mainline metering examples discussed below include the San Francisco-Oakland Bay Bridge in Northern California, the Hampton Roads Bridge-Tunnel in Southeastern Virginia, the Baltimore Harbor Tunnel, and unregulated examples of mainline metering.

Bay Bridge

The Bay Bridge traffic management operation in Northern California is an example of mainline metering being used to increase downstream traffic volumes. The Bay Bridge is one of the few links across the San Francisco Bay that connects the cities of San Francisco and Oakland. During the a.m. peak period, three freeways converge onto the Bay Bridge to San Francisco. This traffic passes through a 22-lane toll plaza, of which several lanes are reserved for HOVs. Approximately 305 m (1000 ft) downstream of the toll plaza a metering bridge regulates the frequency of vehicles approaching the five lanes that cross the bay. HOV vehicles do not pay a toll and may travel through the metering area without stopping. Before the metering operation, downstream throughput on the bridge averaging approximately 8,200 to 8,300 vehicles per hour (vph). Implementation of the mainline metering resulted in downstream throughput on the bridge averaging 9,500 vph and sometimes even approaching 10,000 vph (McCrank, unpublished data). Thus, the Bay Bridge metering system increased downstream traffic volumes by approximately 15 percent.

In this example, the Bay Bridge serves as a bottleneck for commuters merging from three freeways desiring to cross the San Francisco Bay. Without mainline metering, the Bay Bridge experiences a drop in capacity. However, by managing the amount of traffic entering the bottleneck section, vehicle throughput on the bridge is increased.

Hampton Roads Bridge-Tunnel

The Hampton Roads Bridge-Tunnel is one of Southeastern Virginia's most important facilities, providing the only interstate link across the Hampton Roads Harbor. The combination bridge-tunnel-bridge connects the Hampton shore on the north to the Norfolk shore on the south. By July 1983, delays of up to 2 hrs were expe-

rienced, causing cars to overheat and increasing carbon monoxide (CO) levels within the tunnel. In August 1983, manually controlled mainline metering was initiated. This consisted of stopping traffic before the tunnel entrances when the vehicles within the tunnel slowed to 24.2 km/hr (15 mph) or less. When the tunnel was clear of traffic and the CO levels dropped, the traffic was released. In all cases, the vehicles that had been detained caught up with the vehicles that had not been detained before they reached the opposite shore. In effect, motorists who had been detained 5 to 8 min before entering the tunnel arrived at the same time that they would have if they had not been detained. Several benefits were derived from the mainline metering, including:

- Lower CO levels and less ventilation required;
- Lower tunnel temperatures and less stoppages caused by overheated vehicles;
- Free-flow traffic for longer periods with better throughput; and
- Traffic backups of shorter duration and length.

This form of mainline metering was found to be one of the most effective methods of managing the bridge-tunnel-bridge traffic during periods of heavy congestion (*3*). However, because of motorist complaints of being stopped before entering the tunnel, the manually controlled mainline metering operation was not continued.

Baltimore Harbor Tunnel

In the 1970s, the Department of Civil Engineering at the University of Maryland initiated a project entitled "The Study of Traffic Flow on a Restricted Facility." This study, sponsored by the Maryland State Highway Administration and the FHWA, utilized the Baltimore Harbor Tunnel as a test bed to analyze the concepts of traffic flow theory. One of the control strategies analyzed was the effects of a pretimed mainline metering system upstream of the entrance to the tunnel and downstream of the tunnel toll plaza. Traffic signals were located approximately 366 m (1200 ft) upstream of the tunnel portal, and pretimed metering scenarios for cycle lengths of 2, 3, and 4 min were evaluated. When metering was engaged, the red time varied between 7 and 10 sec and amber time was between 3 and 5 sec. The signals were only activated when traffic was congested from the tunnel portal to the merging area just downstream of the toll plaza. This corresponded to vehicular speeds of 32 to 40 km/hr (20 to 25 mph) as motorists passed the metering point. When the metering was not engaged, the signals were a continuous green.

This metering operation resulted in increased speeds within the tunnel bottleneck, in addition to an increased flow rate within the tunnel (*4*). Based on speed flow curves developed from the before and after metering operation, the study noted that the metering system had the potential to increase the capacity per lane by approximately 10 percent above the no-control condition. However, because of lack of political support, the mainline metering operation was not continued.

Unregulated Examples of Mainline Metering

Most commuters experience mainline metering without realizing it. Regulated mainline metering exists when overhead lane-use signals, similar to ramp meter signals, provide red and green indications to motorists. Mainline metering is also achieved in an unregulated fashion as a result of either a freeway incident or a reduction in freeway capacity (a lane drop).

Consider the case of an accident occurring on the freeway. Because of several factors, the accident results in a smaller number of vehicles traveling downstream. Motorists rubbernecking as they pass the accident, the accident itself blocking one or more lanes, and the need to stop traffic temporarily to allow emergency vehicles to access the accident scene are all factors contributing to the reduction in vehicular throughput. However, as motorists pass the accident scene, they find that downstream travel lanes are uncongested, enabling them to travel at free-flow conditions. The accident itself serves as a form of mainline meter. However, a traffic accident is an inefficient form of mainline meter that can overcompensate the desired effect. Accidents can severely restrict the number of vehicles that bypass the incident, resulting in unused downstream capacity that could otherwise be utilized more effectively through proper management.

Another form of unregulated mainline metering occurs where there is a reduction in downstream capacity. Consider a four-lane freeway that has a lane drop resulting in a three-lane facility. If the four-lane freeway is operating near capacity, the lane drop will serve as a bottleneck. As the freeway approaches capacity, a queue usually forms upstream of the bottleneck as vehicles maneuver into the three downstream travel lanes. Downstream of the lane drop, traffic conditions are usually better than upstream. The lane drop essentially functions as a mainline meter. However, as a result of vehicles maneuvering into the three lanes at the bottleneck location, downstream vehicular throughput is reduced to less than what could be achieved. This situation is analogous to the previous case studies. In each case there was a reduction in downstream capacity that limited the number of vehicles that could travel through the bottleneck. However, once a regulated mainline metering system was implemented, downstream vehicular throughput increased.

These examples of unregulated mainline metering were created because of a bottleneck condition. Even the previously discussed mainline metering experiences at the Bay Bridge, Hampton Roads Bridge-Tunnel, and Baltimore Harbor Tunnel were a result of bottleneck conditions at each location. The presence of a bottleneck creates the need to better manage the frequency of vehicles arriving at the bottleneck area. Without this management, experience has shown that there is a loss of downstream vehicular throughput and increased travel time as uncontrolled vehicles attempt to maneuver through the bottleneck location.

MAINLINE METERING RESEARCH OBJECTIVE

Existing freeway operations have demonstrated that, in the presence of a bottleneck condition or reduction in downstream capacity, regulating the number of vehicles through the bottleneck improves freeway operations. This research investigates whether regulating mainline vehicle movements can also provide improved freeway operations without the presence of a bottleneck.

During periods of heavy congestion, traffic density on the freeway can approach and exceed 96.6 vehicles per kilometers per lane (60 vehicles per mile per lane) (*5*). This results in the freeway operating in the unbalanced portion of the density-flow curve and a corresponding reduction in traffic flow efficiency. This traffic flow reduction, in the opinion of this author, can also be considered a reduction in capacity (Figure 1). In Condition 1, in which capacity exceeds demand, there is no congestion and traffic flows smoothly.

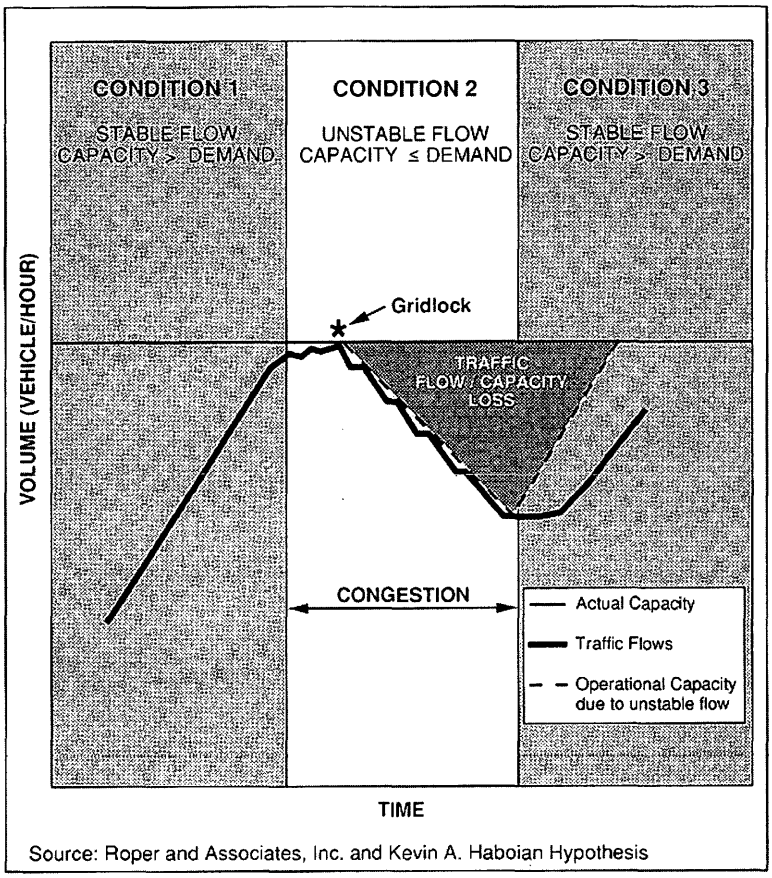


FIGURE 1 Traffic flow or capacity loss caused by congestion.

As demand builds to equal capacity, unstable flows are experienced and inefficient operation takes place. Finally, gridlock conditions occur and traffic flows fall sharply. This reduction in traffic flow can be referred to as the operational capacity caused by unstable flow. Under Condition 2, operation is unstable, inefficiencies continue to bring about a loss in capacity, flows drop off accordingly and congestion persists. This trend continues until a rebalancing of the capacity-demand relationship takes place. The operational capacity approaches the actual capacity, and stable flow with no congestion is restored (Condition 3).

A review of traffic on several freeways under varying congestion conditions in California indicated that freeway efficiency was reduced in some cases by as much as 50 percent as congestion set in, falling from a free-flow rate of 1,800 to 2,000 vehicles per hour per lane (vphl) to, in the most extreme case, a flow of approximately 1,000 vphl under stop-and-go operation (Roper, unpublished data). Traffic flow losses in the 25 to 30 percent range were not uncommon (*6*). If the conditions leading to gridlock could be avoided (i.e., the capacity-demand balance could be maintained), congestion would be minimized, and operational capacity caused by unstable flow would not materialize. In effect, the traffic flow/capacity loss would be preserved, or added back into the system, to serve greater traffic demands.

A typical freeway management response to prevent gridlock has been to institute ramp metering to regulate the entry of vehicles onto the facility. Ramp metering has been proven to be a very successful strategy and has been used to maintain freeway operations in the balanced portion of the density-flow curve. The primary reason behind the success of ramp metering is the dispersion of vehicle platoons entering the freeway. However, when freeway-to-freeway connectors remain uncontrolled, large traffic volumes have access to the freeway. Consequently, during peak travel periods, the freeway facility can again enter into the unbalanced portion of the density-flow curve.

Freeway connectors usually accommodate higher traffic volumes compared with typical freeway on-ramps; however, their basic operational characteristics are very similar. Both provide access to the freeway facility and usually have some amount of queue storage capability. As such, freeway connectors are capable of being metered similar to typical freeway on-ramps. There has been reluctance in certain areas to meter freeway-to-freeway connectors, largely because of difficulty in developing a constituency of supporters. Minneapolis, San Diego, Seattle, and Los Angeles have shown that connector metering can be both feasible and successful.

Assuming that both typical on-ramps and freeway connectors can be metered, the only input that remains uncontrolled is the freeway mainline. By metering the mainline to limit the number of vehicles that can be accommodated optimally given the downstream capacity and ramp volumes, travel delays traditionally incurred when vehicles are maneuvering into the heavily traveled freeway section may be eliminated. It is important to discern whether this mainline metering hypothesis can provide additional benefits over and above typical ramp metering. Because of the uncertainty and expense of field experimentation, traffic flow was modeled through computer

simulation using the INTRAS computer simulation model. A discussion of the simulation model, study area network, and research methodology and results is presented next.

Simulation Model

The INTRAS simulation model was written in 1977 by KLD Associates for the FHWA. It was selected as the analysis tool for this research because its car-following algorithms provide for a realistic simulation of traffic operations on an actual freeway. INTRAS is a microscopic model that simulates the flow of individual vehicles, as opposed to the macroscopic model that simulates the flow of a group of vehicles. On freeways with traffic demand below capacity, traffic flow is smooth and can be modeled reasonably well with the general parameters of a macroscopic model. However, in congested flow, traffic behavior becomes more complex. Because this research focuses on freeway conditions under congested flow, it is important that the vehicular behavior be modeled as accurately as possible. For these reasons, the INTRAS microscopic model was selected. For specific details of the INTRAS simulation model, consult the manuals cited in Reference *(7)*.

Study Area Network

A simple freeway network was established to evaluate appropriately the impacts associated with mainline metering. The network, illustrated in Figure 2, consisted of a three-lane freeway approximately 5.6 km (3.5 mi) in length, with one on-ramp located in the middle of the network. It is recognized that typical freeway segments will also have off-ramps and, in many areas, on-ramps may be located within a distance shorter than 2.8 km (1.75 mi). However, if it is found that mainline metering can provide improved traffic operations within this simple freeway network, the mainline metering strategy may be appropriate for a more typical freeway section with several on- and off-ramps. Conversely, if within the simple freeway network it cannot be shown that mainline metering provides any additional freeway traffic flow benefits, it will not be necessary to extend the research to a more typical freeway condition.

For simulation purposes, the freeway was divided into eight segments of 610 m (2,000 ft) each. Free-flow speed on the mainline was assumed to be 104.7 km/hr (65 mph), whereas free-flow speed for the on-ramp was assumed to be 88.6 km/hr (55 mph). The one-lane on-ramp also contained a 152.5-m (500-ft) auxiliary lane to facilitate vehicles merging from the on-ramp onto the freeway mainline. Although this freeway network is somewhat idealized, it is felt that the qualitative and quantitative results obtained from the simulation are analogous to a typical nonbottleneck freeway condition.

Research Methodology

The mainline metering evaluation was based on a variety of mainline volume and on-ramp control conditions. Previous experience has shown that the freeway mainline operates very well when mainline volumes are lower than 1,800 vphl. Consequently, mainline service volumes lower than 1,800 vphl were not evaluated. Specific mainline service volume rates analyzed included 1,800, 1,850, 1,900, and 1,950 vphl. Freeway lane capacity was defined as 2,000 vphl. Mainline volumes of 2,000 vphl were not analyzed because of limitations in the simulation model when analyzing service volumes equivalent to the capacity of a freeway lane. The vehicular demand on the on-ramp was kept constant at 1,200 vphl. However, the rate at which this on-ramp demand could access the freeway was simulated for the following conditions: no-control; ramp metering at 3 sec; ramp metering at 4 sec; and ramp metering at 5 sec. Ramp metering rates greater than 5 sec were not simulated because as the metering rate becomes more restrictive, the length of the on-ramp queue increases and can extend back to the arterial street, drawing objections from local jurisdictions.

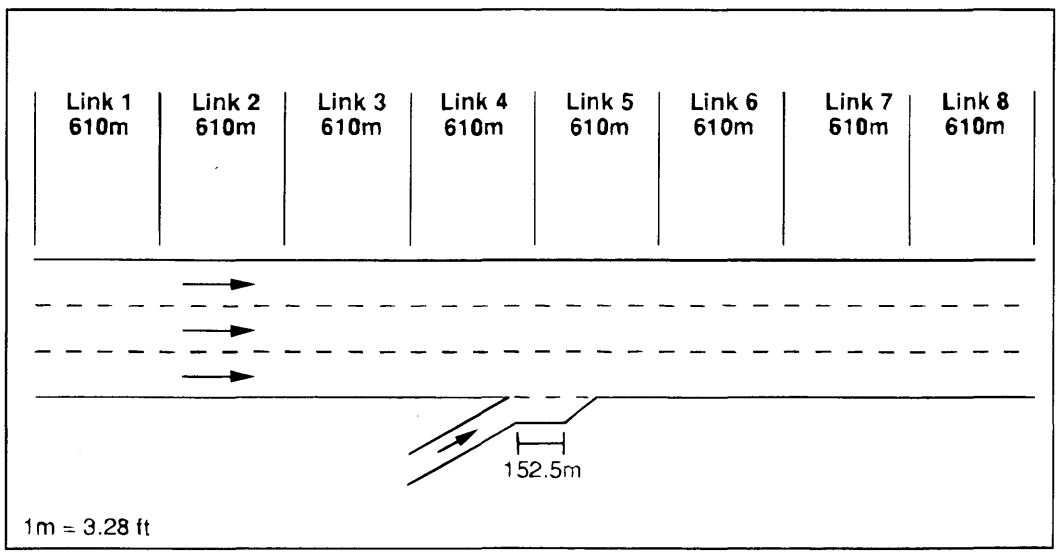


FIGURE 2 Study area network.

Because this evaluation was applied to an idealized freeway network and not actual field conditions, a simulation model calibration was not performed. For each mainline service volume condition, simulations were first performed for a base case assuming no form of control on either the on-ramp or mainline. Simulations were then performed using on-ramp metering rates of 3, 4, and 5 sec. Mainline metering was then simulated and evaluated for each of the three ramp metering conditions. In general, the mainline was metered at approximately 50 to 150 vphl less than the service volume demand. For example, if the service volume demand was 1,950 vphl, mainline metering was instituted to feed the downstream freeway at a rate of 1,900 or 1,800 vphl. Freeway conditions were simulated for a 30-min time period. Measures of effectiveness used to compare the results of these simulations included downstream vehicle throughput and downstream and overall speed.

It should be pointed out that this evaluation assumed that there would be no diversion as a result of either ramp or mainline metering. In actuality, there may be some diversion caused by both ramp and mainline metering, although to what extent is hard to predict. The purpose of this analysis is to determine the effect of mainline metering assuming the same level of demand (i.e., no diversion to local streets). If the simulation indicates that overall travel time is increased through the project area, then there is a strong likelihood that diversion would occur. On the other hand, if the simulation shows that overall travel time is reduced through mainline metering, then the propensity for diversion could be minimal. To appropriately evaluate the merits of mainline metering, the freeway demand is kept constant both with and without the on-ramp and mainline control strategies.

RESULTS

Vehicular Throughput

Figures 3 through 6 depict histograms of downstream vehicle throughput for the various combinations of mainline service volumes and on-ramp and mainline metering rates. It is important to realize that the combination of the mainline and on-ramp demands results in a total hourly demand of 6,600 to 7,050 vehicles downstream of the study area network. In addition, because traffic operations were simulated for a 30-min time period, the maximum downstream vehicular demand would range from 3,300 to 3,525 vehicles.

Figure 3 illustrates that for a mainline service volume of 1,800 vphl, the downstream vehicular throughput remains relatively unchanged for each of the control strategies simulated. Given that there are sufficient gaps in the mainline traffic stream for vehicles merging onto the freeway, implementation of ramp metering does not increase the downstream mainline throughput compared with the no-control scenario. Implementation of mainline metering results in slightly lower downstream traffic volumes compared with the ramp-meter-alone scenario. This result was expected, given that the network was able to effectively handle the ramp and mainline demands under the no-control scenario. With a service volume of 1,800 vphl, mainline metering would only serve to slow down vehicles, causing a reduction in downstream throughput.

Downstream throughput volumes for a mainline service volume of 1,850 vphl are shown in Figure 4. Noteworthy differences from the above results occur for the 5 sec ramp-meter-alone scenario and

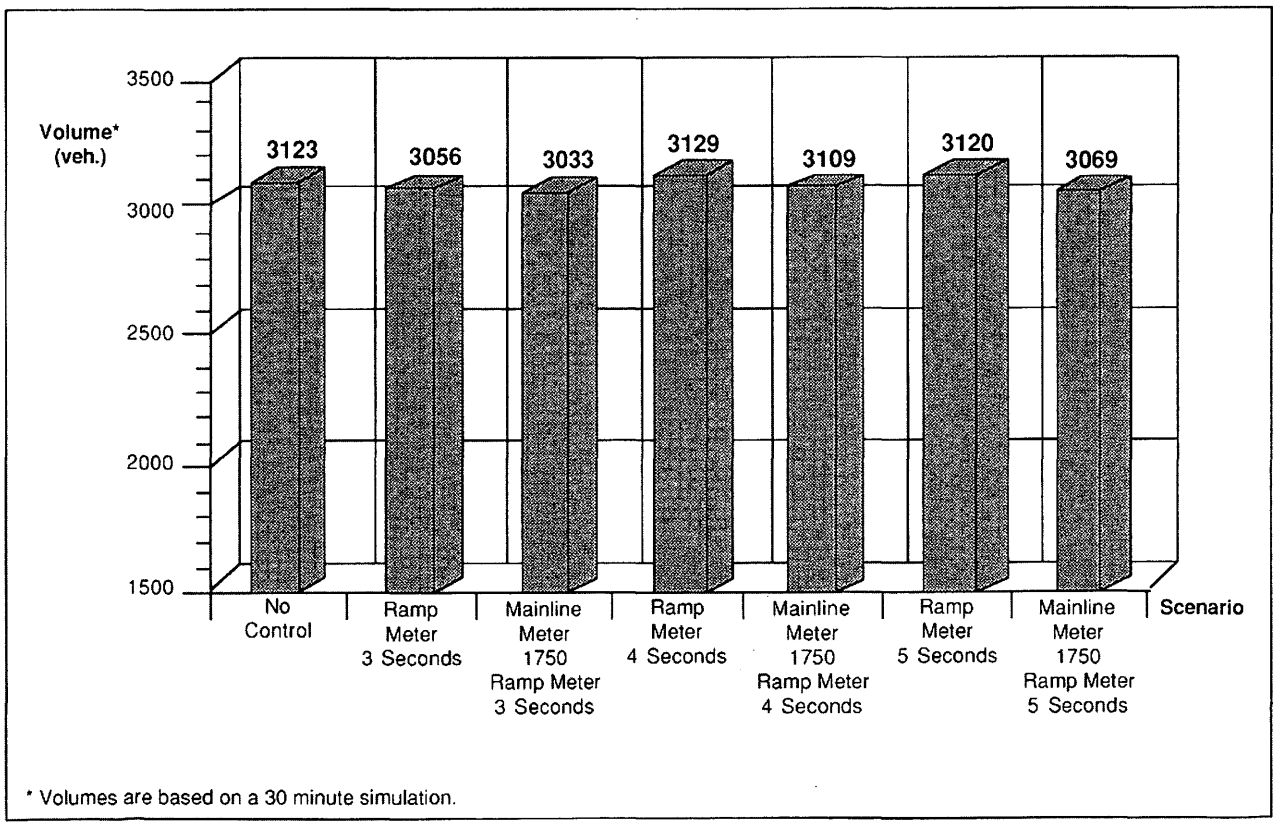


FIGURE 3 Downstream vehicle throughput for mainline service volume = 1,800 vphl.

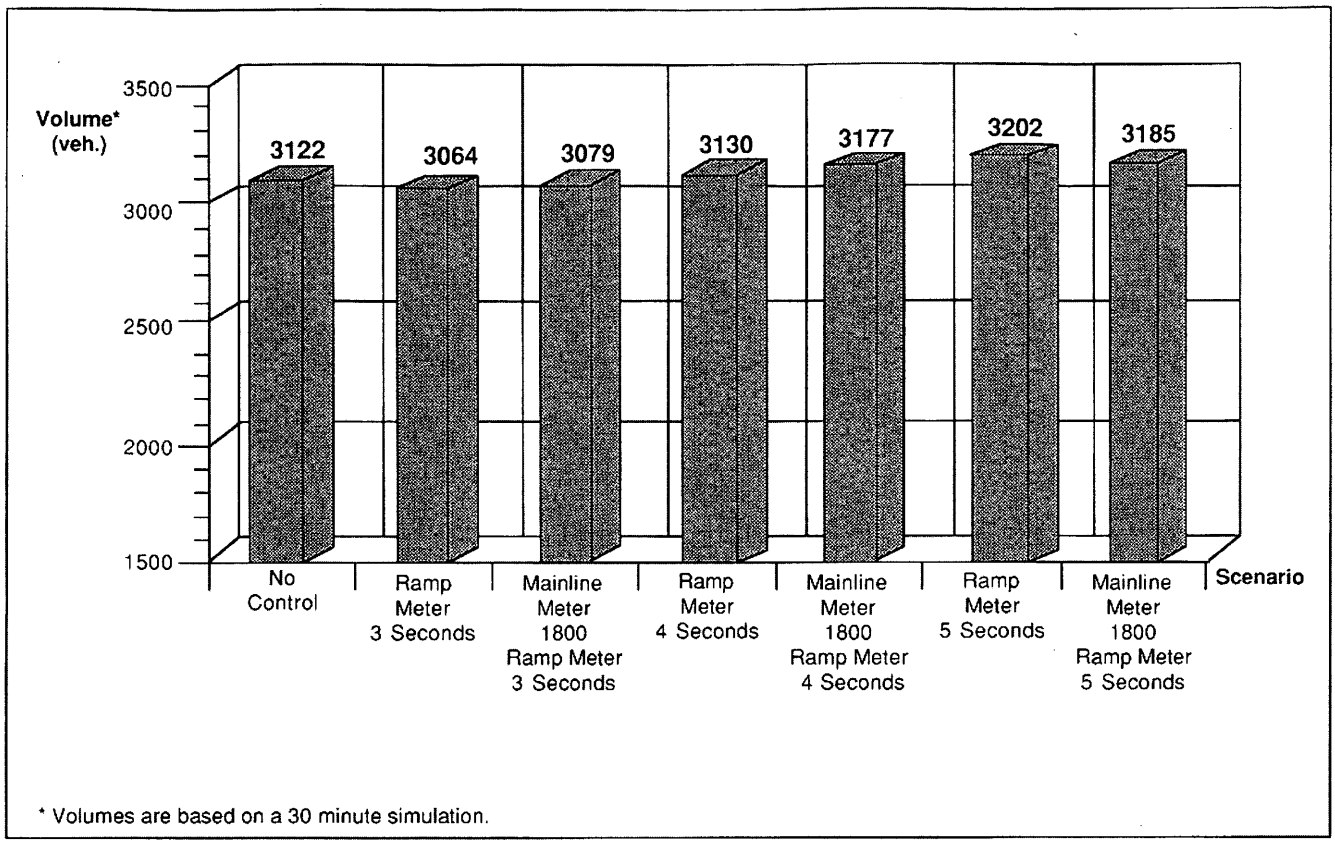


FIGURE 4 Downstream vehicle throughput for mainline service volume = 1,850 vphl.

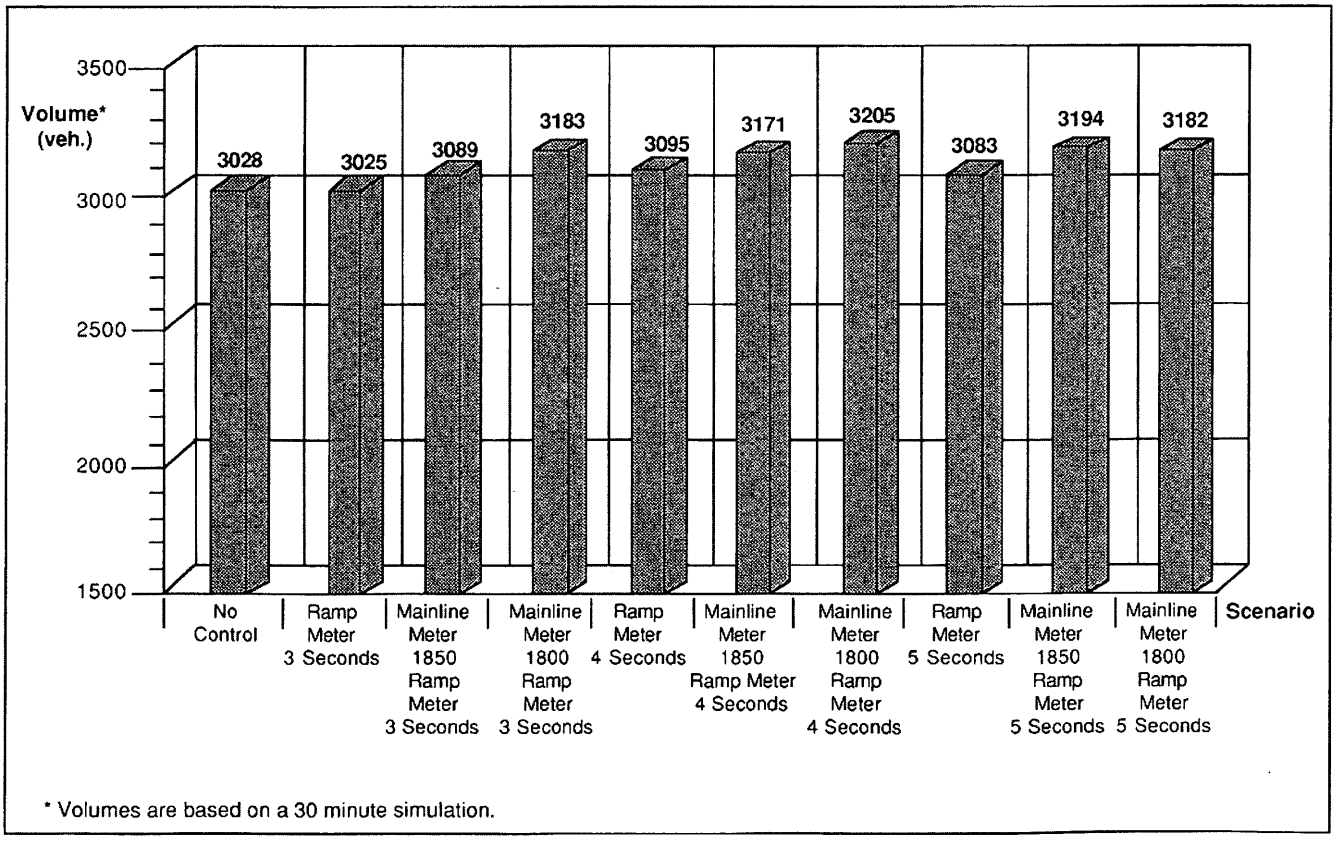


FIGURE 5 Downstream vehicle throughput for mainline service volume = 1,900 vphl.

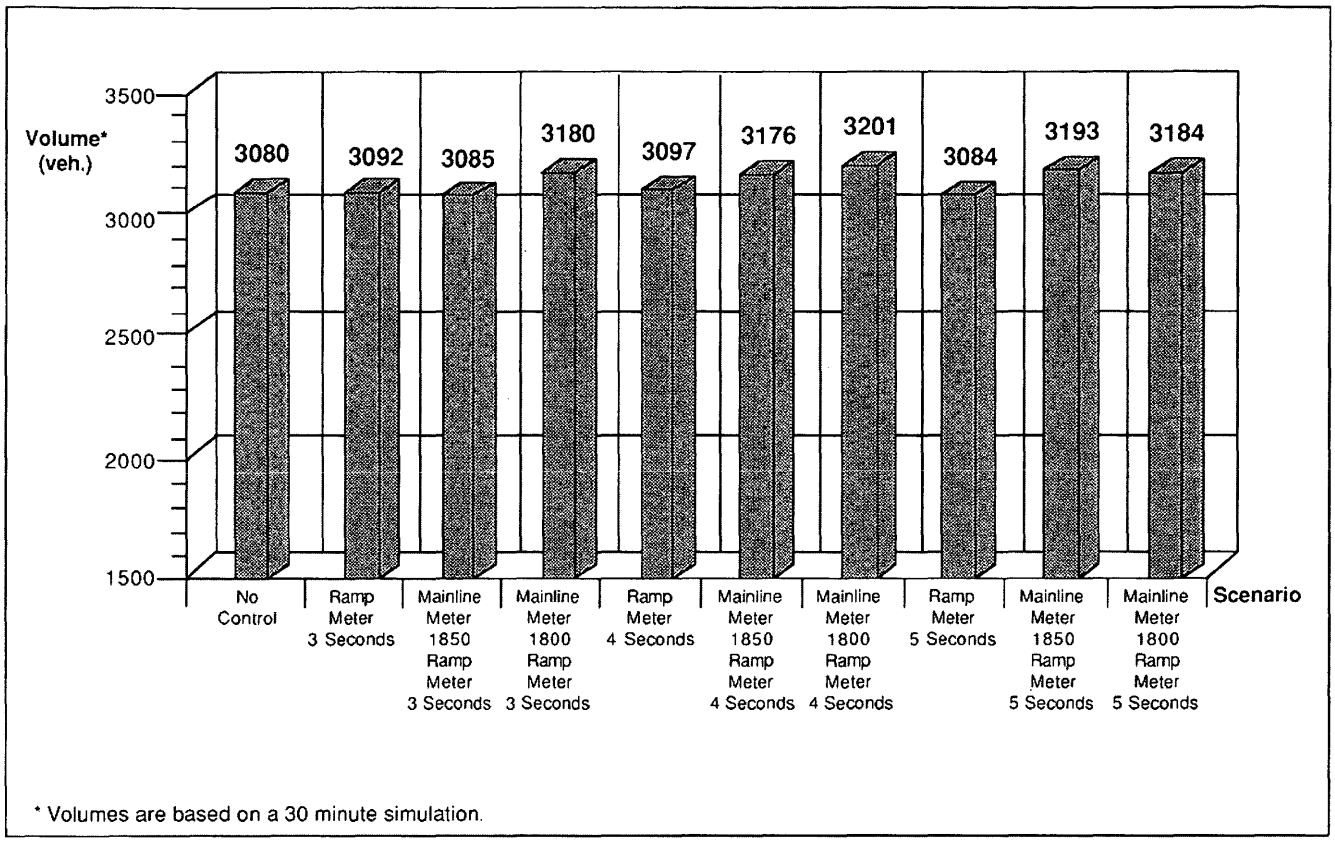


FIGURE 6 Downstream vehicle throughput for mainline service volume = 1,950 vphl.

the combined 4- and 5-sec ramp metering with mainline metering scenarios. These scenarios, which are the last three entries of Figure 4, resulted in an increase (approximately 2.5 percent) in downstream throughput compared with the no-control scenario.

Figures 5 and 6 present the downstream traffic volume results for mainline service volumes of 1,900 and 1,950 vphl, respectively. A review of the ramp-meter-alone scenario indicates the downstream traffic volumes generally increased, in some cases up to 3 or 4 percent, compared with the no-control scenarios. It appears that uncontrolled on-ramp traffic can cause a slight reduction in the downstream traffic volume because of the increase in mainline demand. Appropriately regulating on-ramp demand appears to eliminate the platooning of entering vehicles that can cause the reduction in downstream mainline volumes. When mainline metering is combined with on-ramp metering, the downstream traffic volume appears to increase an additional 3 percent compared with the ramp-meter-only scenarios.

Based on these results, it appears, for the nonbottleneck condition, that as freeway traffic volumes approach the capacity of the facility, mainline metering can increase downstream vehicular throughput above what ramp metering alone can accomplish.

Average Travel Speed

The average speed for each traffic control scenario and mainline service volume are indicated in Figures 7 through 10. In each figure, average speed downstream of the mainline meter and average speed for the network (which includes vehicles entering from the on-ramp and those upstream of the mainline meter) are indicated for each control strategy. For the no-control and ramp-meter-alone scenarios, the average speeds downstream of the mainline meter are indicated for comparative purposes and reflect the average speed downstream of where the mainline meter would have been.

Figures 7 through 10 indicate that for the ramp-meter-alone scenarios, the network speed and speed downstream of the mainline meter increase slightly as the ramp metering rate becomes more restrictive. With the implementation of mainline metering, the speeds downstream of the mainline meter are approximately 10 percent higher for the combined mainline and ramp meter scenarios when compared with the ramp-meter-alone scenarios. For these same scenarios, the overall network travel speeds are approximately the same or, in some cases, higher. The previous results appear to indicate that mainline metering can increase freeway speed downstream of the mainline meter, and the delays incurred upstream of the metering point do not result in any overall travel time increases for the nonbottleneck condition.

Also noteworthy is that as mainline service volumes increase, mainline metering offers larger speed increases downstream of the metering point compared with the no-control and ramp-meter-alone scenarios. These increases, illustrated in Table 1, are accomplished without increasing overall travel time.

Summary of Results

The purpose of the simulation was to determine whether there are any freeway operational benefits, above those achieved through ramp metering, resulting from metering the freeway mainline for

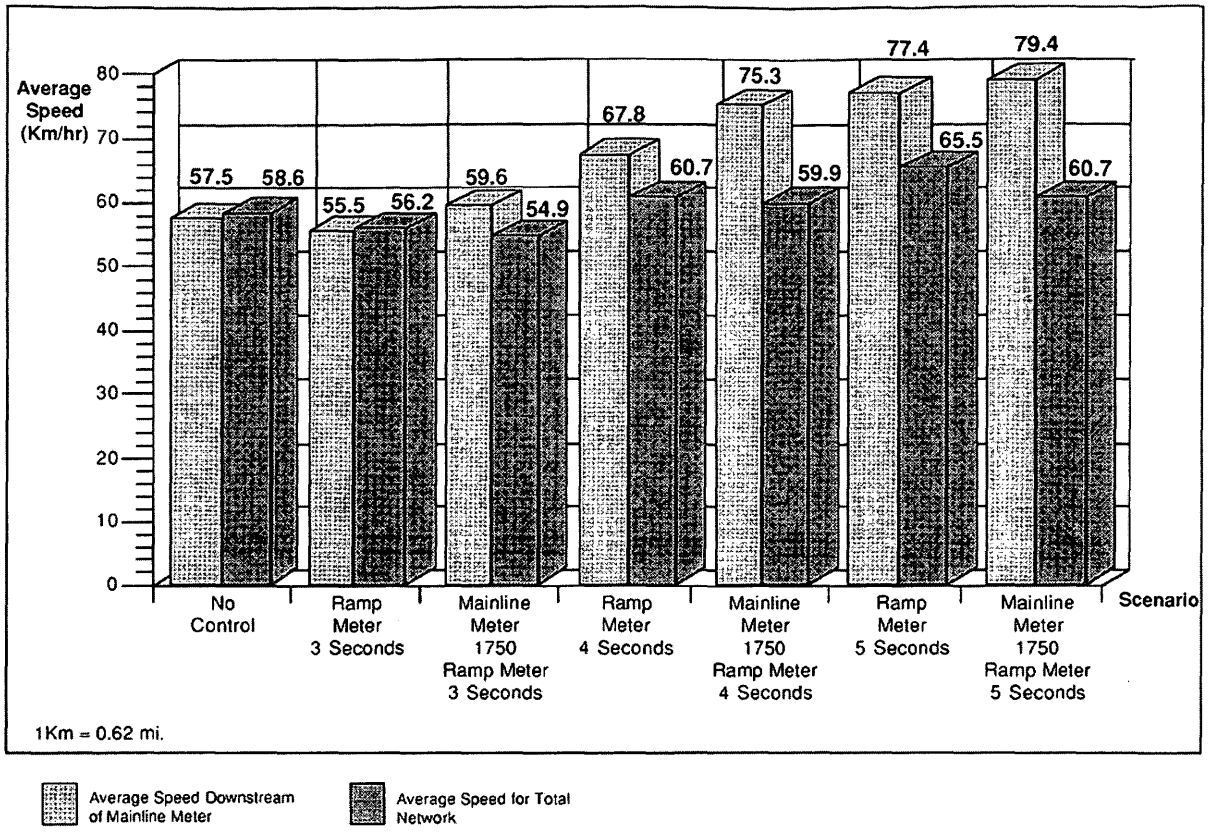


FIGURE 7 Average speed for mainline service volume = 1,800 vphl.

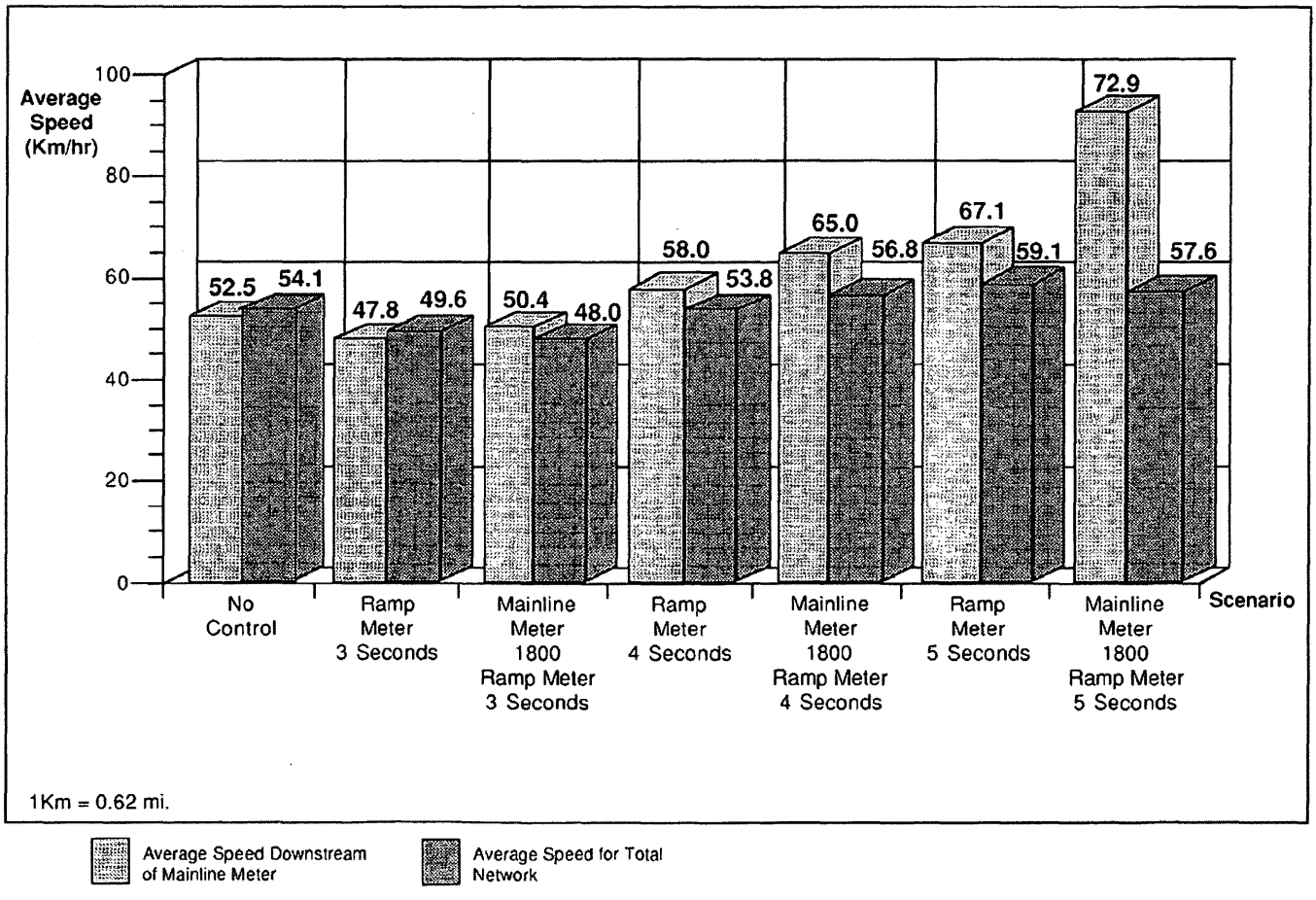


FIGURE 8 Average speed for mainline service volume = 1,850 vphl.

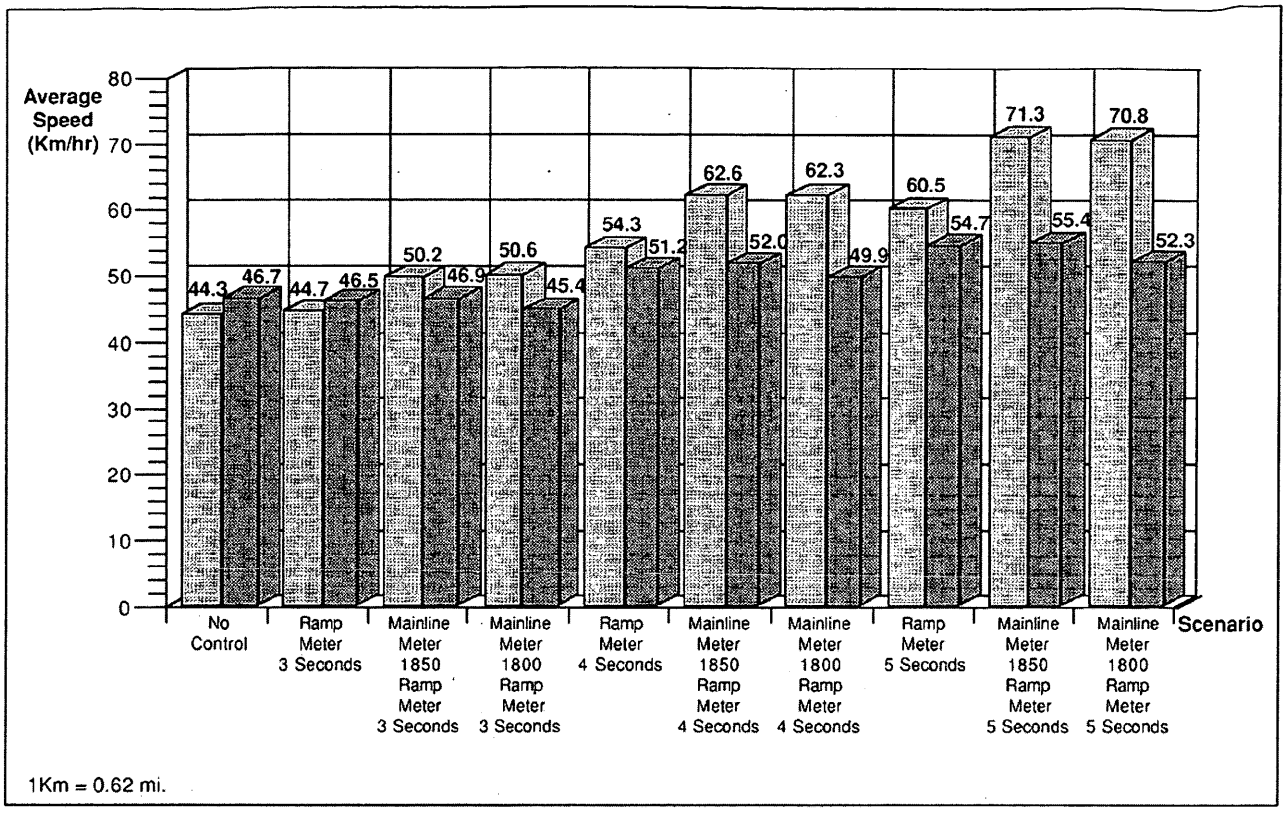


FIGURE 9 Average speed for mainline service volume = 1,900 vphl.

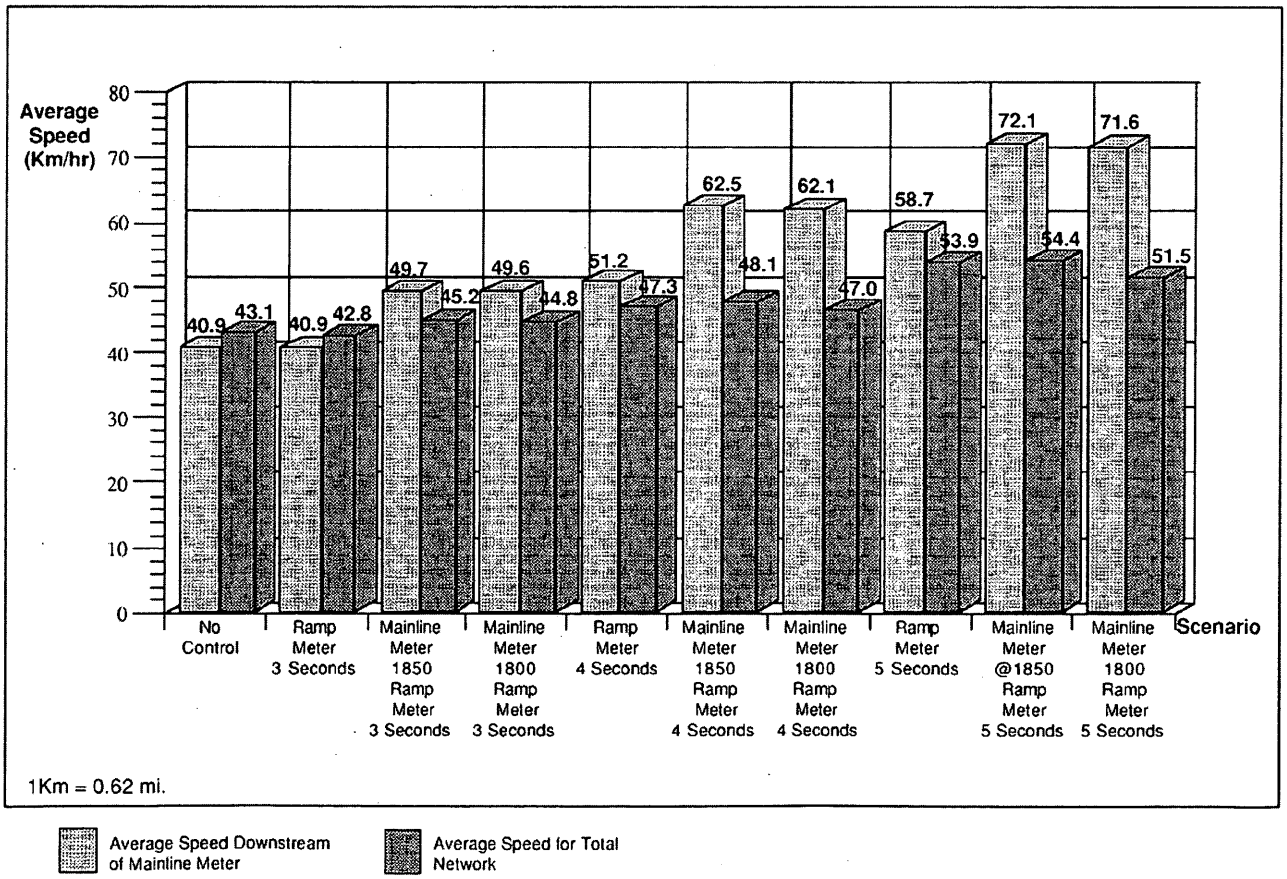


FIGURE 10 Average speed for mainline service volume = 1,950 vphl.

TABLE 1 Percent Increase in Freeway Speed Downstream of Mainline Meter*

Scenario	Mainline Service Volume (vphpl) 1,900	1,950
Ramp Meter 4 Seconds		
Mainline Meter @ 1,850 vphpl	15%	22%
Mainline Meter @ 1,850 vphpl	15%	21%
Ramp Meter 5 Seconds		
Mainline Meter @ 1,850 vphpl	18%	23%
Mainline Meter @ 1,800 vphpl	17%	22%

** Compared to Ramp Metering alone Scenario*
(vphpl) = vehicles per hour per lane

the nonbottleneck condition. Several observations drawn from the previous results include the following:

- As the freeway mainline volume increases, ramp metering appears to increase the downstream vehicular throughput by approximately 2 to 4 percent. This result is consistent with existing experiences with ramp metering to date.
- Ramp metering appears to increase the freeway speed and the average speed for the entire freeway network compared with the no-control scenario. In addition, the more restrictive the metering rate, the better level of operation on the freeway (in terms of higher speeds and lower travel times).
- Combining mainline metering with ramp metering resulted in the same or, in several cases, slightly higher downstream vehicular throughput compared with the no-control and ramp-meter-alone scenarios.
- As mainline service volumes increase, the freeway speed for the ramp-meter-alone scenario decreases. The addition of mainline metering provides much improved freeway conditions downstream of the mainline meter. This improvement is balanced by the added travel time incurred upstream of the mainline meter.
- When mainline metering is combined with ramp metering, the average freeway speed for the total network increases compared with the ramp-meter-alone scenarios. Figures 8 and 9 indicate that this speed increase can be as high as 8 to 11 km/hr (5 to 7 mph).
- Vehicles originating from on-ramps downstream of the mainline meter are provided better freeway conditions compared with the no-control and ramp-meter-alone scenarios. Consequently, vehicles originating from these downstream on-ramps experience lower freeway travel times.

IMPLICATIONS OF RESULTS

These results appear to indicate that mainline metering can provide improved freeway operations downstream of the metering point compared with ramp metering alone. For mainline service volumes of 1,950 vphl, the average freeway speed downstream of the mainline meter was 22 percent higher than the ramp-meter-alone scenario. But, most important, the simulation indicates that this can be accomplished while decreasing freeway travel time. This is an important observation, considering that one of the concerns regarding mainline metering has been the perception of additional delay incurred by vehicles waiting in a mainline queue upstream of the mainline meter. From these results, it appears that this delay is balanced by improved freeway operations downstream of the mainline meter. Based on metering experiences in San Diego (*8*), commuters are willing to wait in a queue (in San Diego's case, it is more of a rolling queue) provided they perceive an improvement in their trip farther downstream.

The results are also important when considering the equity issue. One of the historical arguments against ramp metering is that vehicles originating from entry points closer to their destination incur a greater travel time delay than vehicles originating farther out in the suburbs. Based on this research effort, it appears that mainline metering may not increase the overall travel time through the freeway network, and may be used to distribute more equitably the delay incurred by vehicles ingressing the freeway. For example, consider the morning commute of a line-haul freeway approaching the outskirts of a metropolitan area. A mainline meter may eliminate the need to meter on-ramps upstream of the mainline meter location. The delay experienced by commuters using this particular freeway could be distributed equitably between the mainline meter and metered on-ramps downstream of the mainline meter. In addition, mainline metering can promote ridesharing and HOV travel time savings. Similar to HOV operations at the Bay Bridge, HOVs could gain travel time benefits over single occupant vehicles via HOV bypass lanes upstream of the metering point.

This research appears to indicate that, depending on the goals and objectives of the freeway operating agency and given the right conditions, mainline metering is an appropriate freeway management tool.

REFERENCES

1. Lerner-Lam, E. *Mobility Facts.* The Institute of Transportation Engineers, Washington, D.C., 1992.
2. Jacobson, E. L., and J. Landsman. Case Studies of Freeway to Freeway Ramp and Mainline Metering in the U.S., and Suggested Policies for Washington State. Presented at 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., 1994.
3. Harrison, J. E. Hampton Roads Bridge Tunnel Traffic Control and Surveillance System. Presented at 66th Annual Meeting of the Transportation Research Board, Washington, D.C., 1987.
4. Carter, E. C., and R. C. Loutzenheiser. *Study of Traffic Flow on a Restricted Facility.* Report FHWA-MD-R-77-9. FHWA, U.S. Department of Transportation, 1977.
5. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
6. *Traffic Engineering Handbook,* 4th ed. Institute of Transportation Engineers, Washington, D.C. 1992.
7. KLD Associates. *Development and Testing of INTRAS: A Microscopic Freeway Simulation Model,* Vols. 1,2,3, and 4. Federal Highway Administration, Washington, D.C., 1977.
8. Haboian, K. A. *Freeway Management Strategies.* Parsons Brinckerhoff Inc., New York, 1993.

Publication of this paper sponsored by Committee on Freeway Operations.

Development of a Freeway Congestion Index Using an Instrumented Vehicle

GLEN S. THURGOOD

The purpose of this study, funded by the Utah Department of Transportation, was to produce a freeway congestion index (FCI). The data required for the FCI can be developed with or without the benefit of automated traffic surveillance or data collection systems. A 9.7-km (6-mi) segment of I-15 in the Salt Lake City metropolitan area was used to test the viability of the FCI. The FCI reflects both the extent (length) and duration of congestion on a given freeway segment and can be used to compare congestion levels on different freeway segments or subsystems, and to compare congestion levels on freeway systems of differing sizes. It can also be used to compare changes in the level of congestion as they occur over time (from year to year or between different seasons of the year). Speed was used as the indicator of congestion onset, with acquisition of the needed data for calculation of the FCI being done using an instrumented moving vehicle. It was also found that measurements taken in a single lane can be used to accurately determine the FCI for all lanes of a six-lane freeway.

Current federal legislation (the Intermodal Surface Transportation Efficiency Act) requires that urbanized areas of the United States implement congestion management systems. To implement such systems it is necessary to settle on a definition of congestion and arrive at an acceptable and repeatable means of measuring it. In this study is an outline of one method of measuring freeway congestion that can be employed regardless of whether the freeway has extensive instrumentation for monitoring traffic flow parameters.

Traffic congestion generally can be described as the operating conditions that exist on any roadway at any point in time when the quality of traffic flow (as measured by parameters such as travel time, speed, delay, etc.) deteriorates below a level acceptable to the user. Traffic congestion on urban freeways generally can be categorized into two types: recurring and nonrecurring. Nonrecurring or incident-based congestion is the result of some planned or unplanned event (e.g., a maintenance operation or traffic accident) that temporarily causes a significant reduction in the capacity of any transportation system. It may be as severe as a total closure.

Although any transportation facility may experience congestion at any time due to an incident, recurring congestion is the type that occurs repeatedly and is time-predictable as to its onset, extent, and duration. It is simply the result of demand exceeding the capacity of some point or section of the freeway, which creates a bottleneck. It could even be referred to as the "background level of congestion" or that level of congestion that could be expected to occur regularly on a given day at a specified location. On Utah freeways, the bottlenecks often occur in weaving sections that are too short or otherwise inadequate, in merge areas downstream of on-ramp noses, at the intersection of off-ramps and arterial crossroads causing exiting traffic to back up and obstruct flow on the freeway main lanes, or as a result of some combination of these three problems.

As a result of its time-predictable, repetitive nature, recurring congestion is easier to deal with than nonrecurring congestion. When a freeway is operating near its capacity, traffic flow becomes unstable and even slight surges in traffic demand will cause congestion to occur and travel speeds to diminish. As long as traffic demand exceeds the available capacity, forced flow [level of service (LOS) F] will occur. Freeway traffic congestion manifests itself in severely restricted speeds and the development of long, slowly moving queues in which stop-and-go driving may occur on the freeway, and long queues and delay may develop on the access system.

The urbanized areas along the Wasatch Front in Utah, from Provo on the south to Ogden on the north are no exception to this generality. Although the duration and severity of congestion on Utah freeways may not be as extensive as in other urban areas, it is nonetheless a major concern to the citizenry and public officials.

The primary objective of this study was to develop and test an index that describes both the extent and duration of freeway congestion. The method of measurement was to be cost effective, to be repeatable, and able to be implemented using equipment and skill levels presently available in the Utah Department of Transportation (UDOT). The index needed to be capable of reflecting changes in congestion levels over time and between segments and systems without the benefit of extensive automated data collection.

CONGESTION INDEX

When Does Congestion Begin?

The logical first step in developing a method for measurement of level of congestion on freeways is to reach agreement on some value or condition that describes when congestion begins. This can be very difficult because congestion is as much subjective (qualitative) as it is quantitative. Although congestion is a commonly occurring phenomenon, there is no commonality of definition as to what level of degradation in the quality of traffic flow constitutes congestion. To make quantitative comparisons between congestion levels at different locations, we must settle on a definition of what constitutes congestion and when it begins. The specific definition may vary according to such variables as type of facility, functional classification, and location.

Speed as an Indicator of Congestion

After extensive investigation of the problem of congestion definition and quantification, for this study (which was limited to measurement of recurring congestion on freeways) an onset-of-

Civil & Environmental Engineering Department, Brigham Young University, Provo, Utah 84602.

congestion definition based on the LOS dropping from E to F (breakpoint, E/F) as determined by speed measurement was selected for the following reasons:

1. A significant reduction in speed below that normally expected or desired is an operational parameter to which drivers relate. When a significant speed reduction is encountered during travel, the driver knows that his travel time is going to be increased and he will be delayed in reaching his desired destination if operation at the reduced speed persists significantly in time and in distance.
2. Speed is a traffic parameter that can be measured rather easily, at relatively low cost, using a variety of devices and methods.
3. Speed is the parameter second-most preferred (24 percent) for use by most agencies in measuring congestion. Delay is the most favored (31 percent). The measurement actually used most often, at present, is LOS (90 percent) (*1*).

Although speed reduction was the parameter used in this study for defining congestion onset, the freeway congestion index (FCI) as developed in this paper is flexible enough to accept other definitions and parameters.

For many years the characteristic speed-flow (volume) relationship for freeways was generally accepted as being similar to that shown in Figures 3 and 4 of Chapter 3, Basic Freeway Sections, of the *Highway Capacity Manual* (HCM) (*2*). Examination of the curve shows a gradual decline in speed as flow increases, with a progressively increasing rate of change of speed as capacity is approached. Speed at capacity (LOS E/F breakpoint) was generally believed to be around 56 kph (35 mph). In January 1995, the 1994 updates to the HCM, including a revised Chapter 3, were released by the TRB. Included were new speed versus flow curves for basic freeway sections. Examination of these curves shows that there is relatively little deterioration in speed from the free-flow speed as traffic flows increase. As capacity is approached, only about a 16-kph (10-mph) decrease in speed to 80 kph (50 mph) is experienced (for a freeway having a 100-kph (60-mph) free-flow speed) before reaching the LOS E/F breakpoint and dropping into LOS F, in which flow is forced and speeds substantially decrease. The new, higher LOS E/F breakpoint speeds no doubt are a reflection of the more aggressive behavior of present-day drivers.

The new maximum densities at the LOS E/F breakpoint are 36.7 to 47.9 passenger cars per mile per lane, depending on the free-flow speed of the facility. These density values are considerably less than the 67 passenger cars per mile lane density given for the LOS E/F breakpoint in the present (1985) HCM. In summary, the revised procedures of the HCM seem to indicate that on freeways the LOS E/F breakpoint seems to occur at significantly higher speeds and lower densities than previously believed.

For purposes of this research, it was decided, in concert with the study advisory panel, that the onset of congestion on Utah freeways would be based on traffic stream speeds falling below a threshold speed of 64 kph (40 mph). This is higher than the old breakpoint speed of 56 kph (35 mph) but lower than the new values of 80 kph (50 mph). This decision was based on the premise that a traffic stream speed of 64 kph (40 mph) is a strong indicator that flow is falling into the LOS F (forced-flow) realm. The speed profile studies performed as part of this study seem to verify this perception. A threshold speed of 72 kph (45 mph) or even 80 kph (50 mph) would not likely have changed the results significantly because, in most instances, once speeds fell below 80 kph (50 mph), they also fell below 64 kph (40 mph).

Use of density as the parameter of choice to determine the onset of congestion was considered but was rejected because of the difficulty and cost of directly measuring density. Aerial photography is about the only reliable way of directly measuring density; however, this type of data collection is expensive and is time-consuming to reduce. Some density measurement using oblique photography was done as a part of this study, with densities in mixed traffic of approximately 50 to 75 vehicles per mile per lane being measured in periods identified as being the onset of congestion.

Development of a Congestion Index

The primary objective of this study was to develop and test an index for quantifying recurring congestion that reflects both its extent (length) and duration. Cottrell (*3*) presented the idea of a lane-mile duration index (LMDI), which came close to providing an index that met these objectives.

$$\text{LMDI}_f = \sum_{i=1}^{m} \text{LM}_i \times D_i \tag{1}$$

where

i = a two-way freeway segment of uniform capacity, generally between two adjacent access points;
m = the total number of freeway segments in a given urban freeway system;
LM_i = the total lane-miles in freeway segment i; and
D_i = the duration of LOS F congestion, in hours, on i.

In Cottrell's calculation of the LMDI, traffic volumes (annual average daily traffic) from the Highway Performance Monitoring System data base were used as a basis for determining whether a two-way freeway segment of uniform capacity would be expected to experience LOS F congestion during the day and for how long. If any portion of the segment was congested, it was assumed that the entire two-way segment was congested. The LMDI also makes no provision for comparing segments or systems of significantly differing sizes (i.e., lane-miles).

If Equation 1 is normalized by dividing by the number of lane-miles in a freeway segment, then an index is provided that has the units of lane-mile-hours per lane-mile. This enables a direct comparison of the extent (length) and duration of congestion on different freeway segments having differing lengths (i.e., long versus short segments). It can also be used to reflect the level of congestion on the freeway system in an entire geographical area, such as an urbanized area, and compare it with the system in another urbanized area even though the areas may be considerably different in size (i.e., lane-miles of freeway).

The FCI has been developed to measure, in a quantitative manner, the severity of recurring congestion on Utah freeways. Its mathematical expression is given by Equation 2.

$$\text{FCI} = \sum_{i=1}^{n} \left(\frac{\text{CLM}_i \times D_i}{\text{LM}_i} \right) \tag{2}$$

where:

FCI = Freeway Congestion Index (lane-mile-hours per lane-mile), usually computed per day or per average weekday (AWD);

i = a one-way freeway segment, the length of which is determined by the responsible agency as desired;
n = the total number of freeway segments in a given urban freeway system, or a defined subsystem;
CLM_i = total congested lane-miles in freeway segment i operating at LOS F congestion [e.g., < 64 kph (40 mph)];
D_i = duration of LOS F congestion, in hours, on freeway segment i; and
LM_i = total lane miles in freeway segment i;

Although Equations 1 and 2 appear quite similar, there are some significant differences. First, the segment lengths are defined differently. In the LMDI equation, segment length is for a two-way segment, usually limited in length to the distance between two access points. In the FCI, the segment is directional and its length may be defined as the user desires. Second, the LMDI equation assumes congestion based on two-way volumes and a calculated LOS for the entire link. The FCI equation uses field-measured values for duration and length of congestion, and only that portion of the link that is congested is reported in the calculation. The FCI has the advantage of allowing the summation of multiple segment or lane FCI values. This allows several segments to be grouped together so that system or regional comparisons can be made.

Determining Extent and Duration of Congestion

Any suitable method for determining the time of onset of congestion, how long it lasts, and the number of congested lane-miles with reasonable accuracy can be used to provide the needed inputs for determining the FCI. The one described here is operational below a prescribed speed, but a density criteria could be used as well.

The development of congestion during peak traffic periods is a dynamic process, with the length of the congested area changing from minute-to-minute. As traffic demand volumes begin to approach bottleneck capacity, vehicle speeds decrease and a queue begins to form. As demand continues to increase, the queue is propagated upstream and the congested area lengthens. Initially, only one lane may be affected, but congestion soon spreads to adjacent lanes as drivers shift lanes to avoid the congestion and maintain a higher speed. As long as the vehicle arrival rate at the back of the queue exceeds the departure rate from the front of the queue, the length of the congested area will continue to increase. Once the arrival rate falls below the departure rate, the length of the congested area will begin to decrease until congestion has dissipated.

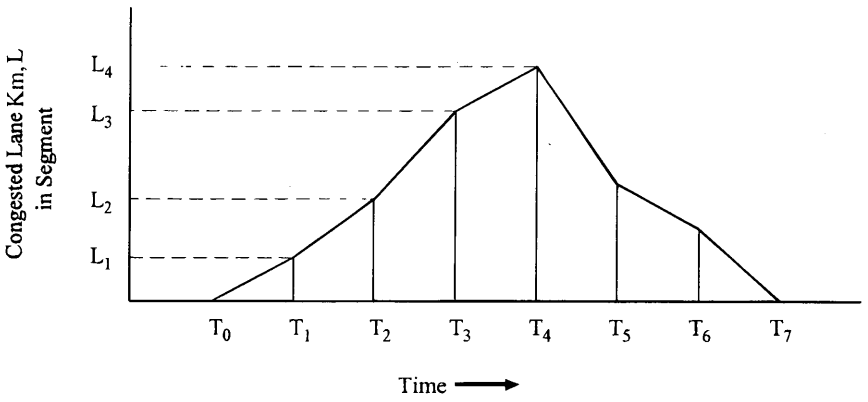
To quantify the extent (length) of the congested area, a sampling process is needed because the extent of congestion will change with time. This process is illustrated in Figure 1. At some time (T_0), congestion, as defined, does not exist but is just beginning to develop. As yet, no major speed reduction has occurred, but congestion is impending. T_0 is determined by the time of the last sample taken during which traffic stream speeds at no point fall below the threshold value of 64 km/hr (40 mph). At some later sampling time (T_1), a slow-moving queue of length (L_1) exists. The length of this queue could also be detected using aerial photography and employing a density criterion for defining the onset of congestion and the length of congested roadway. By periodic sampling (e.g., aerial photography of a given freeway segment at uniform time intervals) a curve such as that in Figure 1 could be developed for a given freeway segment. This can also be done by sampling traffic stream speeds using an instrumented probe vehicle periodically traveling the study segment, as was done in this study, or using some other speed measurement technique. Although the shape of the curve between points of measurement is almost certainly not strictly linear, if the sampling interval ($T_{j+1} - T_j$) is kept short relative to the length of the congested period, a reasonable approximation to the true shape of the curve may be obtained.

The area under this curve represents the product of duration of congestion and length of the congested area ($CLM_i \times D_i$) as required for computation of the FCI. The computational process can, of course, be done stepwise using the relationship

$$CLM_i \times D_i = \sum_{j=1}^{m} \frac{(L_{j+1} - L_j)}{2} \times (T_{j+1} - T_j) \qquad (3)$$

where

j = successive observations of the congested area, 1 to m;
T_0 = the latest observation (time) during which no congestion is detected;



T_0 = the latest sampling time during which no congestion is detected

T_j = the time at which L_j lane miles of congestion is measured

FIGURE 1 Development of congestion on a freeway segment over time.

T_j = the observation time at which L_j lane-miles of congestion is measured;
L_j = number of congested lane-miles measured at time j, and CLM_i and D_i are as previously defined.

This becomes the computation of the area of successive trapezoids in Figure 1. Once this computation has been done for a selected segment (i), then the FCI for that segment is computed by dividing the sum arrived at using Equation 3 by the total number of lane-miles in that segment (LM_i).

For the calculation of the FCI for a larger system, such as the freeway system in a given urban area (or a defined subsystem), the system could be subdivided into logical segments, the product of the number of congested lane-miles and duration for each segment determined as described above, these products summed, and the total divided by the number of lane-miles in the system (or subsystem) to yield an FCI for that system (or subsystem). The FCI would usually be computed for each weekday and averaged for an AWD-FCI. The computed FCI for that system could then be compared with that computed for another system to give a relative value for the level of congestion between the two. Comparison of changes in the level of congestion over time in a particular system could be accomplished by determining and comparing the FCI from year-to-year or season-to-season.

PILOT STUDY

To investigate the applicability of the FCI, a pilot study was performed on a 9.7-km (6-mi) segment of I-15 in the Salt Lake City metropolitan area. Speed, distance, and travel time data were collected during the morning and evening peak periods, in both the northbound (a.m. peak period) and southbound (p.m. peak period) directions, for 1 week during August 1993, supplemented with additional observations during October 1993.

To calculate the FCI, it is necessary to determine the time of onset, duration, and extent (length) of congestion. The onset of congestion was defined as the time when traffic speeds at any point within the study section dropped below some threshold value, in this case 64 kph (40 mph). The duration of congestion was defined as the time from the onset of congestion to the time when traffic speeds within the segment no longer fell below the threshold value. The extent of congestion was defined as the distance between the points where the speed of the traffic stream dropped below, and then went back above, the threshold value. Thus, the measurement method selected would need to track speeds versus distance along the segment, as well as keep track of the times of onset and dissipation of congestion, and times the length of congestion was measured. Several congested subsections could exist within the study segment.

Study Segment Description

The Salt Lake City metropolitan area lies at the crossroads of two Interstate highways, I-80 and I-15. In addition to being the dominant north-south through-traffic carrier, I-15 is the major commuter route serving traffic traveling to and from the central business district (CBD) of Salt Lake City and suburban communities to the north and the south of the city. East-west traffic is carried to I-15 via perpendicular arterial cross streets. This particular segment of I-15, located south of the CBD, is a six-lane facility (three lanes per direction) and is inside the I-215 loop.

In the northbound, a.m. peak direction, recurring congestion occurs beginning at the merge areas of on-ramps from I-215 and from interchanges at 5300 South, 4500 South, and 3300 South. The latter three of these are ramps from arterial cross roads having compressed diamond interchange configurations with two-way service roads, and they exhibit similar congestion characteristics. The former consists of a single-lane loop on-ramp from eastbound I-215 followed by a two-lane on-ramp from westbound I-215 with the two ramp lanes merging into the same outside through-lane of I-15. This particular geometric configuration is a violation of the AASHTO lane-balance criteria and creates a particularly hazardous merge situation during periods of heavy traffic, with slowing and eventual backups occurring in all three through lanes.

During peak periods, mainline slowing occurs at all of these on-ramp merge locations caused, in part, by an insufficient number of gaps in the outside lane to accommodate the number of merging vehicles. This problem is exacerbated by the fact that the signalized intersections at the ramp terminals operate at capacity during the peak periods, requiring the use of long cycle lengths to maximize intersection capacity. This, in turn, results in the release of large queues onto the ramps, which causes a surge of traffic to arrive at the ramp merge area.

In the southbound, p.m. peak direction, the 5300 South, 4500 South, 3300 South, and the I-215 and I-15 interchanges were again included within the study area limits. Mainline slowing at on-ramp merge locations occurs on I-15 at the 3300 South, 4500 South, and 5300 South interchanges. At the I-215 and I-15 interchange, mainline slowing on I-15 occurs because of the merge of single-lane ramps from both eastbound I-215 and westbound I-215.

In addition to exhibiting similar congestion characteristics, these sites were chosen based on the following:

- The causes of congestion observed at each site was representative of the causes seen at other Utah sites.
- This study area was small enough that extensive data could be collected in a cost-effective manner.
- Methods used in collecting data were repeatable.
- Potential remedies to recurring congestion could be studied in conjunction with data collection efforts.
- This site represented an area where recurring congestion causes noticeable effects and has a high impact on commuter traffic. It is one of the more congested segments on the freeway system in Utah.

Northbound data collection began at the westbound I-215 diverge from northbound I-15 and ended at the eastbound I-80 diverge from northbound I-15. The southbound segment began at the westbound I-80 merge with southbound I-15 and ended at the 7200 South off-ramp diverge from southbound I-15.

Data Collection Methodology

There are no advanced traffic surveillance or traffic management technologies presently in place on Utah freeways. The only permanent remote data collection devices in this segment are permanent counting stations for volumes and speeds midway between each of the diamond interchanges. Volume data were collected at these

locations during the study period. Traffic densities, as supplemental data, were also obtained during the study through the use of oblique aerial photography. Three probe vehicles were instrumented to allow them to collect speed, position, and related data.

Vehicle Instrumentation

Instrumentation consisted of a distance measuring instrument (DMI), a laptop computer, and the Moving Vehicle Run Analysis Package (MVRAP) developed by the University of Florida (*4*). The DMI was connected to the transmission of the probe vehicles and to the laptop computer. The transmission gives off a certain number of pulses for each unit of distance traveled by the vehicle and these pulses are converted into speeds and distances. This system keeps track of time, distance traveled, and speed. A speed profile (a continuous plot of speed versus distance) can be obtained for any traveled roadway segment. In addition, the time the vehicle passes the beginning and ending points of the study segment and, thus, the elapsed time to traverse the segment are recorded by the software.

The software records speed information from the DMI at 60-m (200-ft) increments along the test segment and notes the locations of link ends. When plotted, speeds along the segment are printed as points, each at approximately 60-m (200-ft) intervals. Using this plot, in conjunction with run start and end times, it is possible to determine the parameters needed to calculate the FCI.

Data Collection Preparation

The first data collection run, also known as the calibration run, required the vehicle driver to mark the starting and ending point of the segment, as well as each link end location (e.g., merge points, etc.) within the segment, by pushing the computer space bar as the point was passed. Each of these points of interest had been previously marked for easy identification. From this, the MVRAP software was able to set all the distances between the starting and ending point, as well as all link end points. The calibration run must be very precise in locating starting, ending, and link end points, because all subsequent runs are referenced to the calibration run. During subsequent runs, the driver needed only to identify the starting and ending point for the entire pilot segment by pushing the computer space bar as the reference point was passed.

For each run it was important for the vehicle to follow the same path, and for the driver to push the space bar at the same starting and ending point location as for the calibration run. The software is tolerant of small errors and will allow slight adjustments in subsequent run lengths to be made. However, the software will discard all data collected for runs that show a discrepancy of greater than 2 percent of the calibration run length, or 60 feet for link lengths or 120 feet for the total segment length, whichever is less. Driver experience and care become important.

Data Collection Procedure

Proper orientation of the probe drivers before beginning data collection is critical to a successful effort. After orientation, all three vehicles, one following directly behind the other, proceeded onto I-15 and into their preassigned lane (outside, middle, or inside) so that each passed the starting point at approximately the same time.

Speed and distance data for each lane were obtained, enabling the production of speed profiles and computation of the FCI for the entire segment. Comparisons of the outside and inside lanes with the middle lane and with the average of all three lanes could be made to determine lane differences. The hope was that a good correlation could be established between the extent and duration of congestion in one lane and the total for all lanes. Data could then be collected in one lane only and still yield a reasonable FCI for the entire section, thus lowering data collection and reduction time and costs.

Each vehicle then traveled as an "average car" in its respective lane until it passed the segment ending point. They then exited the freeway at the next interchange, reversed direction, and returned to the starting point to begin another run. At the completion of each run, all three cars would once again meet before proceeding with the next data collection run. During the study, no recurring congestion occurred in the off-peak direction, otherwise data would have been collected in this direction as well. The length of the study section was chosen, in part, so that a trip by the probe vehicles could be made through the section every 20 to 30 min.

This method was selected for the pilot project because it could be done within existing budgets using equipment available within UDOT and could be repeated using existing UDOT personnel. It is applicable to any freeway not having advanced technologies in place for monitoring traffic flow conditions.

Data were collected in each of the three lanes during the morning and evening peak period, Monday–Friday, August 16–20, 1993. Data collection began before the usual time of onset of congestion and continued until congestion had dissipated. A minor amount of congestion was sometimes encountered during the first data collection run, in which case the time marking the onset of congestion was estimated based on observations in the adjacent lanes or on the experience of other days.

Because this study was focused on quantifying recurring rather than incident-based congestion, it was essential to record all incidents that were observed by the probe drivers. In addition, traffic reports by local radio stations were monitored for news of such incidents. One member of the study team was able to observe traffic conditions while flying with an aerial traffic reporter, noting any incidents that occurred. In addition, each driver made note of any observed incidents. This became very helpful when interpreting results.

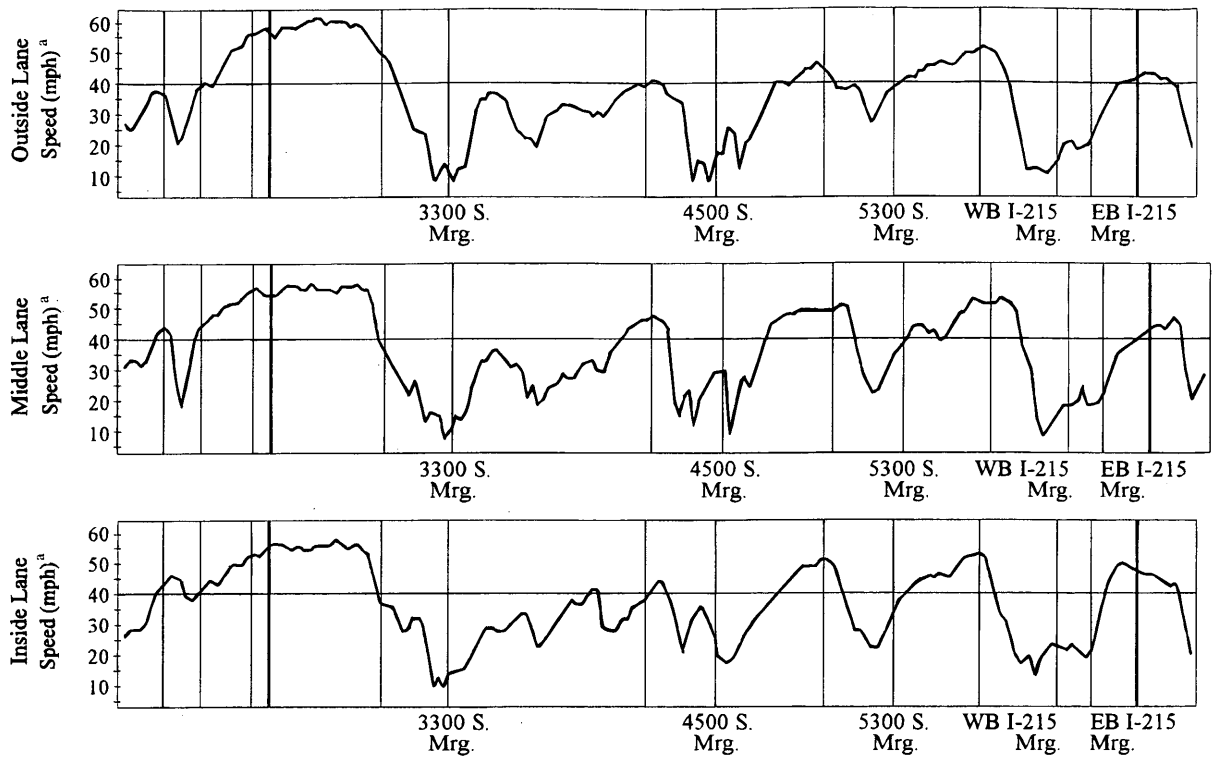
Data Reduction

Speed Profiles

All data collected by the three vehicles were combined into one file, and a speed profile was printed for each run in each lane during the week. A summary was made noting the run start and end times, the travel time between each link, the number of stops within the section, and the average running speed within each link, as well as overall segment average running speed.

After the plots were printed, all of the plotted speed points were manually connected. Each point represented the average running speed of the vehicle over a 60-m (200-ft) travel increment. Each speed profile plot occupied several sheets of paper and therefore were printed using a dot-matrix, continuous-feed printer.

Using the plot and the run information data sheet, the following information was obtained for each run:



[a]1 Km/h=0.6 mph

FIGURE 2 Speed profile well into the congested period, outbound (southbound) direction.

- The overall length of the study segment.
- The length of each congested section [i.e., traffic speed less than 64.4 km/hr (40 mph)] within the segment.
- The time between the end of one run and the end of the following run.

A typical speed profile run during the congested part of the outbound (p.m. peak) direction is shown in Figure 2.

Length of Each Congested Section Within the Study Segment

Link lengths for each link in the section are shown on the left side of the run information data sheet, as is the total or overall length of the segment. To calculate the FCI, the total number of lane-kilometers (lane-miles) within the pilot section is required. In this case, this value is three times the length of the segment because there are three lanes throughout the segment.

Each link length was used to establish a horizontal scale on the speed profiles. Once the speed points on the plot were connected, a line was drawn horizontally across the plot at 64.4 km/hr (40 mph). A congested section was defined by the point at which the vehicle speed line dropped below 64.4 km/hr (40 mph) to the point that it went back above the same line.

The total congested length for each lane was determined for each run by scaling it off the plot. When a plot was missing because of a run length problem, congested lane-kilometers (lane-miles) were estimated based on the values obtained in the other lanes if available. For an operational tool, the MVRAP software could be modified to yield the distance traveled below any selected speed and eliminate the need for this manual calculation. In fact, the software could be modified to yield the FCI itself.

Time Between Runs

The software records the travel time, in seconds, for each link and the total travel time for the entire segment for each run. The start time for the run is also recorded. The run end time (T_j) was obtained by adding the total travel time to the run start time. The difference between the ending time of one run and the ending time of the next, obtained in a similar manner, gave the duration for that run.

SUMMARY OF FINDINGS

Pilot Study FCIs

Using the procedures outlined above, Equation 3 was used to determine the product of congested lane-kilometers (lane-miles) and duration. The time taken by the probe vehicle to make a complete run from the end of the test section and return to the same point is the sampling period duration. The duration of each sampling period ranged from 19 to 44 min depending on the severity and extent of congestion, with most intervals being between 20 and 30 min. If a shorter sampling period is desired, it can be accomplished by send-

ing multiple instrumented probes through the study segment at specified time intervals.

Equation 2 was then used to calculate an FCI for each of the three lanes, by direction, for each day of the week that sampling was done. The lane FCIs were then combined to yield an average FCI for each direction for each day of the week and an AWD-FCI. These values are summarized in Table 1 and are graphically portrayed in Figure 3.

The patterns on Tuesday, Wednesday, and Thursday were similar, with outbound (southbound) FCIs ranging from a low of 0.901 on Tuesday to 1.217 on Thursday evening. The patterns on Monday and Friday were significantly different (higher), with the highest outbound FCI being 1.892 on Friday. For the AWD (Monday–Friday) the outbound AWD-FCI was 1.298 lane-kilometer-hours per lane-kilometer. To lend some perspective, it should be noted that the FCI can range from 0 (no congestion) to a maximum of 24 lane-mile-hours per lane-mile for a 24-hr period. A value of 24 means that all lanes were operating below the threshold speed of 64 kph (40 mph) for all hours of the day.

In the a.m. (inbound) peak direction, the calculated FCIs were substantially lower, ranging from 0.303 to 0.598, as shown in Table 1. An AWD-FCI could not be accurately calculated for the inbound direction because data collection for Friday was terminated as the result of a traffic accident in the northbound lanes, which substantially increased the level of congestion. This deficiency should have been compensated for by collecting data the next Friday, but it was not done.

It should be noted that the inbound FCIs were substantially lower than those for the outbound direction, a reflection of the fact that the evening peak traffic flows persisted substantially longer than the morning peaks. The day-of-week inbound pattern at first seemed to be different from the outbound pattern, with the inbound (a.m.) FCIs appearing to be somewhat more constant. The highest inbound FCI was found on Tuesday, which is not what one would normally

TABLE 1 Freeway Congestion Index (FCI)

Southbound I-15 -- P.M. Peak Period

Date of Data Collection	Inside Lane CLM * D	Middle Lane CLM * D	Outside Lane CLM * D	Average	Standard Deviation	SLM	FCI
Mon. 8/16/93	7.937	8.810	9.837	8.861	0.951	6.066	1.461
Tues. 8/17/93	4.808	5.683	5.899	5.463	0.578	6.066	0.901
Wed. 8/18/93	5.980	5.967	6.641	6.196	0.385	6.066	1.021
Thur. 8/19/93	6.706	7.610	7.826	7.381	0.594	6.066	1.217
Fri. 8/20/93	11.370	11.369	11.695	11.478	0.188	6.066	1.892
Average Weekday FCI							1.298
Tues. 10/26/93		2.772				6.066	0.457
Thur. 10/28/93		1.642				6.066	0.271
Adjusted Average Weekday FCI							0.466

Northbound I-15 -- A.M. Peak Period

Date of Data Collection	Inside Lane CLM * D	Middle Lane CLM * D	Outside Lane CLM * D	Average	Standard Deviation	SLM	UFCI
Mon. 8/16/93	2.517	---	---	---	---	5.686	0.499 *
Tues. 8/17/93	2.911	3.512	3.779	3.401	0.445	5.686	0.598 **
Wed. 8/18/93	1.253	1.971	2.199	1.808	0.494	5.686	0.318
Thur. 8/19/93	1.650	1.910	1.601	1.720	0.166	5.686	0.303
Fri. 8/20/93	---	---	---	---	---	5.686	---
Average Weekday FCI (4-day week)							0.429
Tues. 10/26/93		4.505				5.686	0.792
Thur. 10/28/93		3.043				5.686	0.535
Adjusted Average Weekday FCI (4-day week)							0.632

CLM = Congested lane miles in segment i.
D = Duration of congestion in hours.
SLM = Total lane miles in segment i.
Adjusted Average = Adjusted to an equivalent 5-day sample
* Adjusted to the average of all lanes based on inside lane measurements only.
** Transients (see Table 1) were effecting congestion levels.

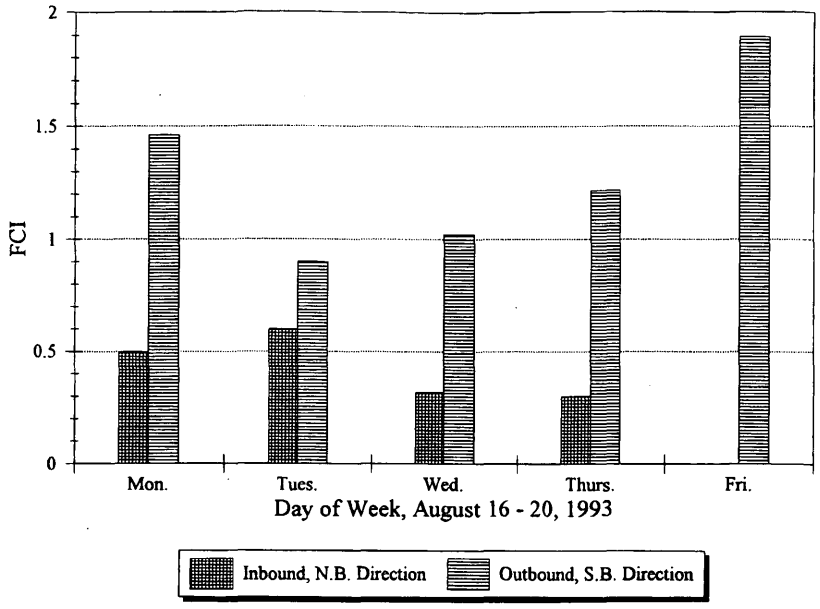


FIGURE 3 Inbound and outbound August FCI values by day of week.

expect. This anomaly is probably explained by the fact that during this particular observation period two transients were observed pushing a shopping cart along the shoulder of the freeway. The probe drivers thought that this may have been causing an increase in the level of congestion that particular morning. The elevated FCI (0.598) was confirmation of that. Taking this situation into consideration, it is probable that the inbound and outbound patterns were similar but with the inbound FCIs being substantially smaller, which was expected.

The question could be asked whether traffic during this particular week represents a typical August week. To answer this question, traffic volume data for the week of the study were compared with data from August 1992 and were found to be very similar. Thus, it was concluded that traffic flows were normal during the data collection period for the pilot study.

Seasonal Variations

Because seasonal variations in traffic flows are a normal occurrence, it would be reasonable to expect seasonal variations in congestion levels, just as there are daily variations. This situation can be handled by making sampling runs for a week during months representative of the four seasons of the year, much as is done with seasonal sampling of traffic volumes.

To give an idea of the seasonal variations, data collection runs were made on Tuesday and Thursday, October 26 and 28, 1993. The directional FCIs for these 2 days are also shown in Table 1. Interestingly, although the outbound peak FCIs of 0.457 and 0.271 for Tuesday and Thursday, respectively, were much lower than for the same days in August, the inbound peak FCIs were 32 and 77 percent higher than the corresponding August values. The only explanation we have is that this was the week of deer hunting season in Utah, an activity in which a lot of people participate, and this may have altered normal traffic behavior.

Determining the FCI Using Data From a Single Lane

It would be much less expensive to determine the FCI if the required data could be obtained using a probe vehicle in only one lane instead of all lanes of the freeway. This possibility was investigated by using linear regression analysis to compare the FCI for each of the three lanes with the FCI for all lanes combined. Regression equations were developed for each of the three lanes as compared with the average FCI for all lanes combined. The coefficients of determination ($r2$) were all very high, with the lowest being 0.968 and the highest being above 0.990, indicating a very good correlation between an FCI value based on data collection in a single lane and the FCI value based on data from all three lanes.

For an overall, bidirectional FCI based on measurement of congestion in the middle lane, the equation is

$$FCI_{all} = -0.006 + 0.990 \times (FCI_m) \quad (4)$$

where FCI_{all} = the FCI for all lanes, and FCI_m = the FCI based on measurement of congestion extent and duration in the middle lane.

For this equation, r_2 was 0.997 at the 95 percent confidence level. Similar equations were developed for the FCI based on measurement of congestion in any lane.

Implementation in a Freeway Traffic Management System

The FCI could readily be determined using automated traffic data collection. Speed measurement could be done using fixed detectors placed in the freeway lanes. A detector spacing of 0.53 to 0.8 km (⅓ to ½ mile) as recommended (*5*) for economical incident detection is suggested. Detector placement should be designed so that speed reductions can be detected in merge areas and other locations where recurring congestion usually begins. Speeds could be sampled at uniform intervals of time to determine the onset of congestion and the

extent (length) of the congested area. Software would need to be developed to permit the automated calculation of the FCI.

CONCLUSIONS AND RECOMMENDATIONS

The following conclusions and recommendations are made based on the results of this research.

1. An FCI was developed and tested that can be used to quantify both the extent (length) and duration of freeway congestion, based on a definition of the threshold of congestion as being freeway traffic stream speeds below 64.4 km/hr (40 mph).
2. The FCI can be used employing a freeway congestion definition based on a parameter other than speed, such as density, as long as the extent of congestion [number of congested lane-kilometers (lane-miles)] and duration (length of time the congestion persists) can be measured at reasonably short time intervals (e.g., 20 to 30 min).
3. Speed profiles created using an instrumented "average" probe vehicle traveling in a single lane can be used to quantify both extent and duration of congestion for use in calculating an FCI. Although the level of congestion generally decreases somewhat going from the outside lane to the inside lane, regression equations have been developed that accurately provide an FCI for all lanes based on measurement of congestion in only one lane.
4. The FCI should be usable in a congestion management system to compare changes in the congestion level on a freeway segment, subsystem, or system over time, including before-and-after comparisons of the effects of congestion management programs. It can also be used to compare levels of congestion on different segments, subsystems, or systems, including comparisons between freeway systems in different urban areas.

REFERENCES

1. Lomax, T., H. Pratt, H. S. Levinson, P. N. Bay, G. B. Douglas, and G. Shunk. *Quantifying Congestion*. Texas Transportation Institute, Interim Report 7-13. College Station, Tex., 1992.
2. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
3. Cottrell, W. D. Measurement of the Extent and Duration of Freeway Congestion in Urbanized Areas. In *Compendium of Technical Papers, 61st Annual Meeting*, Institute of Transportation Engineers, Washington, D.C., 1991, pp. 427–432.
4. *Moving Vehicle Run Analysis Package, A System for Evaluating Urban Congestion with Moving Vehicles*. University of Florida Transportation Research Center, distributed by McTRANS, Gainesville, Fla.
5. Marsden, B. G., and H. B. Wall, III. *Intelligent Data for Vehicle Detection*, SAE Technical Report Series 931924. Society of Automotive Engineers, Warrendale, Pa., 1993.

Publication of this paper sponsored by Committee on Freeway Operations.

New Method for Estimating Freeway Incident Congestion

H. AL-DEEK, A. GARIB, AND A. E. RADWAN

Incidents are a major cause of travel delays on urban freeways. This paper describes development and application of a new method for estimating freeway incident congestion where extensive loop and incident data are available. Using shock wave analysis, a time-space domain is determined for each incident. This is used to define the congestion boundaries of an incident and to decide whether the incident should be analyzed as isolated or as a multiple-incident case. The freeway section is divided into smaller segments, each segment containing only one mainline loop station. Traffic speed and counts at freeway mainline stations and traffic counts at on/off ramp stations upstream and downstream of the incident location are used to calculate incident delay on each segment during small time slices, then cumulative incident delay is calculated. Satisfactory results were achieved when the new method was applied to a sample of isolated and multiple-incident cases collected recently as part of the Freeway Service Patrol Evaluation Project on I-880 in California.

Freeway incident congestion is viewed as a major problem in urban travel. In the short-term incident congestion causes delay to travelers, wastage of fuel, secondary accidents, wear and tear of vehicles and roadways, and environmental pollution. In the long-term congestion adversely affects the economic competitiveness of a region. National statistics indicate that more than 60% of urban freeway congestion is related to incidents (*1*).

Efficient incident management may be achieved by implementing Intelligent Transportation Systems (ITS) technologies (e.g., improved incident detection techniques using video image processing). To evaluate the efficiency of ITS in reducing incident delays, it is necessary to develop methods that can estimate accurately the magnitude of nonrecurring congestion. This paper describes the development and application of a new method for estimating congestion using incident and loop data collected simultaneously.

OVERVIEW OF THE PROBLEM

Incidents include accidents, disabled vehicles, law enforcement and emergency vehicles, and spills. Several methods have been used to estimate congestion caused by an incident. Morales (*2*) developed an analytical method that plots the cumulative arrival and departure curves and calculates the cumulative vehicle hours of incident delay. In this method the congested time period is divided into smaller time intervals during which demand and/or capacity are assumed to be constant. This results in linear arrival and departure curves at the incident bottleneck. The method assumes that initial demand is less than capacity of the freeway section. The HCM (*3*) method uses the same approach of the Morales method with an important modification: it considers cases of incidents occurring during the peak period congestion with initial demand exceeding capacity. The Morales (*2*) and HCM (*3*) methods are widely used by practitioners and researchers to estimate incident delay.

Messer et al. (*4*) used the kinematic wave theory of Lighthill and Whitham to develop a method for predicting individual travel times on the freeway during incident conditions. They divided the time-space plane during incidents into areas representing four different traffic flow conditions: normal flow, queue flow, metered flow, and capacity flow. The boundaries of these areas were defined by linear shock waves, and the speed of each shock wave was derived assuming a linear speed-density relationship developed by Greenshields (*5*). The method was applied to four incidents that occurred on the Gulf Freeway in Houston. It was found that two-thirds of the observed travel times were within 10 percent of the predicted travel times. The linear Greenshields' speed-density model results in parabolic relations for volume-speed and volume-density plots. The major problem with using the parabolic curves is that if they do not match the actual conditions in regions upstream of the incident, then significant errors can be made in calculating the wave speeds. It would be more accurate to use empirical data in calculating wave speeds.

Wirasinghe (*6*) used shock wave theory to develop formulas for calculating individual and total delays upstream of incidents. The formulas are based on areas and densities of regions representing different traffic conditions (mainly congested and capacity regions) that are formed by shock waves in the time-space plot.

Chow (*7*) compared two methods for calculating total incident delay on a highway section: shock wave analysis and queuing analysis. He assumed a *unique flow-density relationship* and derived the equations of total delay, which were found to be identical for both methods. Chow (*7*) concluded that if he had used a time-dependent flow-density relationship, which is more realistic, then the two methods would yield different results.

Wicks and Lieberman (*8*) developed INTRAS, a microscopic freeway traffic simulation model designed for freeway corridor simulations. An enhanced version of INTRAS called FRESIM (*9*) is currently under testing. Both INTRAS and FRESIM have the same fundamental structure, which is based on car-following theory. The most recent calibration of INTRAS by Cheu et al. (*10*) used data from Los Angeles freeways. They concluded that INTRAS may underestimate the occupancy during free flow conditions in the recovery periods after incidents. In addition, they indicated that the car-following theory equation used in INTRAS gave satisfactory results in general, but it failed to produce high volume and occupancy that match collected field values in their study site. INTRAS and its successor, FRESIM, can be used to estimate incident congestion by simulating the freeway with and without the incident and finding the difference in vehicle hours of travel.

Department of Civil and Environmental Engineering, University of Central Florida, P.O. Box 162450, Orlando, Fla. 32816.

The main limitations in the existing methods for estimating incident congestion are summarized as follows. The assumption of static demand is clearly unrealistic under peak hour conditions because it ignores the effects of traffic diversion from the freeway to alternate routes and/or traffic avoiding the freeway system if informed ahead about the incident. Further, assuming that the initial demand level is smaller than the capacity of the freeway is not valid under peak conditions. The theoretical shock wave models assume constant densities throughout each traffic flow region, which affects the accuracy of the models. Most importantly in estimating incident congestion is that macroscopic freeway traffic models have been used to analyze/simulate one incident at a time. In real life, multiple incidents could occur simultaneously or within short periods of time on the same stretch of highway. Consequently, incident queues may merge together, and it becomes very difficult to segregate the effect of one incident from another on the magnitude of congestion. Furthermore, incident queues may merge with other queues of recurring congestion that may be present at the time when the incident occurs. In effect, it becomes very challenging to segregate incident and nonincident congestion. None of the existing methods considers these real possibilities, which puts their accuracy of estimating incident delays in question.

NEW METHOD

This section describes the development of a new method for estimating freeway incident congestion. The new method has two steps:

1. Determination of the time-space domain (or the area of influence) of an incident,
2. Calculation of delays based on speed reduction caused by the incident on freeway segments located within the time-space domain of the incident determined in Step 1. Two types of incident cases are considered: single (isolated) incidents and multiple incidents.

The two steps are described below.

Time-Space Domain of Incident

Shock wave analysis is used to determine the time-space domain of an incident. This is the area that defines the time-space boundaries of congestion caused by a specific incident as shown in Figure 1 (cases A and B). The area has dimensions ($T + D$, X), where T is

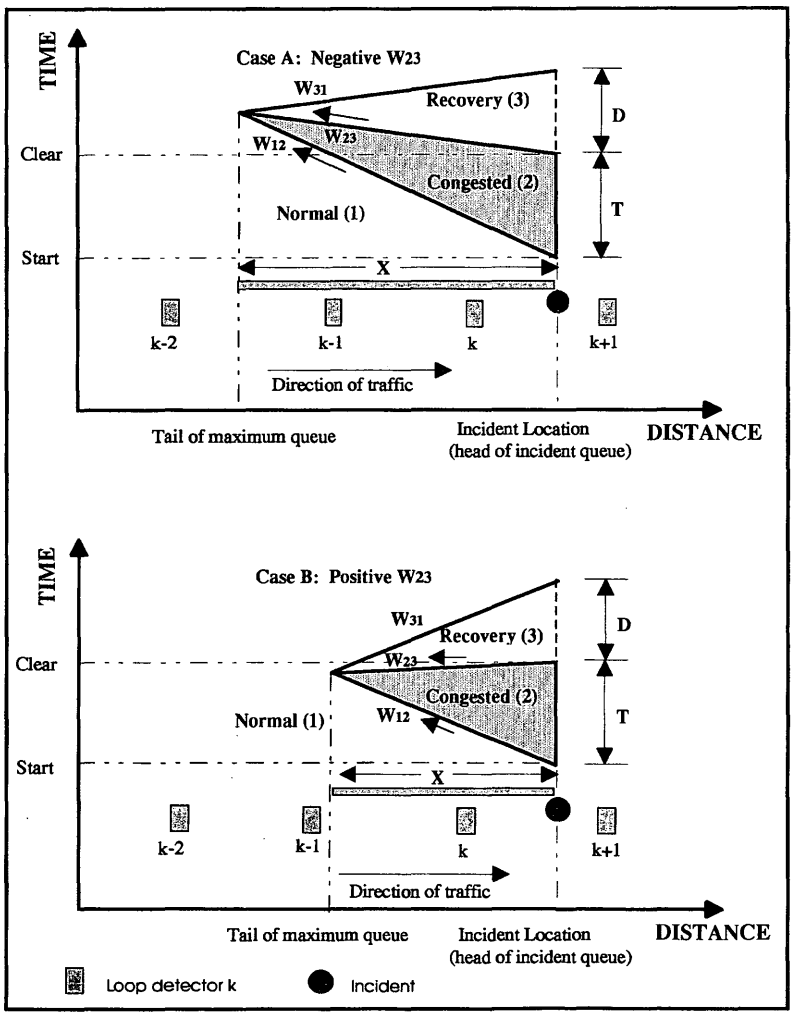


FIGURE 1 Time-space domain of incident.

the incident duration, D is the time to discharge incident queue and return to normal conditions, X is the maximum length of incident queue, and W_{ij} is the speed of shock waves forming along boundaries of traffic conditions i and j. Cases A and B of Figure 1 will be explained later. Because downstream conditions are likely to be affected by platoon dispersion, geometric bottlenecks, and other incidents, they are not considered in this analysis. We have assumed linear shock waves for simplicity. In reality shock waves may be nonlinear. Furthermore, incidents occurring upstream of the subject incident may distort the shock wave diagram by altering the queue length shown in Figure 1. Nonetheless, for the purpose of calculating the time-space domain, these effects are ignored. This is not expected to have a major impact on accuracy of congestion estimates mainly because the equations derived in this section will not be used to calculate congestion; instead, they will only be used to provide a rough approximation for the congestion boundaries of each incident. As will be explained later, incident congestion calculations are based only on reductions in loop speeds on freeway segments located within the time-space domain.

Three shock waves are shown in Figure 1 (with speeds W_{12}, W_{23}, and W_{31}) forming along boundaries of three traffic conditions (normal, congested, and recovery flow). These traffic conditions are defined as follows.

Normal Condition

This is the traffic condition at the first detector *upstream* of the incident location (i.e., detector k in Figure 1) during *the last time slice before the incident occurs, $i - 1$*. The length of time slice can be chosen arbitrarily; however, a 1-min time slice is recommended in this analysis. The traffic condition is defined in terms of three parameters: flow (F), density (K), and speed (V). The density K can be estimated by F/V, where both F and V are known from data of detector k. The three parameters will be denoted as $(F_{k,1}^{i-1}, K_{k,1}^{i-1}, V_{k,1}^{i-1})$, where the subscript 1 denotes "normal traffic condition" (*1*) and the superscript $i - 1$ refers to the last time slice before the incident occurs.

Congested Condition

Two congested traffic conditions are of interest: traffic condition at the first detector *upstream* of the incident location (i.e., detector k in Figure 1) during *time slice i, the first time slice after the incident has occurred*, and traffic condition at the same detector k during *time slice i*, the last time slice before the incident is cleared.* These two congested conditions have parameters $(F_{k,2}^{i}, K_{k,2}^{i}, V_{k,2}^{i})$ and $(F_{k,2}^{i^*}, K_{k,2}^{i^*}, V_{k,2}^{i^*})$, respectively, where the subscript 2 denotes "Congested traffic condition" (*2*). This is indicated by the shaded area in Figure 1.

Recovery Condition

This is the traffic condition at the first detector *upstream* of the incident location (i.e., detector k in Figure 1) during *time slice $i^* + 1$, the first time slice after the incident has been cleared.* This traffic condition has parameters $(F_{k,3}^{i^*+1}, K_{k,3}^{i^*+1}, V_{k,3}^{i^*+1})$, where the subscript 3 denotes "recovery condition" (*3*).

The shock wave defines the boundary separating between two traffic conditions, and its speed equals the difference in flows divided by the difference in densities of the two conditions. Hence,

$$W_{12} = \frac{F_{k,1}^{i-1} - F_{k,2}^{i}}{K_{k,1}^{i-1} - K_{k,2}^{i}}, \tag{1}$$

$$W_{23} = \frac{F_{k,2}^{i^*} - F_{k,3}^{i^*+1}}{K_{k,2}^{i^*} - K_{k,1}^{i^*+1}}, \tag{2}$$

$$W_{31} = \frac{F_{k,3}^{i^*+1} - F_{k,1}^{i-1}}{K_{k,3}^{i^*+1} - K_{k,1}^{i-1}}. \tag{3}$$

Using the geometry in Figure 1, it can be shown that

$$D = \frac{TW_{12}\,(W_{31} + W_{23})}{W_{31}|W_{23} - W_{12}|} \tag{4}$$

and

$$X = \frac{TW_{12}W_{23}}{|W_{23} - W_{12}|}. \tag{5}$$

Note that a wave speed equals the reciprocal of the slope in Figure 1. Also, note that W_{12} has always a negative sign (negative slope) because it is a backward-forming shock wave, whereas W_{31} always has a positive sign because it is a forward-recovery shock wave. But W_{23} can have either a negative sign (Figure 1, case A) or a positive sign (Figure 1, case B). Equation 4 applies if W_{23} is negative, but if W_{23} is positive then Equation 4 should be modified to

$$D = \frac{TW_{12}\,(W_{23} - W_{31})}{W_{31}(W_{23} + W_{12})} \tag{4'}$$

and Equation 5 should be modified to

$$X = \frac{TW_{12}W_{23}}{(W_{23} + W_{12})} \tag{5'}$$

Finally, it is important to mention that Equations 4, 5, 4′, and 5′ use only the absolute magnitude of W_{12}, W_{23}, and W_{31} to calculate D and X.

Search Procedure

The forming (W_{12}) and recovery (W_{23}) waves meet and define the congested region (*2*). Because both waves have to meet, otherwise the queue never diminishes, the following conditions apply:

$$\begin{aligned} &W_{12} < 0, \\ &W_{31} > 0, \\ &\text{If } W_{23} < 0, \text{ then } |W_{12}| < |W_{23}|. \end{aligned} \tag{6}$$

If the data of detector k during time slices $i - 1$, i, i^*, and $i^* + 1$ violate one or more of these conditions, then the actual incident start and/or end times may be slightly different from what has been observed in the field and, therefore, these times should be adjusted.

More specifically, the incident may have been detected a few minutes after it occurred; also, in some cases, an incident may have no effect on speed for several minutes before the reported clearance time. For example, in the I-880 incident database observers drove tach cars on the freeway continuously during 3 hours in the morning (6:30 a.m. to 9:30 a.m.) and 3 hours in the afternoon (3:30 p.m. to 6:30 p.m.) to report start and end times of incidents (*11*). The average headway of the tach cars was 7.5 minutes. Because loops normally are spaced at 0.54 to 0.81 km (0.33–0.50 mi) in most freeway systems, the incident exact location may be as far as 0.81 km (0.50 mi) from the nearest upstream loop detector. This means that the incident effect will take some time after the incident occurs until it reflects in the data of the upstream loop. Further, it is possible in incident cases where the demand level is low that after moving the incident vehicle to the shoulder and emergency vehicles leave the scene, the effect of the vehicle's presence on loop speed is negligible. Add to this the uncertainty in incident duration that has been addressed by researchers as a real concern for inaccuracy in modeling incident congestion (*4*). In an attempt to overcome these problems, we have developed a heuristic search procedure to find the adjusted incident start and end times from loop data for the purpose of obtaining accurate wave speeds that satisfy the conditions listed in condition 6 above. The search procedure is applied to speed and flow data of the nearest detector upstream of the incident and within a few minutes before and a few minutes after the observed start and end times of the incident. It is suggested that one uses 1-min time slices in this procedure. The procedure is described in the following steps:

Step 1. Identify the observed start and end times of the incident and the nearest loop detector station upstream of the incident location from field data.

Step 2. Calculate the wave speed W_{12} according to Equation 1 and using speed flow data of detector k during the first minute before and during the first minute after the incident observed start time. The wave speed W_{12} should always be negative; if not, then the incident start time needs to be adjusted and Step 2 will be repeated with the incident start time being shifted downward by 1 min. This process of downward shifting continues until a negative W_{12} is achieved. It is recommended, however, that the maximum downward shift not exceed 3 min. The 3-min threshold is about half of the tach car headway in the FSP study (*11*). A different (maybe larger) value for the maximum threshold should be used if the incident data are collected via surveillance systems (e.g., CCTV). If the 3-min downward shift does not achieve a negative W_{12}, then the incident start time is adjusted by an upward shift (in steps of 1 min to a maximum of 3 min). The time slice that produces a negative W_{12} is considered to be the adjusted incident start time.

Step 3. Update the incident clearance time by adding the incident observed duration to the adjusted start time found in Step 2 above. Calculate the wave speed W_{23} according to Equation 2 and using speed flow data of detector k during the first minute before and during the first minute after the incident updated clearance time. If the calculated W_{23} is positive, or if it is negative and satisfies condition 6 above, then the calculated W_{23} is used and the clearance time is not adjusted. Otherwise, Step 3 is repeated with downward and/or upward shifts to the observed incident clearance time until condition 6 is achieved. The time slice that satisfies this condition is considered to be the adjusted clearance time of the incident.

Step 4. Calculate the wave speed W_{31} according to Equation 3 and using speed flow data of detector k during the first minute before and during the first minute after the incident adjusted start and clearance times found in Steps 2 and 3, respectively. The calculated W_{31} should be positive; if not, then the above Steps 2 and 3 should be repeated with more shifts to the observed start and/or end times of the incident until W_{31} is positive and the conditions in 6 are satisfied.

The purpose of the above shock wave analysis is to determine where to stop (at what loop detector?) and when to stop (at what time slice?) calculating delays on freeway segments.

Calculation of Delays

Single (Isolated) Incident Delay

Nonrecurrent congestion on a freeway section can be caused by one or more incidents. Isolated incidents are cases where nonrecurrent congestion is caused by only one incident. The default is to analyze all incidents as isolated cases. However, multiple incident cases are possible whenever one or more incidents occur within the time-space domain of another incident. As will be explained later, a special algorithm has been developed to deal with multiple incident cases and to separate the congestion caused by each incident.

The freeway section under study must be divided into smaller segments. Each segment should include no more than *one* mainline station of loop detectors. The segment ends are determined by the midpoint between detectors, if there are no ramps, or by the ramp termination point when ramps exist. If all loops in a station are not working (not producing valid data), then the segment to which the station belongs is eliminated by allocating half of its length to each of the segments directly upstream and downstream of it. The purpose of freeway segmentation is to maximize use of all available loop data. The time period of interest is divided into smaller time intervals (time slices), and delay is calculated for each time slice on each freeway segment. The time-space domain of an incident is divided into a certain number of freeway segments and time slices, then delay is calculated for each time slice on each freeway segment. To estimate delay caused by an isolated incident, the following assumptions are made:

- Traffic speed and volume data are determined from the loop station on the segment, and these data are homogenous throughout the segment.
- The incident delay is calculated with respect to a reference (or base) average speed that reflects normal conditions that may or may not be congested. The reference speed represents a historical speed profile that may be used to segregate incident and nonincident congestion. The historical profile can be determined for each segment using incident free loop data as follows: for each loop detector in

the freeway section and the same travel direction, the reference speed is the average of 1-min speeds for all incident-free days in the data set studied. In other words, during each minute of the incident-free day and for each loop detector within the study section, there are two reference speeds, one for each direction of travel. One-minute speed averages are considered an appropriate level of resolution in the incident delay analysis.

Within the time-space domain of an incident, delay on each freeway segment is calculated only if the speed on the segment drops below the reference speed; otherwise, delay is null. The delay formula for each segment upstream of the incident is given by

$$D_k^i = L_k \frac{\Delta T}{60} F_k^i \left(\frac{1}{V_k^i} - \frac{1}{V_k^{ir}} \right) \quad \text{for } 0 < V_k^i < V_k^{ir}$$

$$D_k^i = F_k^i \left(\frac{\Delta T}{60} \right)^2 \quad \text{for } V_k^i = 0 \qquad (7)$$

$$D_k^i = 0 \quad \text{for } V_k^i > V_k^{ir}$$

where

D_k^i = delay on freeway segment k during time slice i (vehicle-hours),
L_k = length of segment k (km),
ΔT = length of time slice i (min),
F_k^i = flow (from loops) on segment k during time slice i (vehicles/hr),
V_k^i = speed (from loops) on segment k during time slice i (km/hr), and
V_k^{ir} = reference average speed on segment k during time slice i (km/hr).

The total delay on the freeway section that is caused by the incident is given by

$$TD = \sum_{i=1}^{m} \sum_{k=1}^{n} D_k^i, \qquad (8)$$

where

TD = total incident delay on freeway segments upstream of segment k affected by the incident (vehicle-hours),
n = number of freeway segments upstream of segment k (determined by Equation 5 or 5′), and
m = number of time slices with incident congestion (determined by $T + D$, where D is found by Equation 4 or 4′).

Multiple Incident Delays

The assumptions used in single incident analysis are also applicable here.

Separating Congestion of Multiple Incidents. Suppose that two incidents (1 and 2) occur at times t_1 and t_2, respectively, at locations as shown in Figure 2. The inter-arrival time (the time gap

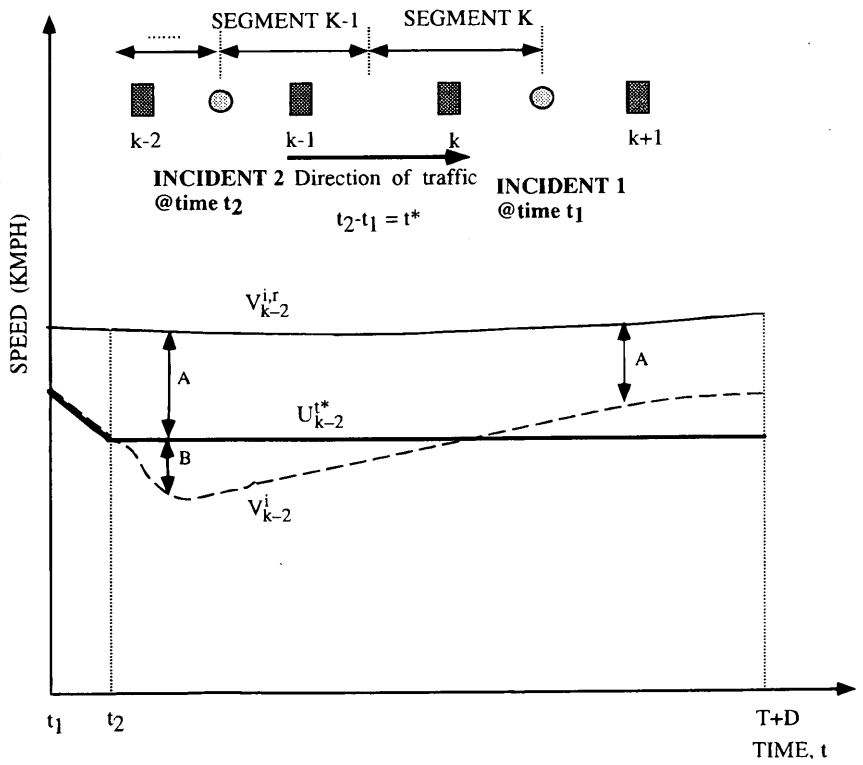


FIGURE 2 Separation of multiple incident congestion using speed profiles.

between occurrences) of the two incidents is t^*. The following illustrates an algorithm for separating the congestion caused by each incident.

Incident 1 Congestion During Time t^*. Congestion of *incident 1* during time slices before incident 2 occurs is found using Equations 7 and 8 above. Note that the two equations can be applied to find *incident 1* congestion as long as *incident 2* has not occurred, that is, these equations can be applied as long as the following condition holds:

$$m < t^*/\Delta T$$

Incident 1 Congestion on Segments Downstream of Incident 2 for Time $> t_2$. Equations 7 and 8 can be applied to find *incident 1* congestion on all segments downstream of *incident 2*. Note that the number of segments here generally will be less than n (mentioned in Equation 8 above).

Incident 1 Congestion on Segments Upstream of Incident 2 for Time $> t_2$. It will be assumed here that the congestion effect caused by *incident 1* on segments *upstream of incident 2* is captured by a drop in speed denoted by A as shown in Figure 2, where A is the difference between the reference average speed on segment $k-2$ (V^{ir}_{k-2}) and the greater of two speeds: the speed during the inter-arrival time t^* (U^*_{k-2}) or the actual speed during *incident 2* on segment $k-2$ (V^i_{k-2}). All segments upstream of *incident 2* ($k-2$, $k-3$, $k-4$, . . .) will be assessed for this type of speed drop. The delay effect of *incident 1* on segment $k-2$ upstream of *incident 2* can be found by

$$D^i_{k-2} = L_{k-2}\frac{\Delta T}{60}F^i_{k-2}\left(\frac{1}{U^i_{k-2}} - \frac{1}{V^{ir}_{k-2}}\right) \text{for } 0 < U^i_{k-2} < V^{ir}_{k-2} \quad (9a)$$

$$D^i_{k-2} = F^i_{k-2}\left(\frac{\Delta T}{60}\right)^2 \quad \text{for } U^i_{k-2} = 0 \quad (9b)$$

$$D^i_{k-2} = 0 \quad \text{for } U^i_{k-2} > V^{ir}_{k-2}, \quad (9c)$$

where

U^i_{k-2} = Maximum $\{U^*_{k-2}, V^i_{k-2}\}$.

These equations can be replicated for segments $k-3$, $k-4$, and so on. Then, the total delay can be found in a way similar to Equation 8 above.

Incident 2 Congestion. *Incident 2* will be considered to have an effect only if it reduces speed below the speed level that prevails under *incident 1* conditions. It will be assumed in this analysis that the speed level under *incident 1* conditions is represented by U^*_{k-2}, which is the speed under *incident 1* conditions before *incident 2* occurs. This speed obviously will vary during different time slices from loop-to-loop on the freeway section studied. For *incident 2* to have an effect on congestion (in addition to the effect of *incident 1*), the speed drop below U^*_{k-2}, which is depicted by the difference between U^*_{k-2} and V^i_{k-2} (shown as B in Figure 2), must be significant at the 95 percent level. The congestion effect of *incident 2* is found by

$$D^i_{k-2} = L_{k-2}\frac{\Delta T}{60}F^i_{k-2}\left(\frac{1}{V^i_{k-2}} - \frac{1}{U^*_{k-2}}\right) \text{for } 0 < V^i_{k-2} < U^*_{k-2} \quad (10a)$$

$$D^i_{k-2} = F^i_{k-2}\left(\frac{\Delta T}{60}\right)^2 - Z \quad \text{for } V^i_{k-2} = 0 \text{ and } U^*_{k-2} > 0 \quad (10b)$$

$$D^i_{k-2} = 0 \quad \text{for } V^i_{k-2} > U^*_{k-2} \quad (10c)$$

where

Z = The delay, D^i_{k-2}, calculated in Equation 9a above (10d)

These equations can be implemented into a simple spreadsheet where delay calculations can be performed for each time slice on any number of segments with incident congestion. Then delay is accumulated over all time slices and all segments affected by the incident to give the total cumulative vehicle-hours of incident delay.

INCIDENT DELAY ANALYSIS

In this section we present the results of applying the new method to cases of isolated and multiple incidents that were selected from the I-880 incident database (*11*). Incident and loop data were collected on a 11.8 km (7.3 mi) freeway section on I-880 as part of the Freeway Service Patrol Evaluation Project (FSP) in Alameda County, California. Loop stations are located approximately every 0.54 km (0.33 mi) on the study section of I-880.

The detailed results of shock wave analysis applied to an isolated incident case (incident #1456) are shown in Table 1, and delay is depicted in Figure 3. Note that the incident duration has been adjusted (it has been reduced by 4 min) using the search procedure described earlier. The calculated maximum incident queue length is 7.1 km (4.4 mi), and the duration of incident congested conditions ($T + D$) is about 55 min. It has been verified through the incident database that no other incident occurred within the time-space domain of incident #1456 (i.e., no other incident was observed along a 7.1 km (4.4 mi) freeway segment upstream of incident #1456 for a period of 55 min from the start of this incident). The estimated maximum incident queue length using loop speeds is 3.4 km (2.1 mi), which indicates that the method has overestimated the incident congestion boundaries (X and $T + D$).

Incidents #651 and #655 of the FSP database represent a multiple-incident case. Results of shock wave analysis are shown in Table 2. The methodology for separating incident delay was applied to segregate the delay for each incident case, and the delay results are shown in Figure 4. A smaller value for the maximum queue length was estimated from loop speeds (8.1 km (5 mi)), which again indicates that the method has overestimated X and, consequently, the incident congestion boundaries. Actually, it has been found that the new method overestimates the incident congestion boundaries in most of the 231 cases analyzed (see Table 3).

Table 3 shows the average and *SD* of delays for each category of incidents during morning and evening shifts. It is clear that "in-lane accidents" have the lion's share in terms of delay in this sample, whereas "right shoulder breakdowns" come second in the list (but these have the largest frequency). There are large variations in

TABLE 1 Isolated Incident Case (Incident #1456, NB I-880)

	Observed		Adjusted		Traffic Condition			
Incident # 1456	Start Time	Duration	Start Time	Duration	Normal	Congestion1[a]	Congestion2[b]	Recovery
Time(hh:mm)	5:04 PM	0:46	5:07 PM	0:42	5:06 PM	5:08 PM	5:48 PM	5:50 PM
Loop					loop11	loop11	loop11	loop11
Flow(vph)					1512	1272	1320	1266
Density(vpkm)[c]					18.5	29.9	16.4	13.7
Speed(kmph)					81.6	42.5	80.3	92.6

Incident #	Shock Wave Speed (kmph)			Time-Space Domain	
	W_{12}	W_{23}	W_{31}	D(min)	X(km)
1456	-21.1	19.5	50.7	13.4	7.1

[a]*Traffic condition at the first detector upstream of the incident location during the first time slice after the incident has occurred.*

[b]*Traffic condition at the first detector upstream of the incident location during the last time slice before the incident is cleared.*

[c]*One kilometer (km) = 0.6 mile.*

delays (e.g., *SD* is more than twice the average delay for most categories). Most studies indicated that incident duration has a large *SD*. Because the incident delay is very sensitive to incident duration, it is not surprising to see large variations in the incident delay estimates.

METHOD DISCUSSION

The following provides possible causes of over-predicting the size of the incident congestion boundaries in the new method. For simplicity, the wave speeds were based on point values of flows and densities obtained from one detector station located immediately upstream of the incident. This implied using constant and linear wave speeds over the freeway segments upstream of the incident as shown in Figure 1. In real life, wave speeds are nonlinear and dynamic. But the estimated wave speeds are likely to be large because of the sharp differences in densities immediately upstream of the incident during the one minute before and the one minute after the occurrence of the incident. Later on this difference in densities will be smaller and, consequently, the actual wave speeds will be smaller than the estimated ones. This translates into smaller congestion envelopes. That is, the congestion envelope produced by linear shock waves is expected to contain envelopes of the more realistic nonlinear waves. It can be demonstrated, using Figure 1, that for the same incident duration the maximum incident queue (X) is larger when the waves are faster than when they are slower.

Overestimation of the incident congestion boundaries does not necessarily result in overestimation of delays using the new method. The formulas for incident congestion do not use the magnitude of

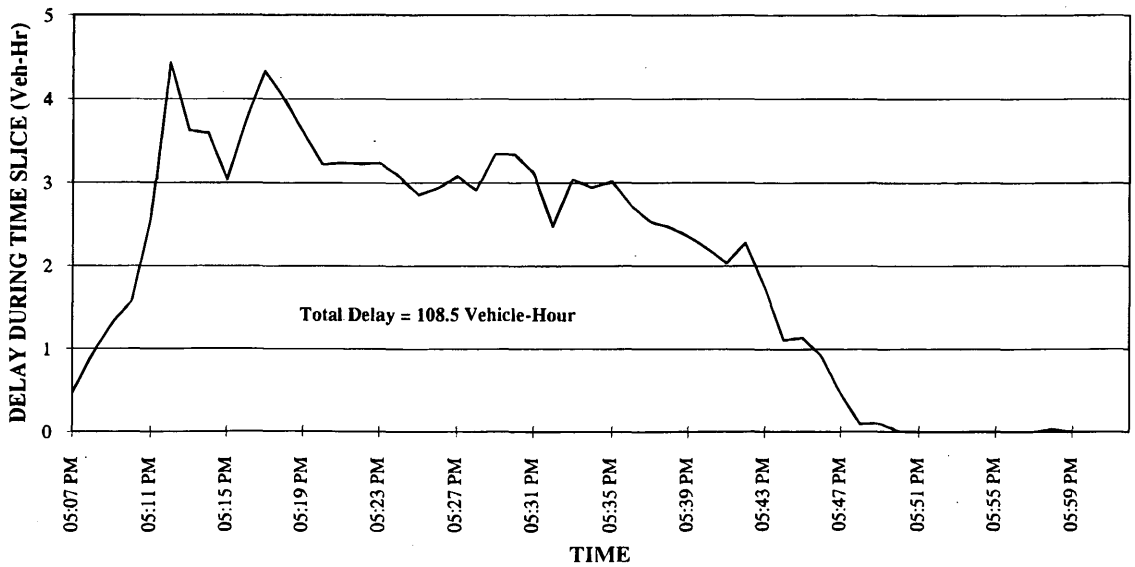


FIGURE 3 Delay for an isolated incident (#1456, NB I-880).

TABLE 2 Multiple Incidents Case (Incidents #651 and #655, SB I-880)

	Observed		Adjusted		Traffic Condition			
Incident #651	Start Time	Duration	Start Time	Duration	Normal	Congestion 1[a]	Congestion 2[b]	Recovery
Time(hh:mm)	7:04 AM	2:24	7:00 AM	2:24	6:59 AM	7:01 AM	9:23 AM	9:25 AM
Loop					loop17	loop17	loop17	loop17
Flow(vph)					1376	1264	1369	1230
Density(vpkm)[c]					14.2	28	14.4	13
Speed(kmph)					96.6	45.1	95	95

	Observed		Adjusted		Traffic Condition			
Incident #655	Start Time	Duration	Start Time	Duration	Normal	Congestion 1	Congestion 2	Recovery
Time(hh:mm)	7:31 AM	0:20	7:31 AM	0:18	7:30 AM	7:32 AM	7:48 AM	7:50 AM
Loop					loop2	loop2	loop2	loop2
Flow(vph)					1542	1374	1656	1296
Density(vpkm)					21.3	28.4	28.6	17.1
Speed(kmph)					72.5	48.3	58	75.7

Incident #	Shock Wave Speed (kmph)			Time-Space Domain	
	W_{12}	W_{23}	W_{31}	D(min)	X(km)
651	-8.1	93.4	111.1	1.8	18
655	-22.5	32.2	64.4	3.7	4

[a]*Traffic condition at the first detector upstream of the incident location during the first time slice after the incident has occurred.*

[b]*Traffic condition at the first detector upstream of the incident location during the last time slice before the incident is cleared.*

[c]*One kilometer (km) = 0.6 mile.*

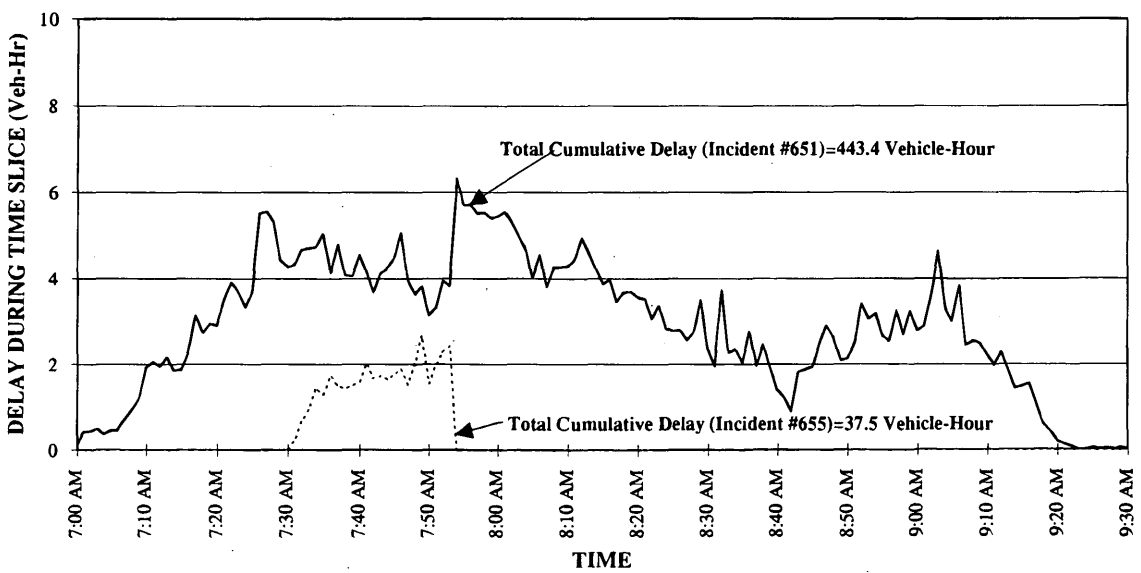


FIGURE 4 Delay analysis for multiple incidents (#651 and #655, SB I-880).

TABLE 3 Incident Delays

	Total			Morning Shifts			Evening Shifts		
Incident Type	**N[a]**	**Average**	**St-Dev[b]**	**N**	**Average**	**St-Dev**	**N**	**Average**	**St-Dev**
Right Shoulder Breakdown	176	2.5	7.1	82	0.8	2.5	94	3.9	9.2
Left Shoulder Breakdown	6	28.8	45.1	2	6.4	9.1	4	40.1	53.5
In Lane Breakdown	4	25.4	41.8	1	0.0	—[c]	3	33.8	46.8
Right Shoulder Accident	21	2.6	4.3	8	4.1	6.4	13	1.7	2.2
Left Shoulder Accident	13	11.4	28.6	6	4.5	9.1	7	17.3	38.5
In Lane Accident	11	55.0	63.8	6	74.6	72.7	5	31.6	48.0
Total Number of Incidents	231			105			126		

[a] *N = Number of incidents for each category.*
[b] *St-Dev = Standard Deviation.*
[c] *– = Standard deviation can not be calculated (one case only).*

the congested area itself. This area serves as a guideline for computations of the actual drop in speeds attributable to the incident over time and space. If there is no drop in speed on a specific segment located within the time-space domain of an incident, then delay is zero for that segment. Hence, although the area of search for a speed drop is larger than the actual one, incident congestion is not overestimated because zero delays are assigned to those segments not affected by the incident. Because over-prediction of the time-space domain is more likely to capture all segments with incident congestion, over-prediction is preferred over under-prediction. The only problem with over-prediction is more computational effort and time spent in checking for speed drops on what will turn out to be zero-delay segments.

CONCLUSION

Most of the conventional methods for estimating incident congestion (except for INTRAS and FRESIM) are incapable of using the detailed loop and incident data that recently became available in several surveillance systems and freeway traffic operations projects around the country. This is either because these methods are designed to deal with summarized types of data and make numerous assumptions or because they are too theoretical and have never been validated with real-life data. This makes them of limited use for practitioners. Moreover, it is not possible to use the conventional *macroscopic* methods for analyzing cases of multiple incidents that occur on the same stretch of highway resulting in multiple queues that merge together. On the other hand, *microscopic* traffic analysis tools such as INTRAS and FRESIM require extensive calibration with loop data and making assumptions about car-following theory.

This paper has presented a new macroscopic method for estimating freeway incident congestion. The method is based on shock wave analysis where the area of influence of a specific incident is demarked. If the time-space domain of incidents overlap, they result in a case of multiple incidents. An algorithm for separating congestion of each incident has been described in this paper. Also, the time-space domain of an incident is used to distinguish between isolated and multiple incident cases. In the new method, incident detection and clearance times collected simultaneously with speeds and traffic counts from mainline loop stations and on- and off-ramp stations are used to calculate incident delay on each freeway segment and for each time slice during the congested time-space region and also to obtain cumulative incident congestion. The method is applied to a sample of incident data from the FSP database of I-880 in Alameda County, California. The sample includes both isolated and multiple incident cases, and the application results are reasonable. Generally, the method overestimates the maximum incident queue length and, consequently, the incident congestion boundaries. However, this does not necessarily overestimate the incident delay. Incident delay is not calculated using the congested areas confined by the time-space domain; rather incident delay is based on the actual drop in speeds on the freeway segments upstream of the incident.

Future research will focus on two main issues:

1. Refinement of this method by calibrating the estimated time-space domain with tach car data. These data include tach vehicle travel times and speeds, which is another source of independent field data collected in the FSP project, and
2. Comparison of incident delay results with those of FRESIM for the same sample of the FSP incident database used in this paper.

Also, the authors will seek applications of the new method in other sites, where similar data have already been collected, such as I-4 in Orlando, Florida (Al-Deek, unpublished data). Although the new method needs further refinements, the authors hope that this paper has accomplished an important step toward bridging the gap between theory and practice in the field of freeway traffic operations.

ACKNOWLEDGMENTS

The authors acknowledge the help of Professor Pravin Varaiya, Dr. Alex Skabardonis, Karl Petty, Hisham Noeimi, and Dan Rydzewski of the University of California at Berkeley in providing the research team with data and comments on this paper.

REFERENCES

1. Lindley, J. Urban Freeway Congestion: Quantification of the Problem and Effectiveness of Potential Solutions. *ITE Journal*, Vol. 57, No. 1, January 1987, pp. 27–32.
2. Morales, J. M. Analytical Procedures for Estimating Freeway Traffic Congestion. *Public Road*, Vol. 50, No. 2, September 1986, pp. 55–61.
3. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1994, pp. 7–17.
4. Messer, C. J., C. L. Dudek, and J. D. Friebele. Method for Predicting Travel Time and Other Operational Measures in Real-Time During Freeway Incident Conditions. In *Highway Research Record 461*, TRB, National Research Council, Washington, D.C., 1973, pp. 1–16.
5. Greenshields, B. A Study of Traffic Capacity. *HRB Proceedings*, Vol. 14, 1934, pp. 448–477.
6. Wirrasinghe, S. Determination of Traffic Delays from Shock Wave Analysis. *Transportation Research*, Vol. 12, 1978, pp. 343–348.
7. Chow, W. A Study of Traffic Performance Models under Incident Conditions. In *Highway Research Record 567*, HRB, National Research Council, Washington D.C., 1974, pp. 31–36.
8. Wicks, D., and E. Lieberman. *Development and Testing of INTRAS, a Microscopic Freeway Simulation Model: Program Design, Parameter Calibration and Freeway Dynamics Component Development*. Report No. FHWA/RD-80/106, FHWA, U.S. Department of Transportation, 1980.
9. FHWA, Office of Research and Development. *FRESIM User Guide*. Beta Version 3.1, 1992.
10. Cheu, R., W. Recker, and S. Ritchie. Calibration of INTRAS for Simulation of 30-Second Loop Detector Output. *Transportation Research Board 72nd Annual Meeting*, Washington D.C., January 1993.
11. Skabardonis, A., H. Noeimi, K. Petty, D. Rydzewski, P. Varaiya, and H. Al-Deek. Freeway Service Patrol Evaluation, California PATH Research Report, UCB-ITS-PRR-95-5, Institute of Transportation Studies, University of California Berkeley, 1995.

Publication of this paper sponsored by Committee on Freeway Operations.

Costs and Benefits of Vision-Based, Wide-Area Detection in Freeway Applications

PANOS G. MICHALOPOULOS AND CRAIG A. ANDERSON

Wide-area detection systems (WADS) through video image processing is gaining worldwide acceptance as a proven technology for IVHS, as well as a preferred emerging technology for replacing loops in many practical situations. This technology has been tested and validated in many real-life applications. The advantages and sophistication of WADS are easily realized at intersections where the large number of detectors and need for wide-area measurements lead to up-front cost justification; this is not so obvious on freeways because of sparse detection and current lack of widespread WADS applications. In this article, a direct comparison of loops versus WADS is made, assuming that WADS is only being used as a direct replacement of loops. Even when ignoring intangible benefits, it is demonstrated that when an economic analysis is performed, WADS can be substantially more cost effective than loops. Intangible benefits include stopped vehicle and incident detectors, automatic extraction of measures of effectiveness and performance measurement, wide-area detection, continuous visual performance verification, accurate speed measurement through vehicle tracking, surveillance at minimal incremental cost, and others.

The need for advanced traffic detection devices that reduce installation and maintenance costs while extracting more traffic flow measurements has lead to the development of machine vision wide-area detection systems (WADS). Machine vision provides "above road" detection and wide-area measurements that replace many conventional inductive loop detectors. Lane closure costs for loop installation and maintenance are effectively eliminated. In addition, the detection performance is easily verified and detectors are easy to reconfigure interactively.

The Autoscope WADS, selected for this cost comparison, has been installed in over 400 sites in North America, Europe, and Asia for both freeway and intersection applications. Although reliability and performance have been assessed through various installations and studies (*1–4*), cost effectiveness on freeways has not been documented. The objective of this paper is to present benefit and cost results based on data collected on a typical freeway detection installation in Minnesota, which is actually unfavorable to WADS because it only requires sparse detection for only two main-line lanes in each direction and on ramps. Furthermore, the installation was intended only for conventional electronic surveillance and ramp metering; as such, only volume and time occupancy are currently measured so that the detector stations consist of single loops rather than the loop pairs that are usually required for speed measurement. In spite of this, the comparison results are very favorable for video detection because they indicate that Autoscope was cost effective even without accounting for many of its intangible benefits. The Minnesota Department of Transportation (Mn/DOT), which funded this project, provided all the data associated with the actual costs of a recently instrumented section of Trunk Highway 36 in Minnesota in which the loops were installed. They also provided up-to-date infrastructure costs associated with loops and video detection and specified the three alternative design options associated with the video. In this article, the overall project objectives and functional specifications of the machine vision device used are presented. This is followed by a description of the site selected for the cost study, the methodology and assumptions, the benefits of the WADS system considered, and the results of the economic analysis.

P. G. Michalopoulos, Department of Civil Engineering, University of Minnesota, Minneapolis, Minn. 55455. C. A. Anderson, Image Sensing Systems, Inc., St. Paul, Minn. 55104.

BACKGROUND

One of the primary objectives of the project completed for Mn/DOT was to compare costs and benefits of video versus conventional loop detection for operational deployment on freeways. The replacement of current loop functionality was a primary requirement. However, the functional capabilities of the Autoscope WADS used in this cost-benefit study exceeded the capabilities of the loop detector alternative against which it was compared. In addition to the individual detections, volume counts, and time-occupancy provided by the loops, the video detection system provides wide-area detection and speed, as well as vehicle length measurements. Furthermore, it classifies vehicles based on vehicle length. These measurements are simultaneously accumulated into time intervals ranging from 10 sec to 1 hr and are made available to the user via serial communications. In addition to these parameters, space mean speed, space occupancy, density, average time-headway per lane, and user-defined level of service congestion grades are generated. Detector outputs can be combined using logical "or," "and," or "nand" operations and can be delayed or extended for user-defined times. These last functions are particularly useful for complex applications, such as adaptive intersection or ramp control based on wide-area detection and reporting alarms for incidents.

Recently, an incident detection algorithm was added to the processor itself. This follows a natural trend to distribute processing within a traffic management system and reduce communication bandwidth requirements. It also allows the algorithm to use data that otherwise might not be available to a central traffic detector server.

TEST SITE AND DESIGN ALTERNATIVES

The site selected for the cost-benefit study was a 4.7-km (2.8-mi) section of Trunk Highway 36 north of St. Paul where a conventional loop detection system and three closed-circuit television (CCTV) surveillance cameras were installed as part of a state construction

project completed in the fall of 1993. This portion of freeway has two main-line lanes in each direction and five interchanges as shown in Figure 1. As built, there are six detector stations in each main-line direction, a detection station on north- and southbound Snelling Avenue, and detection on all 24 adjoining ramps. Each lane or ramp detector consists of a single loop detector, as shown in the Figure 1, for providing only volume and time-occupancy information.

To evaluate the effect of camera placement and coverage on cost, three Autoscope deployment alternatives were chosen for comparison with the actual loop installation. Each alternative was required only to provide detection equivalent to the loops. This requirement underutilizes the WADS capabilities but was done deliberately to assess the worst-case scenario in which Autoscope is used only as a loop replacement for counting applications. Alternatives 1 and 2 use supplemental loop detection on ramps not within the camera's field of view, because only point detection was required on ramps. Alternative 3 uses machine vision exclusively for detection.

Alternative 1 was configured so that the video cameras were located in the median to provide detection for each main-line direction, as well as those ramps within the field of view. The median was as much as 23 m (75 ft) wide in some places, permitting easy access for installation and maintenance. This type of camera placement is not recommended for installations with three or more lanes in each direction when the median exceeds 5 m in width. A total of seven cameras was needed to provide detection on the six main-line stations and Snelling Avenue, as well as on 10 of the 24 ramps. A total of 14 loop detectors were used on ramps to supplement the video detection.

Alternative 2 was configured to have the video cameras located near the outside shoulder of each main-line direction, doubling the number of cameras to 14 to provide detection at all detector stations plus 14 of the 24 ramps. This is a typical camera placement with three or more lanes in each direction and a wide median greater than 5 m. As a result of the added cameras, only 10 loop detectors were used on the ramps, again to supplement the video detection.

Alternative 3 was identical to Alternative 2, except that supplemental detection on 10 ramps was accomplished with 8 additional cameras, bringing the total to 22 cameras. This alternative, therefore, used video detection exclusively throughout the entire roadway.

It should be noted that cost reductions were achieved by moving the locations of the detector stations so that cameras could view the main-line traffic plus exit and entrance ramps wherever possible, in all three alternative deployment scenarios.

METHODOLOGY AND ASSUMPTIONS

This portion of Trunk Highway 36 freeway was selected for the study primarily because current cost data was readily available for the loop detector installation, which had been completed during the 1993 construction season. Although less favorable to WADS than wider urban corridor freeways, this portion of freeway is representative of a major portion of metro area suburban freeways and highways. If the cost comparison were to be favorable on this roadway, it would certainly be even more favorable on wider roadways, consisting of three or more lanes per direction and narrower medians, simply because the wide-area detection capabilities are available at no extra cost.

The Mn/DOT Traffic Management System (TMS) plan sheets from the construction project were used for both the loop and the Autoscope installation alternatives to derive statements of estimated quantities from which the costs were calculated. In addition, because the loop installation required lane closures and were actually installed during daylight hours (9:00 to 14:30), the user delay, in vehicle hours, for main-line detector stations was computed using the KRONOS Freeway Simulation Program, which has been tested and validated for Minnesota freeways for over 10 years (*5,6*).

Ground rules for estimating costs were to include the incremental costs required to exclusively support either loop or video detection. For example, the majority of conduit was laid to support ramp

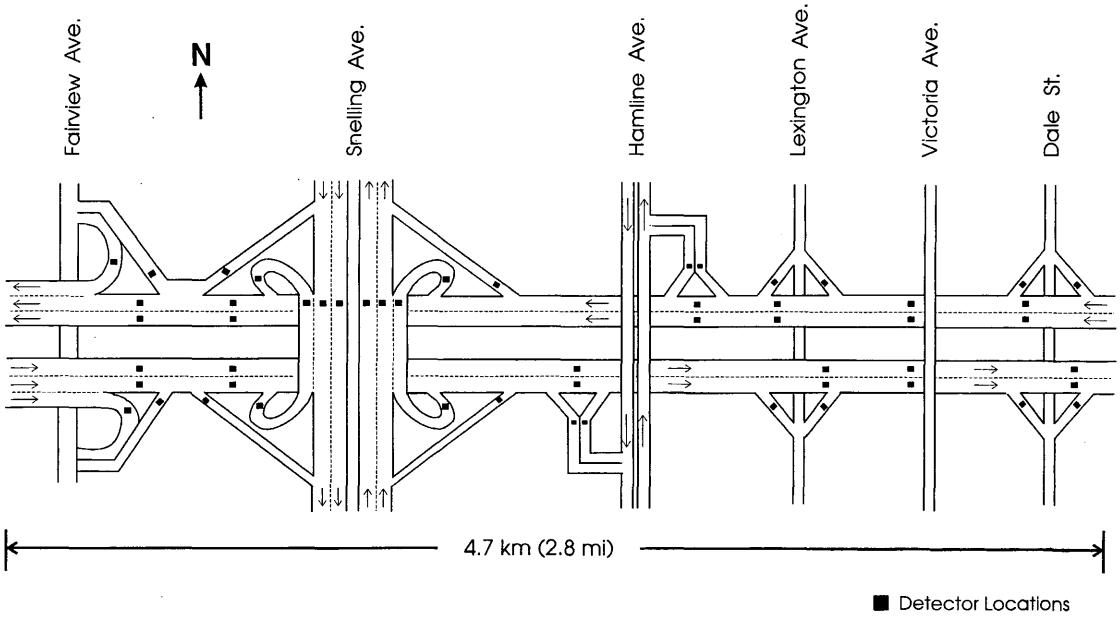


FIGURE 1 Section of Trunk Highway 36 used for detection cost comparison.

meters; therefore, estimated quantities did not include that conduit because it had the capacity to carry lead-in wire for loops and power lines and video cable for cameras. It was clear from the TMS plan sheets that the loops were laid out to maximize the usage of ramp meter conduit in order to minimize total construction costs. In turn, this same consideration was used when camera locations were selected.

Following standard Mn/DOT procedures, all linear quantities measured from the plan sheets were increased by 4 percent to account for grade changes. As a spot check, the total measured length of loop detector lead-in wire was compared and agreed to within 1 percent of original Mn/DOT estimates made before construction.

It was further agreed to use pricing and engineering practices that were used for this specific 1993 construction project. Therefore, all loop and WADS costs are in 1993 dollars. The only significant change in engineering practice is that trenched conduit for camera video and power was required to be rigid steel conduit in 1993, whereas today it is standard practice to use nonmetallic conduit, which is 60 percent lower in price. As a result, the total system costs would be reduced by $3,300 for Autoscope Alternative 1 and by $5,000 for Autoscope Alternatives 2 and 3.

Primary components of the loop detectors are the loop itself, lead-in wire, detector amplifier cards, input files to hold the amplifier cards, and the 170 controllers which processed and transmitted the detection data back to the central Traffic Management Center (TMC). However, the 170 controllers, except for the Victoria Avenue station, were used for ramp control and their cost was not included as part of any detection system. If no ramp control were present, the cost of the loop installation would be increased by the cost of the 170 processors. Each furnished and installed loop cost $560, two conductor No. 14 lead-in wire cost $2.07/m ($0.63/ft), and the four-channel inductive loop detector amplifiers were $153 each.

Primary components of the WADS system are the cameras, including fixed focal length lens, enclosure, and mounting brackets; the WADS processors; and video coaxial cable and power to the camera. The average price of the camera system and the WADS processor used was $7,000 per camera in 1993, in spite of the anticipated price drop as the technology became widespread. The price of the RG-11 coaxial video cable was $2.46/m ($0.75/ft), and the three-conductor No. 10 wires for power were $3.28/m ($1.00/ft).

Note that all prices for loop and video detection components include materials and labor to install. Traffic control costs, which are typically 3 percent of Mn/DOT project costs, were estimated to be $40 per loop, by simply taking 3 percent of the total system cost and dividing by the number of loops. Although no lane closures are required for the WADS alternatives, traffic control costs were nevertheless added to the total system costs. These costs were arbitrarily chosen as $20 per camera, one-half of the traffic control cost per loop, as a conservative estimate.

Finally, the indirect user cost because of lane closures was estimated by multiplying user delay in vehicle hours by a user cost in dollars per vehicle hour. A delay cost of $10.65 per vehicle hour was derived from (*7*), which quoted an FHWA study that used $8.42 per vehicle hour because of lost time and wasted fuel in 1987. This number was increased to $10.65 by assuming an inflation rate of 4 percent per year.

The University of Minnesota Center for Transportation Studies provided assistance in estimating the delay in vehicle hours because of lane closures required to install loops. The roadway geometry and loop detector locations were measured from the plan sheets. Actual traffic demand data from a typical construction day was used for the KRONOS freeway simulation. A day with fair weather and typical roadway volumes was chosen, and then traffic data was extracted from the TMCs detector data base collected by these same loops, which had been installed 6 months earlier. The loops were actually installed between 9:00 hr and 14:30 hr, and the contractor was required to close a minimum of 370 m (1200 ft) of lane upstream of the saw cut for the loop. It typically takes 2 hr per loop installation. A simulation lane closure schedule was set up to complete a full installation in one lane of roadway. It should be noted that because of simulation program constraints, the length of lane closure was limited from 60 m to 245 m (200 ft to 800 ft), in all but one case, instead of the 370 m required. Therefore, the simulation can be expected to underestimate the delay.

An estimated two-lane roadway capacity of 4,800 vehicles per hour and constricted capacity of 1,500 vehicles per hour with one lane closed were specified for the simulation. The capacity was estimated using measured data from all main-line detectors and the constricted capacity reduction of 68 percent was extrapolated from the *Traffic Engineering Handbook* (*7*), which provides the capacity reduction that results from lane closures on a typical freeway.

The simulation results are shown in Table 1 for a baseline run with no lane closures and the three simulator construction runs. The delay is accumulated only for vehicle speeds below 64 kph (40 mph). A total delay of 3,853 vehicle hours resulted for one lane of closure to install loops. This delay was multiplied by four to obtain a total estimate of 15,400 vehicle hours of delay for the entire four-lane roadway. A total estimate of 34,000 L (9,000 gal) of extra fuel was similarly obtained. The $10.65 per vehicle hour delay includes the cost of extra fuel and lost time. The resulting user cost because of delay is $164,000. Note that this is not a direct cost paid by the department, but an estimate of an indirect cost borne by the users of the road. This cost could be reduced by installing loops at night, with modest increases in the direct costs of installation. However, the objective of this comparison was to compare with an actual loop installation that occurred during daytime hours using the provided "furnished and installed" prices, which were for daytime work and not nighttime.

Finally, to allow extrapolated comparison with wider roadways, the cost of adding loop detection for an additional one lane in each direction was computed from which total system costs for three and four lane highways could be derived. The user-delay costs for these wider roadways would likely be greater as volumes also increased, however such increases were not considered in this study, and the same value of $164,000 in user-delay costs was conservatively used for these wider roadway estimates.

Mn/DOT had two system design requests that significantly impacted total system costs that are not included as part of the system costs because they were not considered necessary to provide detection capability functionally equivalent to loops. However, it is worthwhile to discuss these design requests, the reasons for the requests, and the incremental impact on WADS system costs. The first request arose from Mn/DOT's desire to take advantage of a key Autoscope benefit, the capability to verify detector performance as well as video quality, remotely. To realize this benefit, they requested that the video transmission from the camera to the WADS processor use an existing multimode fiber-optic line to transmit the video from the field cabinet to a regional TMS shelter where the WADS processors could be located. In turn, the outputs of the multicamera WADS processors could be selected for transmission on

TABLE 1 KRONOS Simulation Results for Installation of Six Loops in One Lane of Westbound Trunk Highway 36

Length of Roadway: 5624m (18,450 ft)
Time of Simulation: 8:00 A.M. to 3:00 P.M. each day
Lane Closures Permitted from 9:00 A.M. to 2:30 P.M.

Lane Closure Per Loop:
2 hours required
60m to 245m (200 - 1200 ft of lane closed)

Simulation Run	Total Travel Vehicle-Km (vehicle-mile)	Travel Time (vehicle-hr)	Ave. Speed Km/hr (mph)	Delay* (vehicle-hr)	Fuel Liter (gal.)
BASELINE (no lanes closed)	91,732 (57,000)	918	100.0 (62.1)	0	13,601 (3,593)
CONSTRUCTION Day #1	88,232 (54,825)	2,071	42.7 (26.5)	1,014	14,842 (3,921)
CONSTRUCTION Day #2	95,648 (59,433)	3,805	25.1 (15.6)	2,583	20,149 (5,323)
CONSTRUCTION Day #3	93,204 (57,914)	1,225	76.1 (47.3)	256	14,316 (3,782)

* Delay computed only for vehicle speeds under 64 Km/hr (40 mph)

the single-mode, fiber-optic line back to the TMC. This would permit detector layout, performance verification, additional surveillance, and video quality monitoring to be performed from a central location, thereby reducing maintenance costs incurred by trips to the field to troubleshoot system operation.

The incremental cost to realize this benefit consists of multimode fiber-optic transmitters and receivers, a length of fiber-optic line to splice into the fiber-optic backbone, and a splice vault in which to place and protect the splice. The basic costs for each WADS system does not include this cost, because loops do not provide this capability and the comparison would be unfair; however, this feature will be discussed in the benefits-costs comparison discussion. The basic cost for each system does include the cost of coaxial cable runs from the cameras to the WADS processor in the field cabinets.

The second request was that each camera be mounted on separate, specially designed CCTV poles. These innovative "crank-down" poles were designed to support much heavier CCTV surveillance cameras with full pan-tilt and zoom capability and washer-wiper systems, to minimize movement from wind and vibration, and to enable maintenance crews easy access to the CCTV camera system without the services of a bucket truck. Additionally, the poles can be located in places that bucket trucks cannot reach. The incremental cost of these benefits will likewise be discussed in the comparison discussion.

COST AND BENEFIT COMPARISON

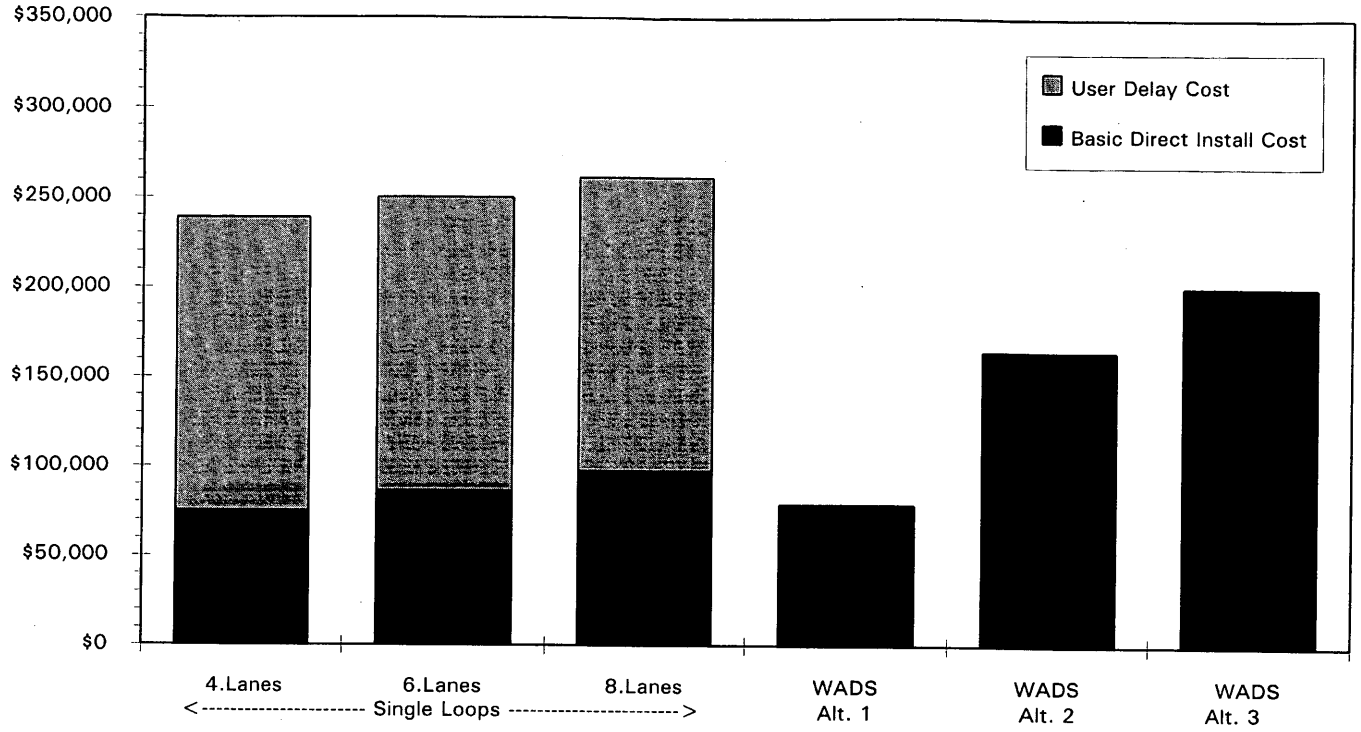
The total system costs are shown in Figure 2 for the loop (four lanes) and CCTV surveillance systems per TMS plan sheets, for the loop installation extrapolation to six and eight lanes, and for all three Autoscope alternatives. Note that the WADS cost is the same for four-, six-, and eight-lane roadway options. The indirect user-delay costs because of main-line lane closures required by the loop installation have been distinguished from the basic out-of-pocket direct costs of installation. The user-delay costs resulting from loop installation ramp closures were not taken into account in this study because of the lack of sufficient diversion information and the limited budget for the study.

It is significant to note that when the road is resurfaced, on average every 8 years in Minnesota, that the loops must be installed again. The direct costs to replace the loops is roughly an additional 45 percent of the original direct cost. However, the user-delay cost will be incurred in full or will be even greater if road usage has increased. These additional costs do not appear in Figure 3. Elimination of significant recurring user-delay costs and recurring loop install costs are a primary benefit of the WADS alternatives.

A further breakdown of costs for the loop and WADS alternatives shown in Figure 2 is given in Table 2. Note that the user-delay cost was only computed for the freeway configuration with two lanes in each direction. A conservative value of $10.65 per vehicle hour of delay was used to convert the delay into dollars because of lane closures for loop installation. This value includes lost time and fuel costs. Values ranging from $10 to $15 per vehicle hour are commonly used to determine the cost of delay.

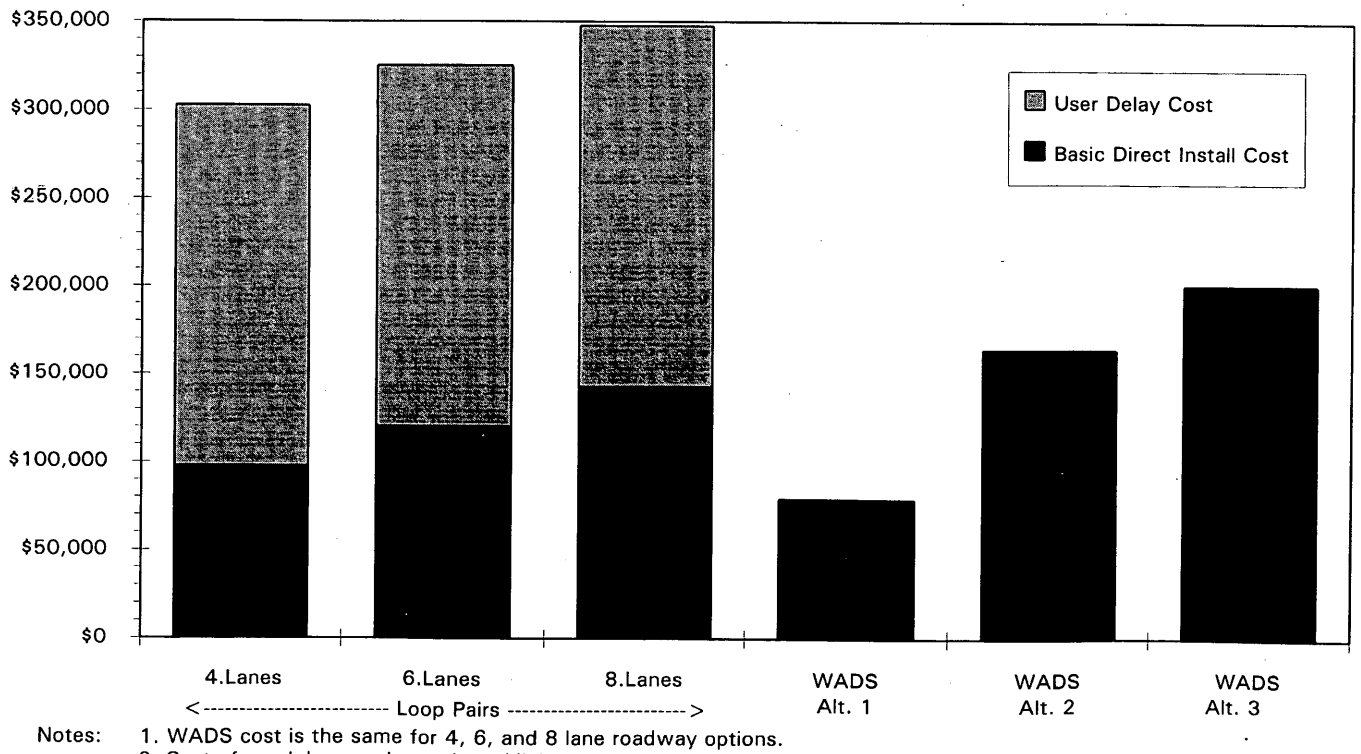
Benefit-cost ratios of 1.25 to 18.4 for the two lanes in each direction roadways were computed by dividing the incremental direct cost of the WADS installation into the WADS benefit; in this case the avoidance of delay cost because of lane closures. For example, WADS Alternative 1 cost $8,896 ($78,890 to $69,994) more than the loop direct costs. Dividing this cost into the $164,000 delay cost not incurred by the WADS system produces a benefit-cost ratio of 18.4. Note that favorable benefit-cost ratios resulted even for WADS Alternative 3, which used video detection exclusively.

The number of detectors used for the comparison of the loop and WADS alternatives also is shown in Table 2. Note that the Autoscope processor has significant unused detection capacity for all three WADS alternatives. The unused capacity, in number of detectors, can be calculated by subtracting the number of detectors used from a potential of 25 detectors per camera. Actually, each Autoscope processor will process up to four cameras simultaneously with at least 100 detectors distributed between the cameras.



Notes: 1. WADS cost is the same for 4, 6, and 8 lane roadway options.
2. Cost of crankdown poles and pan/tilt/zoom cameras would increase WADS cost.

FIGURE 2 Total system cost comparison of WADS versus single loops.



Notes: 1. WADS cost is the same for 4, 6, and 8 lane roadway options.
2. Cost of crankdown poles and pan/tilt/zoom cameras would increase WADS cost.

FIGURE 3 Total system cost comparison of WADS versus loop pairs (for speed).

TABLE 2 Summary of Installation Cost Estimates

Brief Description	Number of Mainline Lanes Per Direction	Loop Cost	VIDS Cost	User Delay Cost*	Total System Cost	B/C	No. of Detectors	Cost Per Detector
Conventional Loops Only	2	69,994	—	164,000	233,994	—	54	4333
	3	80,812	—	164,000	244,812	—	66	3709
	4	91,630	—	164,000	255,630	—	78	3277
WADS Alternative 1								
Median Cameras Supplemental loop detection on ramps	2	19,555	59,335	0	78,890	18.4	54	1461
	3	"	"	"	"	—	66	1195
	4	"	"	"	"	—	78	1011
WADS Alternative 2								
Side view cameras Supplemental loop detection on ramps	2	20,930	143,478	0	164,408	1.7	54	3044
	3	"	"	"	"	2.0	66	2491
	4	"	"	"	"	2.2	78	2108
WADS Alternative 3								
Side view cameras Video detection on all ramps	2	—	200,923	0	200,923	1.2	54	3720
	3	—	"	"	"	1.4	66	3044
	4	—	"	"	"	1.5	78	2576

*Use delay cost calculated only for mainline on 2 lane roadway.

Finally, to assist comparison of each alternative, the system cost has been converted to a "cost per detector" in Table 2. Note that WADS Alternative 1, in which maximal use of the wide-area capability is made by mounting the camera in the median, has the lowest cost per detector of any of the WADS alternatives, as expected, and is lower than the loop installation as well.

In the funded study, speed measurement was not a requirement because speed is not yet used in Minnesota for surveillance and control because of cost and loop maintenance considerations. Loop pairs can be located in the roadway at a known separation distance and sampled at high frequency (100 Hz) to measure the speed of vehicles passing between the loops. Although the most desirable freeway state variable to measure is density, it cannot be measured directly from point sensors, such as loops, but can be estimated from flow and space mean speed. However, it suffers from the sampling noise of the point flow measurement over time and assumes constant flow over a local region. In the absence of sensors to measure density directly, speed can very effectively be used. An important feature of speed is that it can be measured at a point and does not require many vehicles to sample accurately. As a result, rapid breakdowns in traffic can quickly be assessed to enable timely management and control decisions to be applied.

If speed is required, the benefit of WADS takes on significant value. The extra loops required for speed measurement would cause a significant increase in total cost. The added cost of extra loops was computed and is shown in Figure 3. Increases in direct cost are 30 percent, 39 percent, and 46 percent for the four-, six-, and eight-lane roadways, respectively. The user-delay costs were conservatively estimated to increase by 25 percent because of the added delay resulting from the time to install the extra loops.

As discussed in the previous section, supplemental surveillance and centralized video troubleshooting are benefits that result when the WADS system is connected to an already existing fiber-optic communications backbone. The incremental cost to add fiber-optic connections to Autoscope Alternative 1 is $22,700 and is $39,700 for both Alternatives 2 and 3. The same benefit could be accomplished with wireless video transmission alternatives in the absence of fiber-optic lines. The cost of wireless video transmission was not evaluated in this study.

Each of the three WADS alternatives evaluated assume that existing poles or structures are available on which to mount cameras. Adding poles specifically for WADS cameras will increase the total cost. The innovative crank-down poles that were evaluated for this study cost nearly $7,000 a piece. The desire to avoid using bucket trucks for camera-pointing necessitates an additional cost of roughly $1,000 for pan-tilt and zoom capabilities. Although providing flexibility of pole placement and the potential for reduced maintenance costs, providing poles exclusively for WADS main-line cameras significantly increases costs. The incremental cost increase to add special crank-down poles and pan-tilt and zoom capabilities would be 73 percent of the total system cost for WADS Alternative 1, 62 percent for Alternative 2, and 55 percent for Alternative 3. It should be noted, however, that outside of Minnesota, no such expensive poles have been used in connection with video detection.

In addition to the benefits already discussed, there are many more that may be important to consider in a cost-benefit comparison for specific freeway projects. The benefits of WADS over loop detectors are summarized in Table 3. These benefits must be evaluated in each specific freeway project where WADS is under consideration. What is a valuable benefit to one agency may be of less value to

TABLE 3 Comparison of Video WADS versus Loops for Benefit Evaluation on Freeways

Function	WADS	Loops
Year round install, maintenance	Yes, except for underground conduit and wiring	No
Lane closures required	No, maybe shoulder	Yes
Usable during reconstruction	Yes	No
Susceptible to deterioration	No	Yes
Visual detecton monitoring	Yes	No
Reliable speed measurements	Yes	Yes, with speed trap pair
Stopped vehicle detection over wide area	Yes	Not practical
Wrong way vehicel detection	Yes	Yes with second loop
Vehicle classification	Yes, 3 classes	Yes, with speed trap pair
Spatial occupancy, density measurement	Yes	No
Queue length measurement	Yes	Yes, with added loops
Delay, extend, combine detector outputs	Yes	No
Provide MOE's, stops, delays, etc.	Yes	No
Incident detectors	Yes	Yes, if processed at central location
Visual surveillance capability	Yes	No
Off-line video processing capability	Yes	—

another. For example, an agency with no surveillance camera capabilities would benefit greatly from the surveillance available from WADS, whereas an agency with existing CCTV cameras in place, such as Mn/DOT, would, from the surveillance point of view, benefit only marginally. Using recommended mounting heights of 10 m (30 ft) or more, the top of the camera field of view can typically be set to just below the horizon to provide a surveillance view with detectors in the near field at the bottom of the field of view to maximize detection performance.

CONCLUSIONS

This study demonstrated that when user costs are taken into account, Autoscope is more cost effective than conventional loop detectors, even for sparse detection requirements on a two-lane roadway. Benefit-cost ratios ranging from 1.25 to 18.4 were obtained for three alternative Autoscope configurations on a freeway with two lanes in each direction and when speed was used for accurate assessment of traffic state. As expected, benefit-to-cost ratios are even higher when cost data is extrapolated to three and four lanes in each direction because the multiple lane detection capability of video detection and when speed is used for accurate assessment of traffic state. Although the direct cost to install conventional loops, in most cases, is less than the costs of the WADS alternatives, the total loop cost, including the indirect cost to users because delay from lane closures, is greater than the cost of all three WADS alternatives. Even though this analysis derived user-delay costs for loops installed during the day, there are many sections of roadway that cannot be installed at night without causing significant delays. Cost trade-off between nighttime and daytime loop installation were not part of this study.

Furthermore, as detection requirements grow, the cost of using conventional loop detection will increase. Although loop costs have been minimized over the last 30 years, the cost of video WADS should continue to decline as production levels increase and manufacturing costs are reduced, which will further increase the cost effectiveness for freeway applications. Finally, recurring loop replacement required by road resurfacing will lead to continued direct costs and even greater indirect user costs that only make video detection more favorable. Documented mean time between failures by the CCTV camera manufacturers is in excess of 20 years, which experience has thus far supported.

The cost effectiveness of video WADS will be driven by site- and application-specific requirements. The system planners must weigh all the costs and benefits of WADS versus conventional loop detec-

tion. The cost of installing WADS will be competitive with the cost of installing loops in most cases, except for those in which simple measurement such as volume or occupancy are required at a single or a few points, such as on a freeway ramp. Even where the cost is higher, the intangible benefits and advantages of wide-area detection can justify the additional cost.

Wide deployment of WADS will enable other IVHS traffic management technologies to take root and will eventually lead to more efficient management of traffic, saving time and money and reducing congestion and pollution levels. The development of automatic measure of effectiveness extraction; incident detection; and continuous, real-time performance monitoring that is possible with WADS will greatly aid in the evaluation of traffic management schemes.

ACKNOWLEDGMENTS

The authors would like to thank Patty Bednarz and Terry Haukom of the Mn/DOT for their assistance in the cost comparison and Eil Kwon of the Center for Transportation Studies at the University of Minnesota for his assistance with the KRONOS freeway simulation. Financial support for this cost comparison was provided by the Mn/DOT and the FHWA.

REFERENCES

1. Michalopoulos, P. G., B. Wolf, and R. Benke. Testing and Field Implementation of the Minnesota Video Detection System. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 11–19.
2. Michalopoulos, P. G. Vehicle Detection Through Video Image Processing: The Autoscope System. *IEEE Transactions on Vehicular Technology*, Vol. 40, 1991, pp. 21–29.
3. Michalopoulos, P. G., R. D. Jacobson, C.A. Anderson, and J. C. Barbaresso. Integration of Machine Vision and Adaptive Control in the FAST-TRAC Intelligent Vehicle Highway System Program. *Transportation Research Record 1408*, TRB, National Research Council, Washington, D.C., 1993, pp. 108–115.
4. Klein, L. A. and M. K. Mills. Evaluation of Modern Traffic Detector Technologies for IVHS Applications. *Proc. NATDAC '94, National Traffic Data Acquisition Conference*, 1994.
5. Michalopoulos, P. G., E. Kwon, and J. G. Kang. Enhancement and Field Testing of a Freeway Simulation Program. *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 203–215.
6. Kwon, E., H. Xie, S. Pong, and P. G. Michalopoulos. *Enhancements of the KRONOS Simulation Package for Geometric Design, Planning, Operation, and Traffic Management in Freeway Networks/Corridors (Phase 2)*. Final Report for Minnesota Department of Transportation, 1994.
7. Pine, J. L. ed. *Traffic Engineering Handbook*, 4th Ed. ITE. Prentice-Hall, Englewood Cliffs, NJ, 1992, p. 394.

Publication of this paper sponsored by Committee on Freeway Operations.

Caltrans Interstate 15 Reversible High-Occupancy Lanes: 1994 Status

GEORGE E. GRAY, STUART HARVEY, JOEL HAVEN, AND WILLIAM A. DILLON

This paper details the present status of a 12.9-km (8-mi) two-lane reversible high-occupancy-vehicle facility implemented in the median of Interstate 15 freeway in the city of San Diego, California, in 1988. It covers the background of the project, design, and operation of the resulting facility, the results of a 1991 major project assessment, the present traffic service conditions, special uses of the lanes, traffic data, and an approved congestion pricing pilot program utilizing the lanes. Implementation of this demonstration program is expected to begin during the summer of 1995.

In October 1988, a 12.9-km (8-mi) section of reversible high-occupancy-vehicle (HOV) lanes on Interstate 15 (I-15) in the city of San Diego, California, was opened to traffic. The lanes are situated in the median of eight-lane I-15 and run without intermediate access or egress from the junction of I-15 and State Route 163 (an eight-lane freeway serving the San Diego central business district) to State Route 56. They operate in the southbound direction (inbound) in the morning from 6:00 a.m. to 9:00 a.m., and in the northbound (outbound) direction in the afternoon from 3:00 p.m. to 6:30 p.m. Their use is limited to car pools (two or more occupants), van pools, buses, emergency vehicles, and motorcycles. The geometrics of the lanes showing the a.m. and p.m. configurations are demonstrated as Figures 1 and 2.

Growth of HOV use has been moderate but steady, as shown in Table 1, with the percentage of person trips using the lanes remaining at approximately 20 percent for the past few years, or roughly equivalent to a fifth lane on the existing freeway.

The freeway in this area is being improved operationally with the addition of ramp metering, some HOV ramp meter bypass lanes, and auxiliary lane construction. There are no present plans to provide interim access to the reversible lanes; however, the implementation of the ramp metering should improve present problems of congestion that occur downstream of the end of the facility in both the a.m. and p.m. peak periods. This, coupled with the delays to those using the ramp meters, should provide marginal incentives for increased use of the reversible lanes.

Since late in 1991, the lanes have been used for a variety of tests and special events when not in HOV service. Most of the tests have involved automatic braking systems, whereas much of the special event use has been to conduct bicycling events.

Approval has recently been received to use these reversible lanes as a congestion pricing demonstration under the FHWA newly expanded congestion pricing pilot program established by the Intermodal Surface Transportation Efficiency Act (ISTEA). In 1993, the California Legislature approved and the governor subsequently signed Assembly Bill 713 (Goldsmith), providing legal authority for single-occupancy automobile use of the I-15 reversible lanes to test a premium travel lane concept. A draft work plan for this test program, titled Project Feasibility Tasks, is presented herein as Table 2.

G. E. Gray (retired), Caltrans, 9720 Oviedo Street, San Diego, Calif. 92129. S. Harvey (Traffic Operations), J. Haven (Project Development North), and W. A. Dillon (Planning Studies Branch), California Department of Transportation, 2829 Juan Street, San Diego, Calif. 92110.

BACKGROUND

I-15 has one of the fastest growing traffic volumes in the entire San Diego regional highway network. Figure 3 schematically indicates the 1988 and 1993 traffic volumes of the subject corridor at the limits and at significant intermediate points. The recent recession, along with recent cutbacks in the defense industry, has had considerable adverse impact on the San Diego region. This is an especially large factor in slowing the region's growth, which had been causing the I-15 traffic volumes to increase by approximately 10 percent per year from 1980 to 1990. Since 1990, traffic growth in this corridor has slowed to approximately 1 percent per year, which is still above the area's current population growth rate.

The I-15 freeway was identified in the early 1970s as a corridor for mass transit applications. It is for this reason that all subsequent construction has included a dedicated 21.3-m (70-ft) median for possible future transit use. Early identification of a car pool/transit strategy in this corridor was a key item in pursuit of this project. The HOV strategy, as developed in the early 1980s, was a compromise between a fixed guideway mass transit facility and highway interests.

The subject HOV facility is a segment of the I-15 freeway between several bedroom areas of San Diego on the north ends (Penasquitos, Carmel Mountain Ranch, Poway) and a high-employment area on the south end (Kearny Mesa and downtown San Diego). Another unique feature is that these lanes pass through the Miramar Naval Air Station. This naval base severely constrains alternate route development in this corridor. The reversible lanes provide an attractive alternative for the traveling public.

The construction of the express lanes was accomplished with three separate contracts. First, two contracts were let to modify and construct bridges in late 1984 and early 1985. When they were completed in March 1987, the roadway and control systems contract began. The total cost of these projects was $32 million.

DESIGN AND OPERATION FEATURES

The HOV facility consists of two 3.7-m (12-ft) lanes, with 3.2-m (10.5-ft) shoulders on both sides separated from the main lanes by either New Jersey-type barriers or fencing, where there is an ample median. Figure 4 shows the facility during peak-period operation. The lane entrance-exit geometric features are indicated in Figures 1

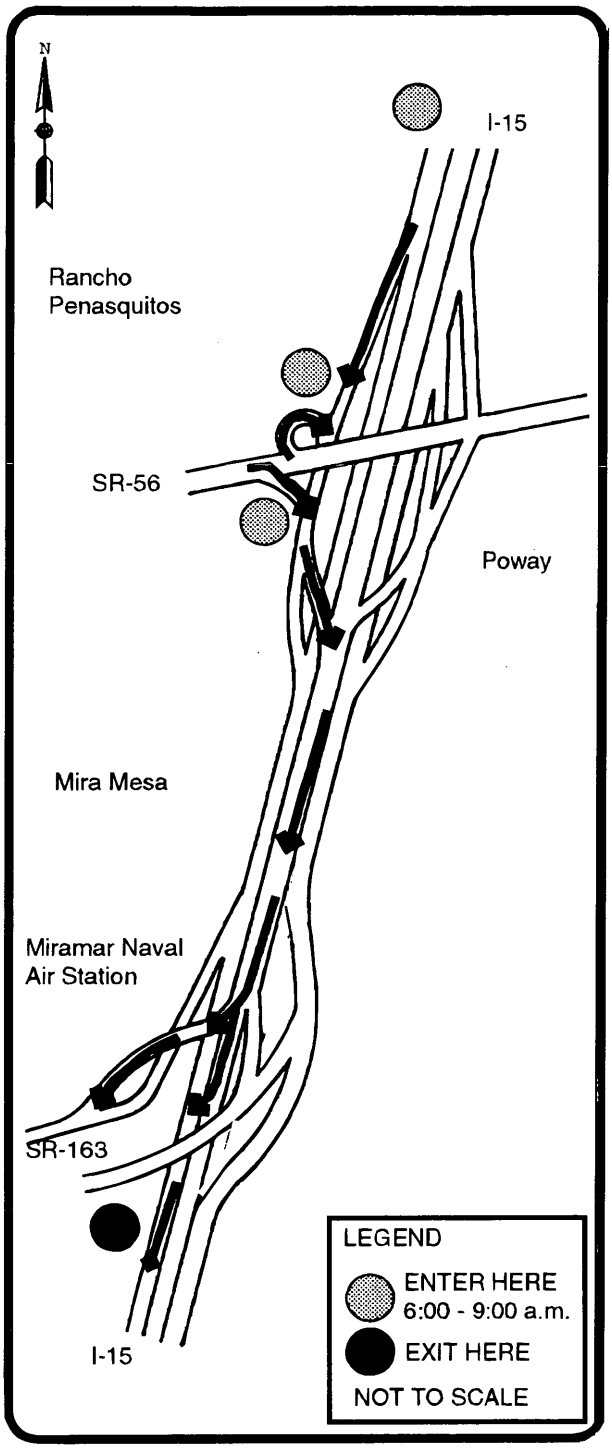


FIGURE 1 A.M. operation of I-15 reversible HOV lanes.

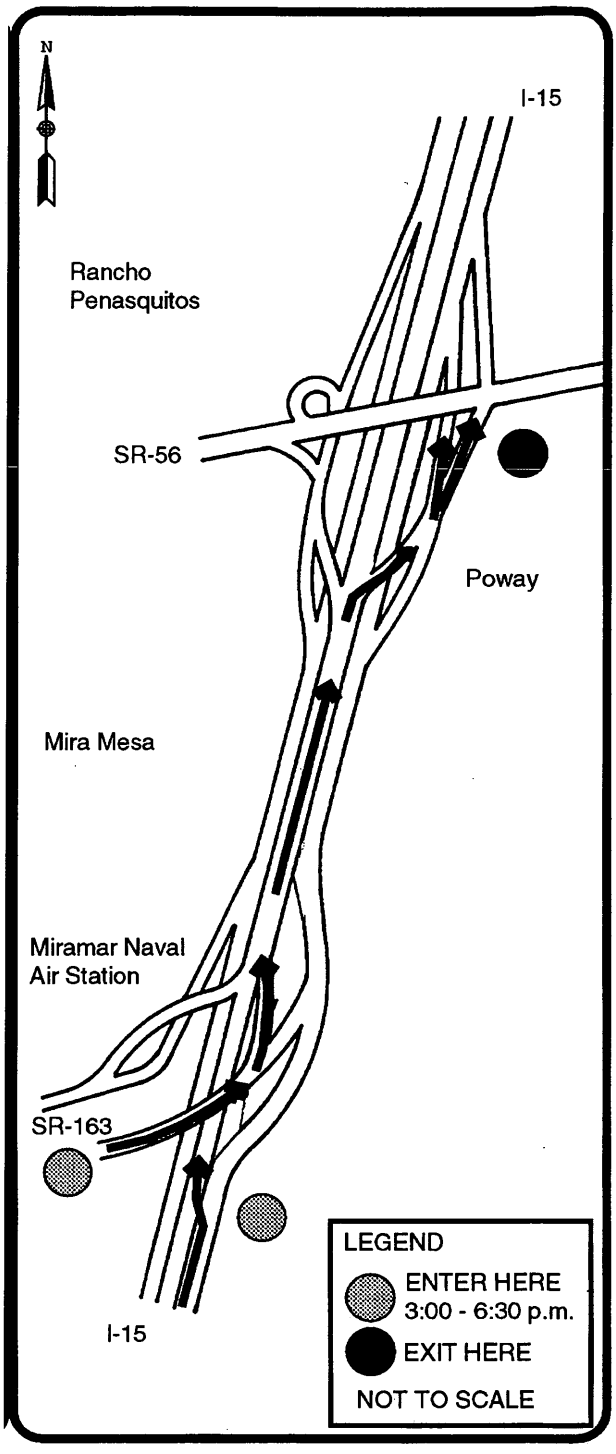


FIGURE 2 P.M. operation of I-15 reversible HOV lanes.

and 2. Figure 5 indicates the direct connector ramps at the northerly end of the HOV facility.

The original plan for the control system design and the development of the control system software was to be completed by consultant contract, because state employees had limited expertise in this field. However, only one bid was received on the contract, and in December 1987, that contractor went bankrupt.

The hardware design was only partially complete and the software barely started. The remaining implementation time was extremely short, and critical deadlines existed because of the state's contractual obligation to furnish the control system as part of the major construction project. A team of state employees was formed to accomplish the task of designing the control system, developing the software, and implementing the system. Various interdependent

TABLE 1 I-15 EXPRESS LANES FACILITY SUMMARY

Month/ Year	Volume Vehicles			Person Trips			
	HOV Peak Hour	HOV[a]	Main Lanes[a]	HOV[a]	Main Lanes[a]	Total[a]	%[b]
Nov. 1988	1,944	4,434	58,269	9,134[d]	64,096[d]	73,230	12.5
May 1993	3,433	8,676	62,295	18,341[c]	68,964	87,305	21.0
July 1993	3,241	8,414	61,958	17,814	68,153	85,967	20.7
Sept. 1993	3,127	8,223	64,444	17,416	70,888	88,304	19.7
Nov. 1993	3,215	8,301	58,448	17,902	64,293	82,195	21.8
Jan. 1994	3,518	8,982	62,436	19,194	68,679	87,873	21.8
Mar. 1994	3,644	9,155	69,286	19,485	76,215	95,700	20.4
May 1994	3,171	8,133	57,866	17,309	63,653	80,962	21.4
Jun. 1994	3,480	8,979	68,220	19,109	75,042	94,151	20.3

[a] *a.m. (0600–0900) + p.m. (1500–1830)*
[b] *HOV lanes as % of total.*
[c] *From May 1993 through June 1994, HOV Person Trips have consistently been twice the value for November 1988, the first full month of operation.*
[d] *The Total Person Trips (HOV + Main Lanes) grew 28.6% between November 1988 and June 1994. The HOV Person Trips grew 109.2% in the same period.*

efforts had to be performed concurrently, where normal work flow would have dictated sequential completion of dependent tasks. The tight schedule was maintained, and the lanes were opened to traffic on October 20, 1988.

Opening-closing devices and emergency features were built into the project as follows:

1. Overhead flip disc changeable message signs. Four signs are upstream of each entrance location, displaying a variety of messages. Figure 6 indicates one of these changeable message signs.
2. Several rows of pop-up delineators. These are operated pneumatically, with positive pressure necessary for raising or lowering the delineators. There are also loop detectors located to gather vehicle speeds and volumes on the main lanes, as well as on the HOV lanes. Additional detectors are placed in advance of the entrances to detect a gap in traffic before raising the pop-ups and closing an entrance.
3. Standard luminaires. These draw lights illuminate the entrances and exits when they are open, and remain dark when they are closed.
4. Semaphore bridge gates. The gates look like railroad gates, but do not have the breakaway features. They lock into the barrier rail and consist of two 1.3-cm (0.5-in.) aircraft carrier cables surrounded by an aluminum shell. These gates are closed at an entrance when the express lanes are open in the opposing direction and, as of this report, have not been hit in the 8 years of operation. Figure 7 indicates a barrier gate and a row of pop-up delineators.

The emergency features consist of the following:

1. Call boxes placed at 0.8-km (0.5-mi.) intervals. They were financed on all freeways in San Diego County by additional vehicle registration fees. These telephones allow a disabled motorist to contact the California Highway Patrol (CHP) and obtain assistance.
2. Motorcycle openings or removable guardrail openings at approximately 1.6-km (1-mi.) intervals. The 1.8-m (6-ft) motorcycle openings in the barrier have crash-cushion protection for barrier end sections in both directions. This allows CHP motorcycle ingress or egress from the facility for enhanced enforcement and quicker emergency response. The removable guardrail openings provide a 3.8-m (12.5-ft) opening by simply loosening wing nuts. This can be done in approximately 1 min, and permits entrance and exit by larger emergency vehicles.

There are five field locations at which opening-closing devices are operated. At each, there is a device control unit (DCU) for the devices within the general vicinity. The DCU consists of a Versa Module Eurocard bus M68000 microcomputer mounted in a modified Type 334 traffic control cabinet. The DCUs control barrier gates, pneumatic pop-up traffic delineators, and entrance-exit lighting. They control roadway opening-closing sequencing and maintain total system status based on feedback from gate arm position switches, pressure limit switches, analog pressure sensors, and vehicle detectors. Each DCU location also has a manual control unit that houses manual controls for emergency and maintenance operations. Air reservoirs provide temporary redundancy for operation during compressor or air delivery line failures.

TABLE 2 Project Feasibility Tasks

	Task	Responsible Agency[a]		MM[b]
		SANDAG	CAL-TRANS	
1	Determine if project violates state or federal laws or regulations.		X	3
2	Determine how project affects goals and objectives of HOV program.		X	3
3	Coordinate proposal with federal and any other agencies who might have interest in proposal.		X	3
4	Conduct surveys of existing and possible users in corridor to determine their attitudes to congestion pricing.	X		4
5	Determine long-term effects which the proposal will have on regional HOV programs.	X		4
6	Determine transit improvements resulting from implementation.	X[c]		4
7	Determine federal and state design and safety requirements necessary to implement project.		X	4
8	Determine long-term impacts which congestion pricing might have on HOV operations at other locations.		X	4
9	Make preliminary investigations on the aspects of access, egress, merging, toll collection, and enforcement.		X	5
10	Investigate possibility of truck usage.		X	5
11	Determine capacity of HOV lanes to ensure they can operate at LOS B or better after implementation of project. (Include merging conflicts at north and south ends.)		X	5
12	Make preliminary estimates of the amount of the user's fee for single occupant vehicles.	X	X	5
13	Determine the cost/benefit of proposal.	X		5
14	Determine the amount and source of any additional funds above those collected which may be necessary to implement, operate, and maintain project.	X		5

(contineud on next page)

TABLE 2 *Continued*

	Task	Responsible Agency[a]		M[b]
		SANDAG	CAL-TRANS	
15	Make a preliminary policy decision to proceed or abort project.		X	5
16	Determine how fees will be collected.	X	X	6
17	Determine method of enforcement.[d]		X	6
18	Determine cost of enforcement.[d]		X	5
19	Determine operating strategy.		X	6
20	Determine other facilities in the state where this strategy could be implemented if it proves feasible.		X	5
21	Identify improvements necessary for the main lanes, toll areas and the merging and converging areas near the north and south access points.	X	X	6
22	Develop a method to measure and monitor changes in level of service for HOV and mixed-flow lanes resulting from project.		X	6
23	Determine methods necessary to measure improvements in air quality and fuel savings resulting from project.		X	6
24	Determine how and when program will be terminated if necessary.	X	X	6
25	Conduct a public hearing on proposal.	X	X	7
26	Develop a marketing strategy for all users in corridor.	X	X	7
27	Prepare an environmental document if required.	X	X	7
28	Make final policy decision to proceed or abort project.		X	7
29	Final report.	X	X	9

[a] *San Diego County Regional Transportation Commission (SANDAG) or California DOT (CALTRANS).*
[b] *Number of months from initiation to completion.*
[c] *In cooperation with Metropolitan Transit Development Board.*
[d] *In cooperation with California Highway Patrol.*

Three DCUs are connected to the south-end control building's field control unit (FCU), which is a computer system similar to the DCU. The FCU supervises the three DCUs and the eight changeable message signs on the south end via a serial line communications system over twisted pair cables and line drivers. Each DCU sends complete device status information to the FCU every 15 sec and whenever a device status change is detected. The changeable message signs are polled for correct status every 6 mins.

Two DCUs are connected to the north-end control building's FCU, which serves a similar function as the south-end FCU. There are four changeable message signs at the north end. Both control buildings house compressors with filter systems, pressure tanks, automatic backup diesel generators, and power distribution and ground fault interruption equipment. In addition, the north-end control building has a maintenance and spare parts storage facility.

The two FCUs are connected via leased dedicated telephone lines and modems to a traffic system unit (TSU) in the Transportation

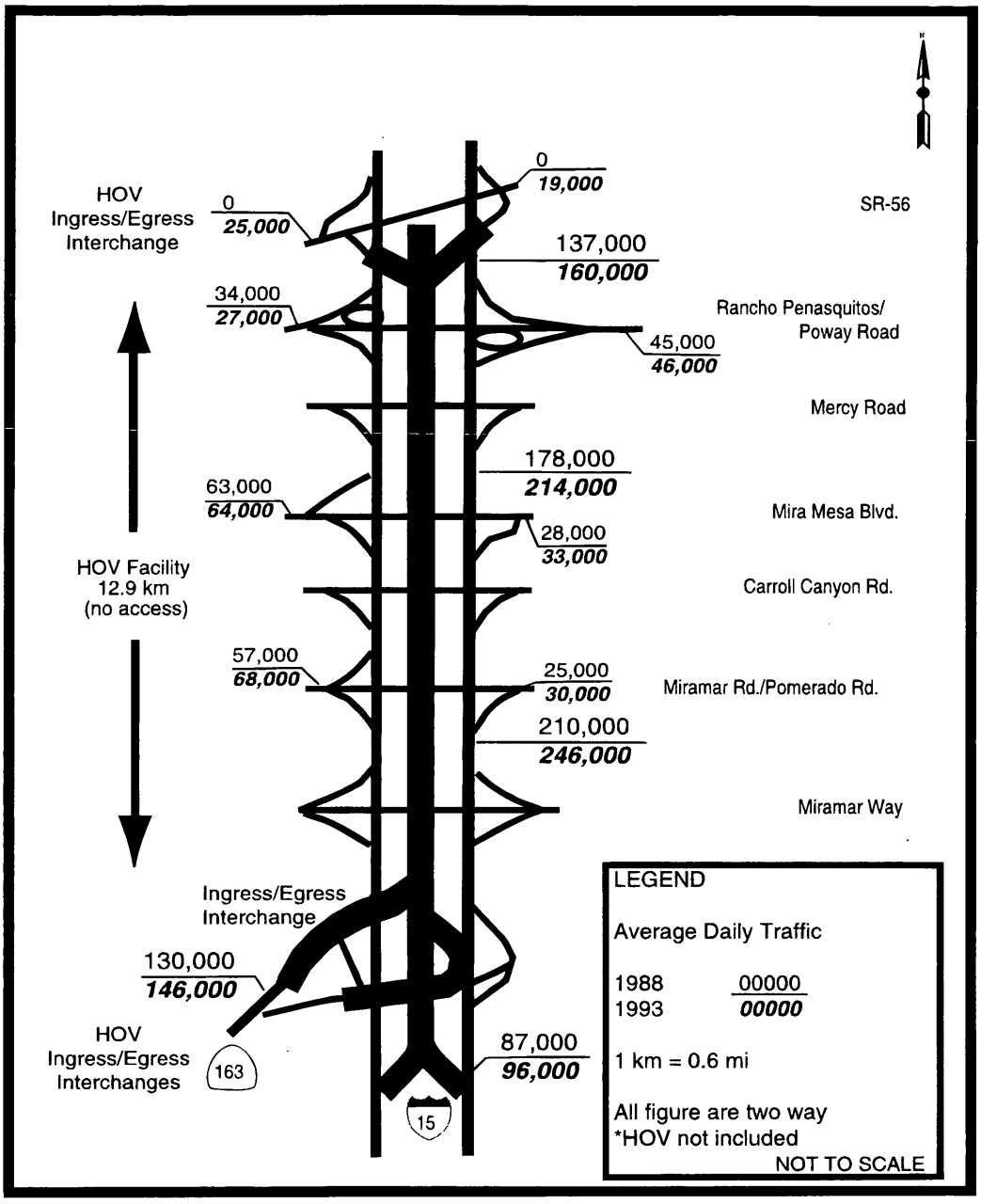


FIGURE 3 I-15 HOV schematic corridor traffic volumes and interchanges.

FIGURE 4 HOV during operation. Reversible HOV facility in median. Northbound traffic Level of Service (LOS) F at p.m. peak period near southerly entrance.

Management Center in the Old Town area of San Diego, which is approximately 16 km (10 mi) from the south end of the HOV lanes. The TSU computer is similar to the other VME bus computers in the system. The system is controlled completely by an operator at this location.

The software for each of the eight microcomputers in the control system is identical. On start up, the software performs hardware validation checks and identifies which unit is in the system. The operating system is a small real-time multitasking operating system called OSENGINE. Most of the application software is written in C language. Communications drivers and device drivers were written in 68000 ASSEMBLY language. All of the application software was developed under the UNIX operating system.

For roadway operation, the operator utilizes prompted English

FIGURE 5 Direct connector ramps at northerly end of HOV facility. State Highway 56 interchange is at top.

language commands (e.g., closure of northbound I-15 entrance is effected by entry of the command "Close south-end 15 entrance"). A security system that assigns specific security levels to each operator is incorporated. The security levels govern access to the various types of commands available in the system. A system error log is maintained.

FIGURE 6 Changeable message sign.

FIGURE 7 Entrance ramp to HOV. This is northbound entrance from State Highway 163. It indicates hydraulic pop-ups and barrier gate. At this time, HOV facility is closed. When access is allowed in this direction, pop-ups are down. When HOV is open in opposite direction, pop-ups are up and gate is down.

Great care was taken in the design and development of the system to ensure safe operation. The software prevents an operator from operating the devices in an improper sequence, such as lowering a barrier gate without the corresponding pneumatic pop-up delineators in place. Each computer in the system monitors the status of all closure device sensors in the system, and prevents operation of closure devices unless complete system status is known and proper.

The performance of the reversible roadway communications system under marginal line conditions is currently being improved, along with mechanisms to improve the operation and safety of the system under conditions of sensor malfunction.

Further modifications to the system are proposed, incorporating the reversible roadway control system into a new and expanded TMC. The modifications would integrate the reversible roadway control system with the new TMC computer systems and make the twelve changeable message signs in advance of the north-end and south-end entrances available to the TMC for other transportation management purposes.

1991 PROJECT ASSESSMENT

Field work for a report based on extensive analysis undertaken over 3 1/2 years by the Department of Civil Engineering, San Diego State University, was finalized in 1990 (*1*). The study, published in 1991, is reported in seven parts as follows: Part 1, executive summary; Part 2, volume/occupancy study; Part 3, speed/delay study; Part 4, land use study; Part 5, park and ride study; Part 6, bus study; and Part 7, attitudinal study.

The major findings of the assessment were:

- The lack of interim access/egress to the reversible lanes has caused some reverse travel to utilize the facility.
- The 2+ occupants per vehicle HOV designation is appropriate.
- The decision to build the HOV lanes as a reversible facility is justified, as the minor direction of travel during the peak periods is only approximately 60 percent of the volume of the major travel direction.

- During the first 2 years of the HOV lane operation, a conservative $30 million was saved through decrease in delay in the corridor.
- The then existing park-and-ride facilities were not adequate. In addition, only two of the eight facilities between Route 78, 17.7 km (11 mi) north of the project, and Mira Mesa Boulevard were judged well placed to serve the HOV lanes.
- Bus and van pool use was found to be relatively low for such a facility, as exemplified by the overall occupancy rate of 2.3 passengers per vehicle.
- The March 1991 bus survey indicated an average increased ridership of 53 percent over the first 3 months of service on the facility.
- Attitudes toward the HOV lanes were found to be "strongly positive" or "positive," but declined slightly from 1988 (77%) to 1990 (70%).
- The most common complaints were: "Lanes should be open to all traffic" and "I would like to use the lanes, but I cannot access them."
- A survey of new homeowners in the area identified that over 22 percent indicated that their home purchase was related to the availability of the HOV facility.
- The average car pooler believed the HOV lanes saved 22 min and approximately $3 per day. The observed time savings at the present time is approximately 15 min for the round trip during the most congested time frames. The cost savings is subjective and is perceived very differently by the users.
- Average speed on the mixed-flow lanes increased from 61 kph (38 mph) pre-opening to 90 kph (56 mph) in 1990 (*1*).

This assessment report concludes that the primary factors contributing to the success of the HOV express lanes are:

- The HOV lanes were added, not taken away from the main lanes.
- The relaxed car pool definition (2+ persons per vehicle) is very helpful at the current stage of the facility use. The fraction of vehicles with 3+ persons is still rather small. Contributions from buses and cars with 3+ passengers are growing fast, but are still relatively low.
- The technical performance of the HOV facility has been excellent, and the system is enjoying a positive public image.
- As Level of Service A is virtually always offered, the reward for using the facility is well defined and reliable.
- The HOV facility is long enough, 12.9 km (8 mi), for commuters to notice the advantages of use.
- This solution to increasing HOV commuter traffic is compatible with transportation and environmental policies, and is considered right for the region by the vast majority of the population.
- The lanes have received mostly positive media attention (*1*).

PRESENT TRAFFIC SERVICE CONDITIONS

At the present time, the reach of I-15 served by the reversible HOV lanes is a full eight-lane freeway, with auxiliary lanes at several locations and operating ramp meters at all of the southbound ramps and most of the northbound ramps. Ramp meter HOV bypass lanes exist at many of the above southbound lanes, but they are also controlled by the metering. Therefore, some of the advantage of the bypass is lost, especially for buses, which often must come to a stop at the meter.

The provision of well-placed and well-sized park-and-ride lots offers an opportunity to increase the percentage of car pools and transit riders in the corridor. Figure 8 identifies the existing park-and-ride lots within the subject portion of the I-15 corridor. The limits shown are State Route 52, 1.6 km (1 mi) south of the southern end of the HOV lanes, and State Route 78, 18 km (11 mi) north of the northern end of the HOV lanes. The lots shown are those within 1.6 km (1 mi) of I-15. However, the lack of intermediate access to the HOV lanes negates or reduces the value of several of the lots. Four lots are located between 4.8 and 18 km (3 and 11 mi) north of the HOV facility.

The following chart shows the park-and-ride facility regional identification number, the number of parking spaces, the percentage of occupancy, and the availability of local or express bus service. The lots have been grouped using a subjective evaluation of their accessibility to the HOV lanes.

Good Access to HOV [located upstream (northerly) of HOV facility]

Regional Lot No.	*No. of Spaces*	*Percent Occupied*	*Served by Local Bus*	*Served by Express Bus*
3	20	100	Yes	No
11	140	34	Yes	No
54	200	39	Yes	Yes
65	16	100	Yes	No
Subtotal	376			

Reasonable Access to HOV (near upstream end of HOV facility)

Regional Lot No.	*No. of Spaces*	*Percent Occupied*	*Served by Local Bus*	*Served by Express Bus*
4	104	30	Yes	Yes
18	103	25	Yes	Yes
26	125	30	Yes	Yes
31	67	95	Yes	Yes
53	93	61	Yes	Yes
57	132	19	Yes	Yes
Subtotal	624			

Poor Access to HOV (out of travel direction)

Regional Lot No.	*No. of Spaces*	*Percent Occupied*	*Served by Local Bus*	*Served by Express Bus*
16	103	39	Yes	Yes
51	44	48	Yes	Yes
Subtotal	147			

No Access to HOV

Regional Lot No.	*No. of Spaces*	*Percent Occupied*	*Served by Local Bus*	*Served by Express Bus*
6	221	33	No	Yes
58	44	77	Yes	No
Subtotal	265			

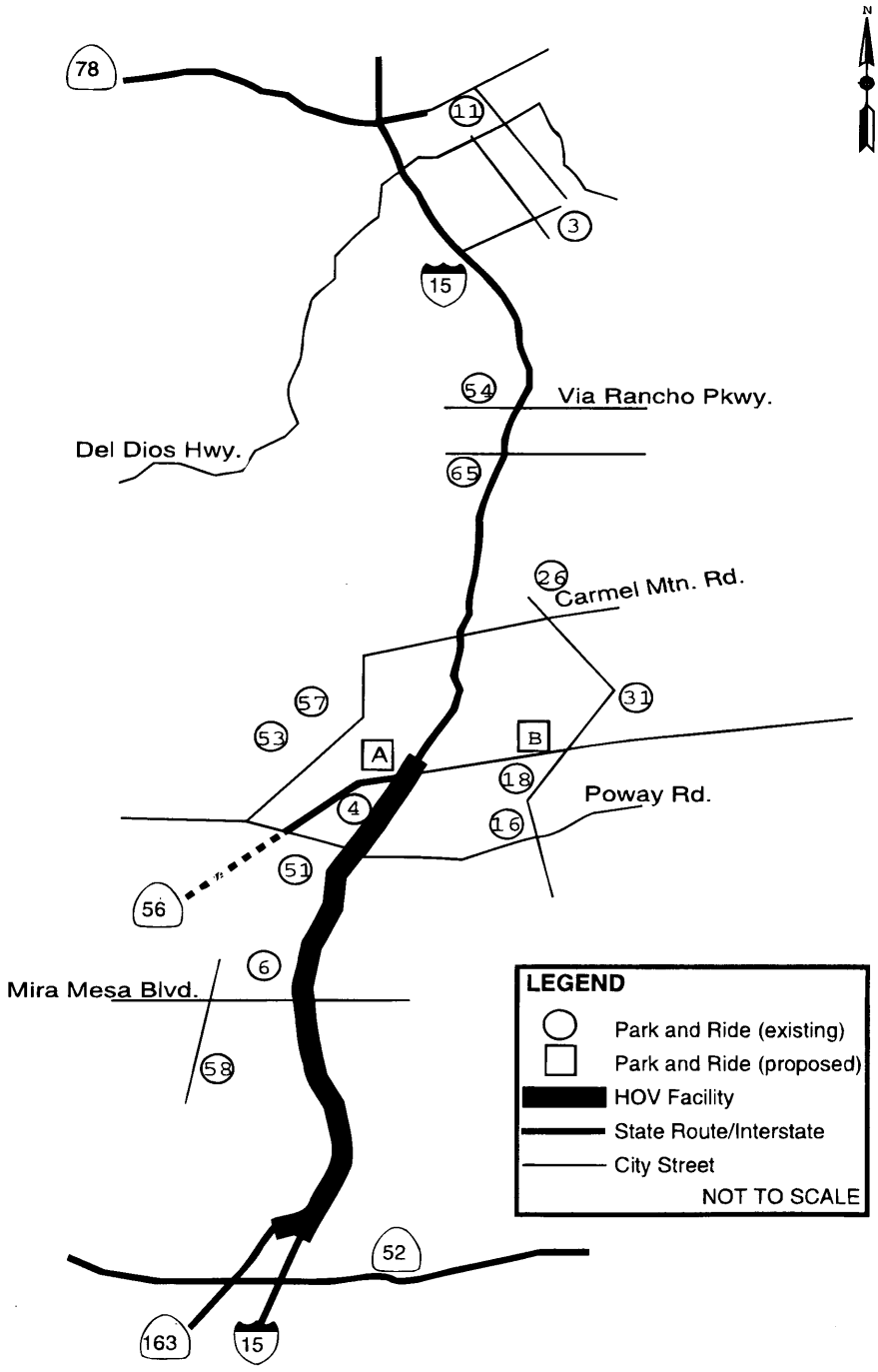


FIGURE 8 Park-and-ride location I-15 corridor.

The project report that was prepared before the construction of the HOV lanes identified two park-and-ride locations to be built as part of the HOV project. These locations are identified in Figure 8 as Sites A and B. However, funds were not programmed for these facilities. Lot A was to be sited within excess freeway right of way, and would have provided 250 spaces with excellent HOV access. Lot B would have provided reasonable HOV access, but required the purchase of right of way. This lot was planned to provide 400 spaces. A smaller lot, No. 18, with 103 spaces has been built at this location, and opened in January 1995. Undeveloped land is no longer available to expand this location.

Bus service along the corridor is nominal. San Diego County Transit System (CTS) operates six southbound express bus trips from Escondido to downtown San Diego. These trips use the HOV lanes in the morning peak period and return northbound on the HOV lanes during the afternoon peak period. This route serves park-and-ride Lot Number 54. CTS also provides four southbound express trips from Poway to downtown San Diego, using the HOV lanes and

returning on the HOV lanes during the evening. This route serves park-and-ride Lots 16 and 18.

San Diego Transit (SDT) also provides peak-period express bus service on the HOV lanes. Six morning trips are provided from the North County Fair shopping mall in Escondido (adjacent to park-and-ride Lot 54) to downtown San Diego. Return trips are provided on the HOV during the evening peak period. Service is provided to park-and-ride Lots 26, 31, and 54. SDT also provides six southbound express bus trips from Rancho Penasquitos to downtown San Diego, using HOV lanes during the morning and returning on the HOV lanes during the evening peak period. This route serves park-and-ride Lots 4, 53, and 57. SDT provides all-day express bus service on one additional route that operates on the mixed-flow lanes of I-15. This route serves park-and-ride Lots 4, 6, 51, 53, 54, and 57.

Greyhound Bus Lines uses the HOV lanes during the evening peak period for two intercity routes going northbound to Riverside County and beyond.

Local bus routes operated by North County Transit, SDT, and CTS provide feeder service to the park-and-ride lots, as indicated in the previous charts. They also provide transfer opportunities at various bus stops shared by the express bus and local bus routes.

The project report identified that the exclusive reversible HOV two-lane facility would provide needed increased freeway capacity in the peak direction of traffic and would encourage increased ride sharing, thus decreasing total vehicle demand. Both of these expectations have been at least partially fulfilled. However, as the adjacent freeway lanes have not reached the projected level of congestion, incentive to use the HOV facility has not reached the expected level. Two factors have kept the present congested level lower than projected: the overall reduction in traffic growth attributed to the present economic slump and the effects of improved freeway operations resulting from ramp metering and recent auxiliary lane construction.

SPECIAL USES

One unanticipated use of the facility has been as a testing facility for automatic braking systems and for video cameras.

Because the express lanes are a long, high-standard 112-kph (70-mph) design speed barrier-separated facility, they are an ideal test site for various new technology applications. The following represent some of the testing and research activities that have utilized the express lanes:

- Radar-controlled collision avoidance systems.
- Vehicle performance evaluation on wet pavement.

Automatic lateral guidance systems.

- Laser-controlled lane stripe maintenance.
- Close circuit television testing.

These tests (many of them are ongoing) have been sponsored by a variety of private and academic institutions working in cooperation with Caltrans.

TRAFFIC VIOLATIONS, ENFORCEMENT, AND ACCIDENT RATE

Since opening in 1988, the average vehicle occupancy in the express lanes has been approximately 2.2 persons per passenger vehicle. If public transit bus ridership is included, then the occupancy rate increases to 2.3 persons per vehicle.

The California Highway Patrol reports that its task of enforcing the minimum vehicle occupancy rate is relatively simple for the following reasons:

- The 12.9-km (8-mi)-long express lanes have no intermediate points of egress or access. Once a vehicle enters the system it must remain there in clear view of officers for the entire trip. Relatively low-lane density makes it easy to spot violators.
- Although most enforcement effort is concentrated at the points from which vehicles leave the system, the generous roadway cross section makes enforcement easy and safe at any point.
- The penalty for receiving a citation is substantial. In 1994, the fine was approximately $270, and the citation counts against the driver's record as a safety infraction (failure to obey an official sign).

For these reasons the violation rate in the express lanes is remarkably low. Manual counts show a violation rate of approximately 1.5 percent of total express lane traffic.

Operation of the express lanes has been almost accident-free. From opening day in October 1988 through mid-1994, the facility was operated approximately 1,500 days. In that time, there have been only eight recorded accidents. The accident rate is 0.07 accidents/million vehicle-kilometers (MVkm) (0.12 accidents/million vehicle-miles (MVM). By comparison, the adjacent freeway mixed-flow lanes have an accident rate of 0.4 accidents/MVkm (0.66 accidents/MVM) during the same hours in which the express lanes are open.

A special feature of express lanes operation is that they may be opened to all traffic under specified emergency conditions when freeway main lanes may be blocked by an accident. This bypass feature has been brought into play approximately once per year, when accidents cause two or more of the adjacent mixed-flow lanes to be blocked for an extended time during the commute peak period.

CONGESTION PRICING PILOT PROGRAM

In late January 1993, the regional planning agency, the San Diego Association of Governments (SANDAG), submitted an application (2) for participation in the FHWA ISTEA-established congestion pricing pilot program. Earlier, in 1991, SANDAG applied to the Federal Transit Administration (FTA) for a transit development and congestion pricing demonstration grant. Since approval of the FTA grant request in late 1992, public and private interest in a congestion pricing demonstration program utilizing the I-15 reversible lanes has grown.

In 1993, state legislation was introduced and subsequently passed into law to allow implementation of roadway pricing on this facility if it were to be designated as a federal congestion pricing pilot program. This legislation stipulates a number of conditions, including that the program apply only to the I-15 reversible lanes; that the resulting available revenue be used for transit; and that, although single-occupant vehicles are authorized to use the high-occupancy lanes for a fee, this not reduce the use or access of the lanes by HOVs.

The SANDAG approach to congestion pricing in general is based on the following principles:

- It is a tool by which to achieve wide regional objectives, such as traffic congestion relief, improved air quality, and improved mobility.
- It should be implemented in stages, each of them based on technical analysis, public involvement, and political acceptance.
- Each stage must be a balance between what is technically and theoretically desirable and what is politically feasible (*2*).

The specific objective of the I-15 pilot program was identified as to promote maximum utilization of the HOV lanes and to reduce corridor congestion through the use of a market approach providing a premium price for single-occupancy vehicle (SOV) use, based on time savings and replacement costs associated with the use of uncongested HOV facilities.

The 1993 SANDAG grant request was denied by the FHWA in February 1994 on the basis that it did not qualify because it was an HOV buy-in project to allow SOV use of the HOV lanes. At that time, this concept was not allowable. The rejection letter, however, announced the extension of the solicitation for projects and a program to fund development of potential pricing projects (G.J. Jeff, unpublished data. Then the FHWA, by *Federal Register* notice of May 25, 1994 (*3*), expanded and liberalized the program to the point that by letter of June 24, 1994, SANDAG resubmitted its original proposal and proposed that federal congestion pricing pilot program funds be used to support the initial engineering, technology, identification of possible transit enhancement, and evaluation of the pilot project. Furthermore, SANDAG proposed that the detailed budget and matching funds determination be subject to negotiation of a cooperative agreement.

The original SANDAG proposal lacked the way very important planning stage needed to guide the decision to implement the program and the details of such an implementation. Consequently, SANDAG and Caltrans have worked together on a plan to address issues and procedures. The resubmitted application has recently been approved. The pilot program will be implemented in two stages over a period of 3 years.

The first stage would involve low-level technology using prepurchased identification allowing drive-alone users to enter the HOV facility without stopping to pay a cash toll.

In the second stage, an electronic toll collection and traffic management system would be utilized. Extensive testing and use of such strategies as automated vehicle identification and electronic toll and traffic management concepts could be involved in this second phase.

The work plan is presently being finalized and is expected to address the 29 tasks included in Table 2. The final work plan for this program is being developed cooperatively by SANDAG and Caltrans, and will be approved by both agencies.

FUTURE POSSIBILITIES

Besides the planned implementation of the first phase of the proposed congestion pricing pilot program in the summer of 1995, a variety of related studies involving this portion of I-15 are either underway or proposed. These studies include:

- An ISTEA Section 1005 study of the economic lifeline corridor.
- A study of possible upgrading of transit services between Escondido and San Diego, with several options being considered, including light rail.
- A Caltrans regionwide update study of the HOV/ramp meter system.
- A Caltrans review of the park-and-ride system throughout the area.

Depending on the results of these various studies, actions will be taken to further enhance the effectiveness and efficiency of the subject current reversible HOV facility.

REFERENCES

1. Supernak, J. S. *Assessment of the Effectiveness of the Reversible Roadway for High Occupancy Vehicles on Interstate Route 15. Part 1: Executive Summary.* Final Report. Department of Civil Engineering, San Diego State University, San Diego, Calif., May 1991.
2. *Application for Participation in the Federal Highway Administration ISTEA-Established Congestion Pricing Pilot Program.* San Diego Association of Governments, San Diego, Calif., Jan. 1993.
3. Federal Highway Administration, U.S. Department of Transportation. Participation in the Congestion Pricing Pilot Program. *Federal Register,* No. 100, May 1994, pp. 27098–27099.

Publication of this paper sponsored by Committee on High-Occupancy Vehicle Systems.

Evaluation of Minnesota I-394 High-Occupancy-Vehicle Transportation System

ALLAN E. PINT, CHARLEEN A. ZIMMER, JOSEPH J. KERN, AND LEONARD E. PALEK

Construction of the major elements of the I-394 transportation system was completed in fall 1992. This system includes high-occupancy-vehicle (HOV) reversible lanes, concurrent HOV lanes, three car pool parking garages in downtown Minneapolis, five transit stations, seven park-and-ride lots, improved transit service, an extensive automated traffic management system, and a number of other HOV-related services, including enforcement and marketing. The Minnesota Department of Transportation began a three-phase evaluation of the I-394 system in 1985, when an interim HOV lane was constructed. The first phase of evaluation addressed the effectiveness of the interim HOV lane before major construction. The second phase looked at HOV use and operation during construction. The third phase has been underway since the facility was completed in late 1992. This evaluation effort includes data collection related to bus ridership, vehicle occupancy, traffic volumes, occupancy compliance, park-and-ride use, garage use, and travel times. The results of the I-394 evaluation will be used to develop incentive programs to encourage further car pooling and bus ridership in the I-394 corridor, to fine tune operational elements of the system to provide safer and more efficient traffic flow, and to provide guidance for the development of other HOV facilities in the Twin Cities metropolitan area.

I-394 was the first Interstate project in the United States to fully integrate the funding and construction of highways with the construction of high-occupancy-vehicle (HOV) lanes, transit facilities, parking garages, and elevated and enclosed walkways called skyways. These physical features are supported proactively by a myriad of programs including transit and ride-share services, traffic management systems, enforcement programs, parking incentives, and public information and marketing activities. The intent of the I-394 transportation system, as this unique combination of physical facilities and programs has come to be known, is to maximize the number of people carried by aggressively encouraging car pooling and bus ridership in a heavily congested highway corridor.

The Minnesota Department of Transportation (Mn/DOT) completed construction of I-394 in fall 1992. I-394 is an 11-mi facility that fully integrates transit and highway systems and is designed to maximize incentives for HOVs, including buses, car pools, and van pools. HOVs are defined as vehicles with two or more people. The key components of the I-394 transportation system are illustrated in Figure 1 and include the following:

- Three miles of reversible HOV lanes.
- Eight miles of concurrent-flow HOV lanes.
- An automated traffic management system.
- Eight HOV meter bypass lanes.
- Three parking garages in downtown Minneapolis that have direct access to/from I-394, reduced parking fees for car poolers, and transit stations.
- Elevated and enclosed walkways called skyways connecting the downtown Minneapolis parking garages to each other and to the downtown Minneapolis skyway system.
- Three additional transit transfer stations.
- Seven park-and-ride lots.
- Expanded express and timed-transfer local bus service.
- Ride-share matching.
- Enforcement activities.
- An extensive marketing program to increase car pooling and transit ridership.

U.S. Highway 12 between Wayzata and downtown Minneapolis was designated initially to be upgraded to Interstate standards in 1968. It was a politically and publicly controversial project that encountered several stumbling blocks and milestones before construction began in 1984.

I-394 INTERIM HIGH-OCCUPANCY-VEHICLE LANE

On November 19, 1985, the Minnesota Department of Transportation opened the I-394 interim HOV lane. Initially, the HOV lane was a physically separated, single reversible lane in the median of Trunk Highway (TH) 12, a four-lane, signalized highway. It evolved through several combinations of the separated HOV lane and concurrent HOV lanes over the following 7 years of construction, transitioning ultimately into the permanent I-394 transportation system.

The interim HOV lane was built in 1985 to (*a*) introduce the concept of an HOV lane in advance of the permanent HOV lanes, (*b*) generate public support for car pooling and bus ridership, and (*c*) provide additional people-carrying capacity during the reconstruction of Highway 12 into I-394.

Because of the uncertainty surrounding public acceptance and use of the HOV lane, Mn/DOT and its I-394 Policy Committee and Corridor Management Team made a decision to review periodically the use and operation of the I-394 system. Four distinct time periods were established for evaluation:

1. First year: first year of operation of the interim HOV lane (1986).
2. Construction: period during which the interim HOV lane was affected by the construction of TH 12/I-394 (1987–1992).
3. Start up: first 18 months after completion of construction (1993–1994).

A. E. Pint and L. E. Palek, Minnesota Department of Transportation, Waters Edge Building, 1500 West County Road B2, Roseville, Minn. 55113. C. A. Zimmer and J. J. Kern, Strgar-Roscoe-Fausch, Inc., One Carlson Parkway North, Suite 150, Minneapolis, Minn. 55447.

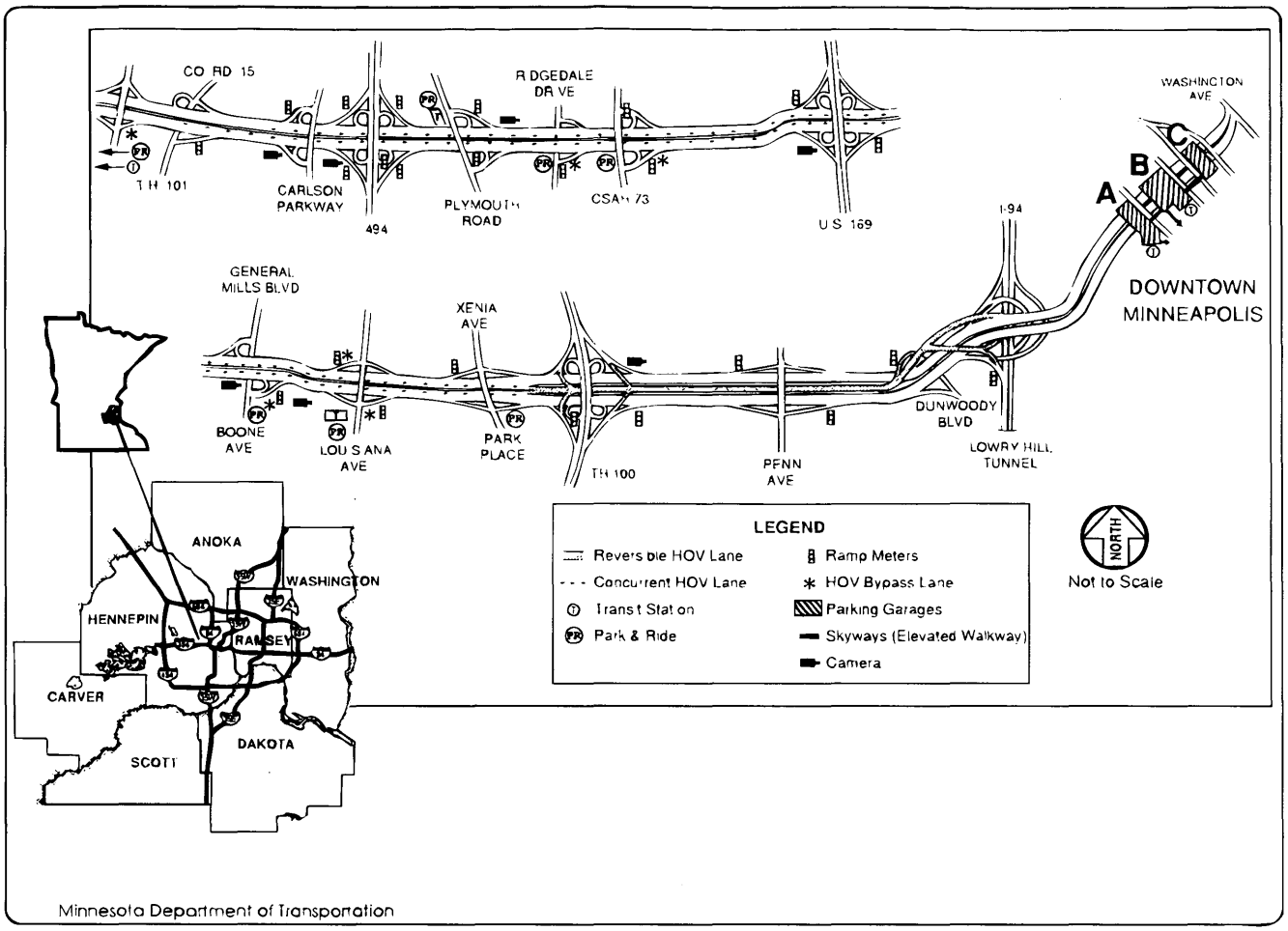


FIGURE 1 I-394 HOV transportation system.

4. Stable operation: ongoing operation of permanent system to initial forecast year (1995–2000).

I-394 CASE STUDY EVALUATION

I-394 was one of the first HOV facilities in Minnesota and included some unique features, particularly the downtown car pool parking garages. Therefore, it has been the subject of a case study evaluation since 1984, when base line information was collected on old TH 12 and parallel roadways before construction. The case study evaluation involved three of the previously described time periods:

Phase I, which was completed in October 1987, analyzed the effectiveness of the interim HOV lane during its first year of operation. The results of the Phase I evaluation were documented in the Phase I report published in October 1987.

Phase II was conducted midway through the construction of I-394. It focused on the effectiveness of the interim HOV lane during the construction period and the problems associated with keeping the HOV lane operational during roadway construction. A Phase II report was published in July 1990.

Phase III was completed in late 1994. Its purpose was to examine the operation and use of the roadway with all of the system elements in place, to identify any operational problems, and to recommend changes and fine tuning to improve the operation of the roadway and the effectiveness of the HOV facilities and programs.

PERFORMANCE OBJECTIVES

Several performance objectives were established in 1984 as part of the original I-394 transportation system management plan. These initial objectives for I-394 were to:

- Increase the peak-hour car pool/van pool modal split for the I-394 corridor.
- Increase the peak-hour transit modal split for the I-394 corridor.
- Improve the level of service for car pools and van pools on I-394.
- Improve the provision of transit service in the I-394 corridor.
- Maintain or improve the existing level of service for mixed traffic on I-394.
- Decrease the accident rate along I-394.
- Achieve and maintain high-occupancy compliance in the I-394 HOV lanes.
- Construct a cost-effective HOV facility on I-394.

A comparison of these stated performance objectives and actual performance to date is provided in Table 1. In general, the use of the

TABLE 1 Performance of I-394 HOV Transportation System

Objectives[1]	Performance Measures[1]	Actual in 1984[1]	Objectives for Start Up Period (1993-1994)[1]	Actual in April 1994[2]	Forecast Conditions (2000)[3]
Increase Peak Hour Car pool/Van pool Usage	Car pools/Van pools	535	1,075	1,596	1,585
	Car pools as Percent of Autos	19%	25%	22%	29%
	Auto Occupancy Rate	1.15	1.30	1.29	1.6
Increase Peak Hour Bus Usage	Buses	21	42	70	57
	Bus Ridership	1,000	2,000	2,256	2,700
Improve the Level of Service for Car pools/Van pools	Average Peak Hour Speed	47 km/hr	69 km/hr	89 km/hr	89 km/hr
	Travel Time from TH 101 to Downtown[4]	23 min	12 min	11.8 min	12 min
Improve Provision of Bus Service in I-394 Corridor	Average Express Bus Speed	47 km/hr	81 km/hr	89 km/hr	81 km/hr
	Travel Time from TH 101 to Downtown[4]	25 min	13 min	11.8 min	13 min
Maintain Improve the Level of Service for Mixed Traffic on I-394	Average Peak Hour Speed	47 km/hr	63 km/hr	80 km/hr	63 km/hr
	Travel Time from TH 101 to Downtown[4]	23 min	17 min	13.5 min	17 min
Decrease Accidents along I-394	Accident Rate per Million km	2.7	1.1 or less	0.59[5]	0.81 or less
Maintain High Compliance with Occupancy Requirements	Compliance Rate	Not applicable	More than 90%	87% 96%	More than 90%

[1] Source I-394 Transportation System Management Plan, 1986
[2] Source Surveys and Counts conducted in spring, 1994
[3] Source Year 2000 forecasts from Mn/DOT TA-M307, 1984
[4] Does not include delays at ramp meters
[5] As of September 6, 1993
Conversion Factor is 1 km = 1.61 miles

HOV lane has been higher than projected and the level of service in the HOV lane has been excellent. Although the use of the mixed traffic lanes has been higher than projected, higher-than-expected average speeds have been maintained. HOV use continues to climb during the peak hours, whereas mixed lane use appears to have peaked related to available capacity.

VEHICLE VOLUMES

Peak-hour volumes for the peak direction are presented in Table 2. Both the daily traffic and peak-hour traffic along the I-394 corridor have increased significantly since before construction began. The morning peak hour has demonstrated the largest percentage increase, whereas the increase during the afternoon peak is similar to the daily increase. There was a dramatic increase in both daily and peak-hour traffic volumes immediately after completion of the new highway.

The following are general observations regarding changes in vehicle volumes from 1984 to 1994:

- Daily traffic volumes on I-394 have risen substantially. At Penn Avenue, the peak-load point of the roadway, daily volumes have increased from 86,000 vehicles per day to 143,000 vehicles per day, a gain of 66 percent. At the western end, there has been an increase of 9,000 vehicles per day, a 35 percent gain.
- Morning peak-hour inbound volumes at Penn Avenue have risen by 68 percent (from 4,000 vehicles to 6,700 vehicles); at Xenia/Park Place the volumes have risen by 124 percent (from 2,500 vehicles to 5,600 vehicles).
- Afternoon peak-hour westbound volumes at Penn Avenue have risen by 75 percent (from 3,600 vehicles to 6,300 vehicles); at Xenia/Park Place the volumes have risen by 115 percent (from 2,600 vehicles to 5,600 vehicles).

TABLE 2 Peak Hour Vehicle Volumes on T.H. 12/I-394

	April 1984	Nov 1992	April 1993	Sept 1993	April 1994	% Change 1984-1994
A.M. Peak Hour (7:00 - 8:00 a.m.) (Inbound)						
Penn Avenue	4,000	4,700	6,300	6,200	6,700	68%
Xenia/Park Place	2,500	1,700[1]	5,100	5,100	5,600	124%
Plymouth Road	2,500		4,500		--	80%
I-494	--	4,500	4,800	5,000	4,500	-
P.M. Peak Hour. (4:45 - 5:45 p.m.) (Outbound)						
Penn Avenue	3,600	4,200	6,000	6,300	6,300	75%
Xenia/Park Place	2,600	4,700	4,600	5,400	5,600	115%
Plymouth Road	3,000		3,800			27%
I-494		4,600	4,700	4,900	4,200	--

[1] Significant congestion levels during count period resulting in low peak hour volumes

HIGH-OCCUPANCY-VEHICLE LANE VEHICLE VOLUMES

Mn/DOT has collected daily and monthly data on vehicle volumes in the HOV lane since the interim HOV lane opened in late 1985. Before the opening of the reversible section between TH 100 and downtown, this data was collected between Turner's Crossroads (near Xenia/Park Place) and Florida Avenue (the peak-load point for the interim reversible HOV lane that ended just east of TH 100). Since the reversible HOV lane opened, data has been collected at Penn Avenue (the peak-load point for the permanent HOV lane). Figure 2 indicates historic traffic volumes in the HOV lane during the morning peak hour and peak period.

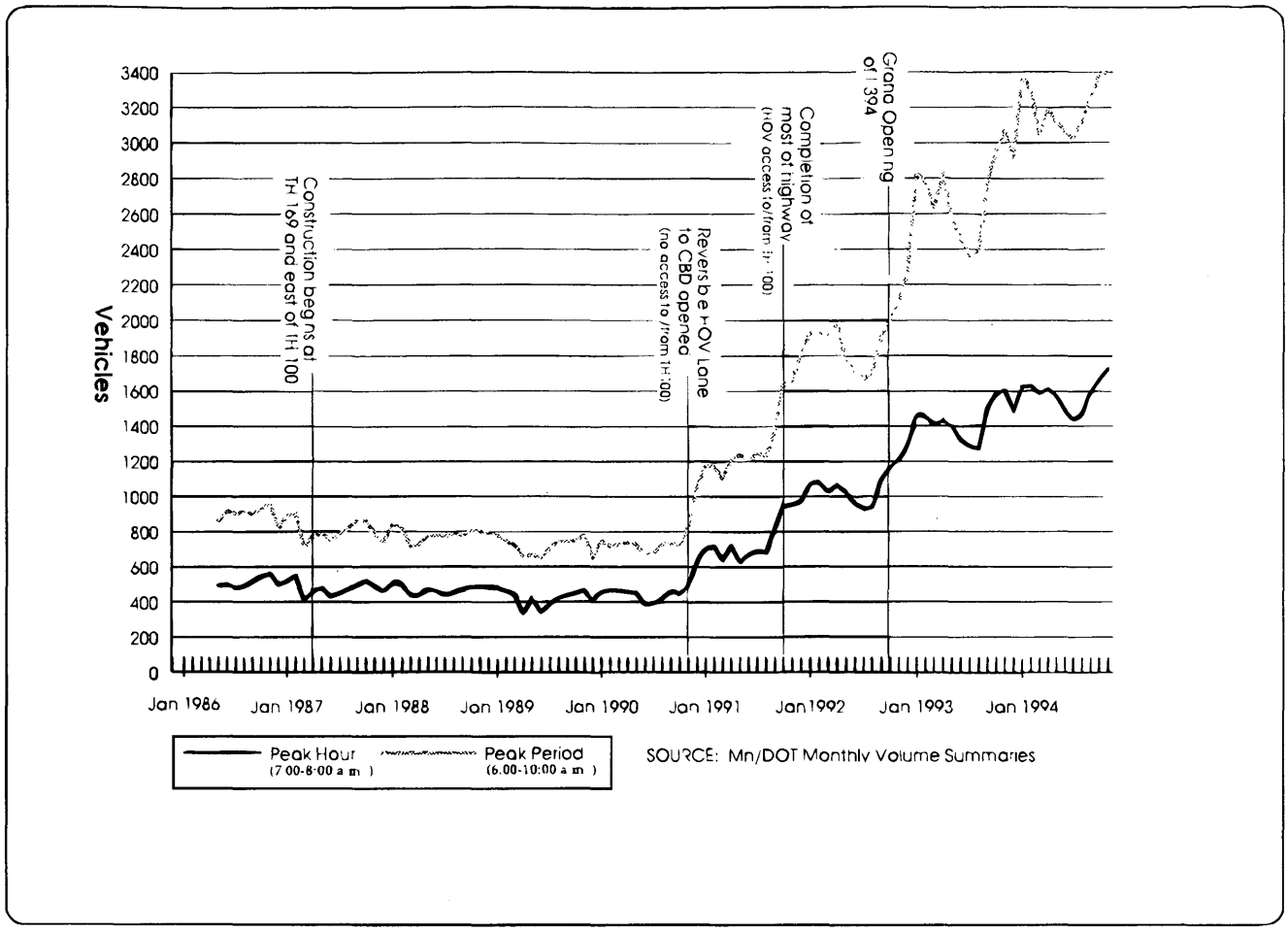


FIGURE 2 I-394 HOV lane traffic volume. At the peak load point—a.m. eastbound.

The following observations are based on an analysis of these historic HOV traffic volumes:

- In April 1994, 22,800 people used the HOV lanes on an average weekday. The HOV lanes are open eastbound from 6:00 to 10:00 a.m. and westbound from 2:00 to 8:30 p.m.
- Sharp increases in HOV lane volumes have coincided with the completion of major portions of the HOV lanes. The sharpest increase in HOV use to date occurred immediately after the I-394 grand opening in October 1992.
- The volume in the reversible lane at Penn Avenue during the morning peak hour (1,596 vehicles) was 19 percent higher than during the afternoon peak hour (1,343 vehicles) in April 1994.
- At Penn Avenue, during the morning peak hour, HOV lane volume increased from 1,139 vehicles to 1,596 vehicles, constituting a 40 percent increase from November 1992 to April 1994. The regular lanes at this location exhibited a 42 percent increase in total traffic volume for morning peak hour inbound, rising from 4,700 vehicles to 6,700 vehicles.
- The afternoon peak-hour volume at Penn Avenue in the HOV lane increased from 795 vehicles to 1,343 vehicles, a 69 percent increase from November 1992 to April 1994. The outbound total volume (regular and HOV lanes) during the same time period increased from 4,200 vehicles to 6,300 vehicles, or 50 percent.

VEHICLE OCCUPANCY RATES

One of the primary objectives of the I-394 transportation system is to increase vehicle occupancy in the I-394 corridor through the provision of incentives for car pooling. Vehicle occupancy rates have been collected at various locations at different times to measure progress in achieving this objective. Only one location is directly comparable throughout the entire project, and that is the segment of roadway east of TH 100. Table 3 shows the change in vehicle occupancy at this location from 1984 to 1994. Vehicle occupancy at the peak-load point of the roadway increased between 1984 and 1994 for both the morning and afternoon peak hours. In 1984, the morning peak-hour occupancy was 1.15, and by April 1994 it was 1.29.

TRANSIT USE

Since 1984, there has been a 126 percent increase in transit ridership on I-394 during the morning peak hour, in part because transit service has been increased and some buses have been rerouted to use the I-394 HOV lane (see Table 4).

The bus routes diverted from other roadways include Route 91 before the fall 1992 data collection and Route 53E after the fall 1992 data collection. These routes operate 10 buses during the peak

TABLE 3 Vehicle Occupancy Rate East of TH 100

	A.M. Peak Hour (Inbound)	P.M. Peak Hour (Outbound)
April 1984	1.15	1.12
May 1986	1.16	1.13
April 1989	1.16	1.13
November 1992	1.27	1.26
April 1993	1.25	1.28
September 1993	1.31	1.29
April 1994	1.29	1.27

period and eight buses during the peak hour. Additional bus service was also diverted from other corridors in spring 1994.

The use of express routes serving the I-394 park-and-ride lots was also evaluated. Reserve capacity on express buses to downtown is available at all lots during both the morning peak hour and peak period. Currently, the most heavily used routes are those providing direct service from the park-and-ride lots to downtown without intervening stops. The last bus trip departing from individual lots scheduled to arrive downtown before 8:00 a.m. is the best used of the buses departing from that lot.

Although there has been a substantial increase in both transit service and transit ridership since the beginning of I-394 construction, the transit service levels envisioned for the corridor upon completion of the roadway have not yet been achieved.

The Metropolitan Transit Commission is moving toward the implementation of a full timed-transfer system using the I-394 transit services and facilities plan as a guide, and will continue to improve service incrementally as resources permit.

TABLE 4 Transit Ridership on TH 12/I-394 During the a.m. Peak Hour East of TH 100

	Ridership	% Change From 1984	Buses	% Change From 1984
April 1984	1,000		27	
May 1986	1,160	16%	35	30%
November 1992	1,492	49%	50	85%
April 1993	1,633	63%	53	96%
September 1993	1,717	72%	49	81%
April 1994	2,256	126%	70	159%

PERSON TRIPS

The number of automobile person trips is added to the number of transit riders to determine total person trips using the corridor. In April 1994, there were an estimated 10,403 person trips in 6,679 vehicles (automobiles, buses, trucks, and motorcycles) at Penn Avenue during the morning peak hour. Fifty percent of all person trips were in the HOV lane.

From 1986 to 1994, vehicle trips east of TH 100 during this time period increased by 57 percent, from 4,250 vehicles to 6,679 vehicles, whereas person trips increased by 72 percent. The increase in person trips predominantly occurred in the HOV lane. From 1992 to 1994, person trips at Penn Avenue during the morning peak hour inbound increased 41 percent, from 7,376 to 10,403.

In the morning peak hour, the HOV lane at Penn Avenue has the ability to carry many more persons, whereas the mixed lanes at the same location have already reached their vehicle trip capacity. In order for there to be an increase in person trips at Penn Avenue, there must be an increase in people using buses or car pools.

PARK-AND-RIDE LOTS

Table 5 indicates the initial capacity of park-and-ride lots along TH 12, along with the current capacity and use of I-394 lots. Seventy-five percent of the 301 spaces available in the corridor were being used in March 1986, or approximately 225 cars. In April 1994, there were 677 vehicles parked in I-394 park-and-ride lots, an increase of 200 percent. Sixty-six percent of the 1,021 available park-and-ride lots spaces are now being used.

TRAVEL TIMES

In terms of travel time, the construction of I-394 had two goals: (*a*) to reduce travel time for everyone using the roadway, and (*b*) to provide a travel-time incentive for people to car pool and ride the bus.

For travel from I–494 to Penn Avenue, the travel time in the mixed lanes during the a.m. peak hour has dropped from 17.4 min

TABLE 5 Park-and-Ride Spaces in I-394 Corridor

	Initial Capacity	Current Capacity	Current Use	Percent Utilized
Xenia/Park Place	--	60	31	52%
Louisiana Avenue	--	173	127	73%
General Mills Boulevard	--	112	34	30%
CSAH 73 (north and south)	182	467	317	68%
Ridgedale Mall	30	--	--	--
Plymouth Road	--	111	97	87%
Wayzata	89	98	71	72%
Total	301	1,021	677	66%

in 1984 to 8 min in April 1994, a savings of 9.4 min or a 54 percent improvement (see Table 6). These times do not include delays at ramp meters. For those now using the HOV lane, travel time in the a.m. peak hour has dropped from 17.4 min to 7.3 min in April 1994, a savings of 10.1 min. This represents a 58 percent improvement.

In the spring of 1994, there was a 1.7-min time savings in the HOV lane over the mixed traffic lanes during the morning peak hour from TH 101 to downtown Minneapolis (see Table 6). Travel times in the HOV lanes are very consistent, whereas travel times can vary considerably in the mixed lanes, depending on driving conditions and incidents. In addition, vehicles using the mixed traffic lanes have additional delays at the metered ramps.

The mixed-lane travel times in spring 1994 were exceptionally fast, and the use of a relatively small sample taken under ideal (i.e., no incidents) conditions does not best represent the observed travel time differences. The variability in speeds in the mixed traffic lanes versus the speeds in the HOV lanes is indicated in Figure 3. This figure depicts the range of expected speeds along the facility based on several years of sample data. Clearly, the overall speeds in the HOV lanes are higher than those in the mixed lanes, and at the same time have significantly less variation.

Since the opening of I-394, Mn/DOT has been adapting its traffic management system for the corridor to alleviate or minimize operational problems. As part of this effort, entrance ramp-metering rates are adjusted constantly on the basis of downstream flow rates. This in effect allows the mainline to operate relatively smoothly in periods of good weather and when incidents are not present. As a result, traffic destined to the mixed lanes is spending more time at the ramp meters, whereas HOVs are given either meter bypass lanes or direct ramps to avoid the queue. The typical delay at the TH 100 ramp meter for mixed traffic is 6 to 8 min during the morning peak hour. Therefore, the total travel time difference between HOVs and mixed traffic, incorporating both ramp delay and mainline speed differences, is 7 to 9 min.

TABLE 6 Travel Time in Minutes on TH 12/I-394 a.m. Peak Hour

	Mixed Lanes	HOV Lanes	Travel Time Savings In HOV Lanes
I-494 to Penn Avenue			
• April 1984	17.4	--	--
• May 1986	14.9	9.7	5.2
• November 1992	10.9	6.7	4.2
• April 1993	9.9	6.6	3.3
• September 1993	11.8	7.6	4.2
• April 1994	8.0	7.3	0.7
TH 101 to Third/Fourth Street			
• November 1992	17.4	13.4	4.0
• April 1993	16.4	13.0	3.4
• September 1993	16.7	11.6	5.1
• April 1994	13.5	11.8	1.7

COMPLIANCE WITH OCCUPANCY REQUIREMENTS

The occupancy requirement for the I-394 HOV lane is two or more persons per vehicle. In addition, motorcycles are permitted to use the HOV lane and HOV meter bypass lanes. Enforcement of occupancy requirements is very important to protect the integrity of the HOV lane, both to maintain high operating speeds and to maintain an incentive for ride sharing. The compliance rate is the percentage of vehicles in the HOV lane that meet the automobile occupancy requirement of two or more people. Violation rate refers to the percentage of vehicles that do not meet the occupancy requirement.

Reversible High-Occupancy-Vehicle Lane

Compliance in the reversible HOV lane (both interim and permanent) has ranged between 93 and 98 percent during the eastbound morning peak hour since 1986 (see Table 7), and between 92 and 97 percent during the afternoon westbound peak hour at the peak-load point. For the period between November 1992 and April 1994, compliance in the reversible section during the morning and afternoon peak hours has been stable.

Concurrent High-Occupancy-Vehicle Lanes

Compliance rates are slightly lower in the concurrent HOV lane segments (87 to 97 percent) at Louisiana Avenue, but even lower at the western end of the corridor, during both the morning and the afternoon peak hours (see Table 7). However, compliance rates in the concurrent-flow HOV lanes are consistent with or higher than concurrent-flow HOV lanes facilities in other cities.

CAR POOL PARKING IN DOWNTOWN MINNEAPOLIS

Three garages with a total capacity of 5,923 parking spaces have been constructed over the eastern end of I-394 near downtown Minneapolis (see Figure 1). These garages have direct access to/from I-394. Approximately 90 percent of capacity (5,302 parking spaces) is available for monthly contract parking. Car pools from I-394 are eligible for a reduced contract parking fee of $25 per month. Car pools not from I-394 and all single-occupant vehicles pay a monthly fee of $90.

As of July 1994, 3,115 contract spaces in the three garages had been sold. This is 59 percent of the 5,302 spaces available for monthly contract parkers (10 percent of spaces are held for hourly parkers). Sixty percent of these are I-394 HOVs. The remainder are HOVs from other routes or single-occupant vehicles (SOVs). Garage A is 52 percent full whereas Garage B is 90 percent occupied. Garage C, which currently has a utilization rate of 38 percent, was opened in November 1992.

PERFORMANCE OF HIGH-OCCUPANCY-VEHICLE SYSTEM

Overall, the HOV system has performed beyond initial expectations:

- Use of the HOV lane has exceeded the start up objectives and continues to grow.

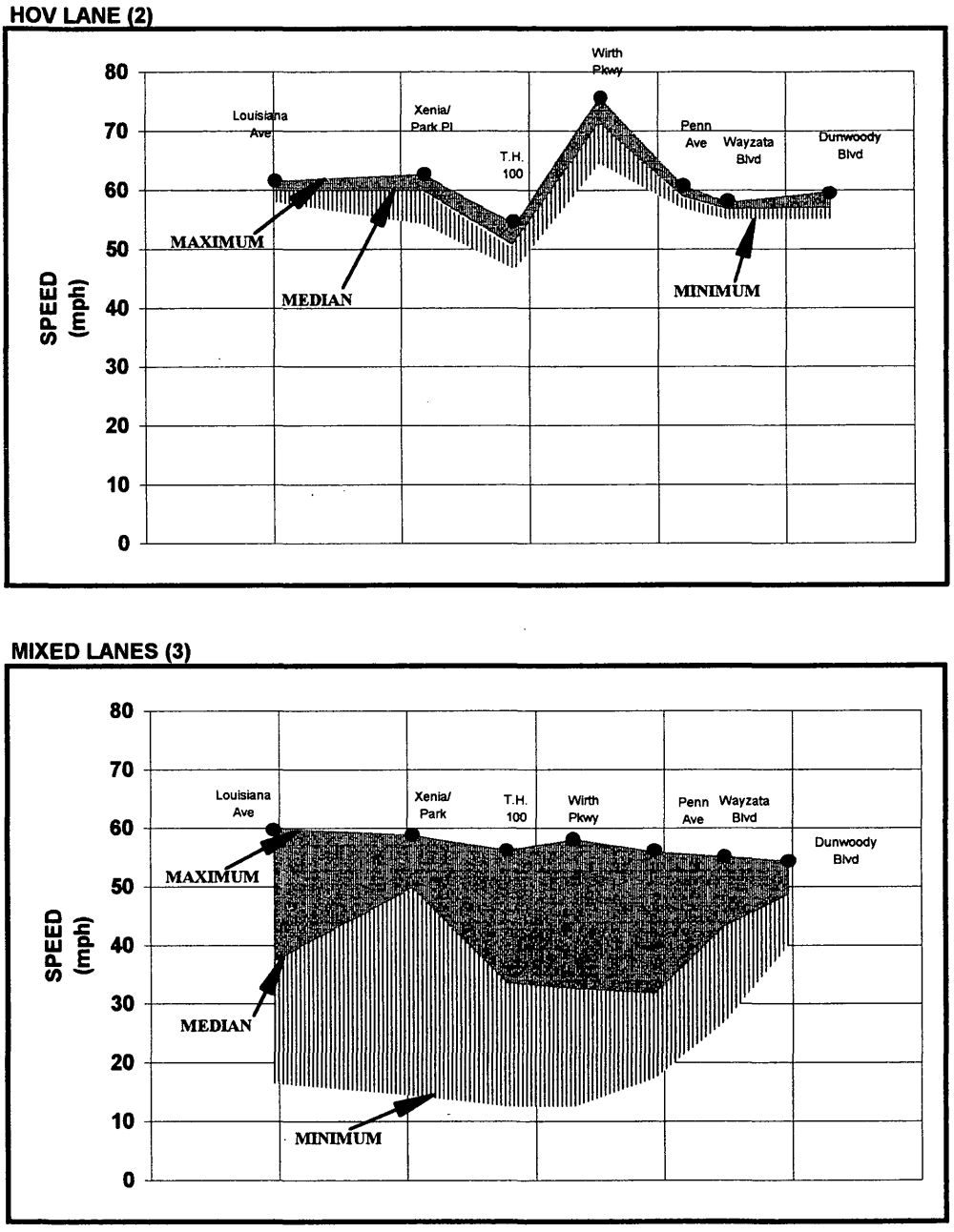


FIGURE 3 Peak hour a.m. speeds on eastbound I-394. Louisiana Ave. to Dunwoody Blvd. 1) Plotted speeds represent averages for the segment defined by the points shown, i.e., the speed for Penn Ave. is that for the segment between Wirth Parkway and Penn Ave. 2) Speeds computed using runs from Fall 1992, Spring 1993, Fall 1993, and Spring 1994. 3) Speeds computed using runs performed in Fall 1994.

- Vehicle occupancy objectives for the start up period have been met. However, the overall auto occupancy rate declined when the automated traffic management system became functional as a result of an increase in the peak-hour capacity of the mixed traffic lanes.
- Bus ridership has increased 126 percent, fulfilling the objective for the start up period. However, substantial route changes have occurred to take advantage of the time savings in the HOV lane. Total service area ridership changes are not presently known.
- Travel speeds are higher than expected in both the HOV and the mixed traffic lanes. Peak-hour speeds in the mixed traffic lanes increased significantly when the automated traffic management sys-

TABLE 7 HOV Lane Compliance Rates

		Reversible HOV Lane (1)		Concurrent HOV Lanes (2)	
		A.M. Peak Hour	P.M. Peak Hour	A.M. Peak Hour	P.M. Peak Hour
May	1986	95%	97%	--	--
April	1989	98%	--	--	--
November	1992	95%	94%	96%	89%
April	1993	96%	92%	97%	90%
September	1993	93%	96%	96%	89%
April	1994	98%	93%	95%	87%

(1) *At peak load point.*
(2) *At Louisiana Avenue.*

tem became fully functional, thus reducing the speed differential between the HOV lane and the mixed traffic lanes.

- Travel time savings in both the HOV lane and the mixed traffic lanes are greater than expected when compared to the "before" condition. However, travel time differences between the HOV lane and the mixed traffic lanes are less than expected, even though volumes in the mixed traffic lanes are higher than projected.
- Although travel time savings between the HOV lane and the mixed traffic lanes declined significantly when the automated traffic management system became fully functional, the time savings at the HOV bypass lanes on the metered ramps increased significantly. However, there are no HOV bypass lanes at the I-494 and TH 169 interchanges, which have long ramp queues and delays.
- Travel times in the HOV lanes are very consistent, whereas travel times in the mixed traffic lanes may vary dramatically from day to day. This factor may also contribute to perceived average travel time savings, which are higher than measured average travel time savings.

CHANGES IN DATA COLLECTION AND INTERPRETATION

Based on the findings of the I-394 evaluation to date, the following changes are recommended in the collection and interpretation of traffic data related to HOV systems, particularly those operated in a managed freeway environment.

- Increases in peak-hour car pooling are a complex mix of mode shift, route changes, time-of-travel changes, and changes in car pool participants. Field traffic data reflect all of these factors. It is not possible to determine the exact level of mode shift from field traffic data. Modal shifts can be better determined through surveys.
- Auto occupancy rates are influenced more quickly by vehicle volume changes in the mixed traffic lanes than in the HOV lane. Because the volume and speed of traffic in the mixed traffic lanes are influenced significantly by the automated traffic management system, small changes in vehicle occupancy rates should not be interpreted as changes in car pooling. Long-term trends in the volume of activity in the HOV lane itself are a more accurate representation of changes in car pooling in the corridor.
- Because of circulation bus route changes that appear to improve overall bus travel time savings, changes in ridership should be evaluated on a service area basis rather than a linear corridor basis. In addition, bus travel time savings should include the savings associated with circulation route changes, as well as the mainline travel time savings.
- It is important to collect and report travel time variability in the mixed traffic lanes as a measure of the predictability and reliability of the HOV lane. This has been identified as an important factor in mode choice in recent focus groups.
- In a managed freeway environment, travel time savings at the HOV bypass lanes on metered ramps are a significant portion of overall travel time savings. It is important that travel time data collection is designed to include these time savings, as well as mainline time savings.
- Enforcement of occupancy requirements in the I-394 parking garages has been very difficult, in part because drop-offs are allowed before entering the garages. An automated method for verifying vehicle occupancy and travel on the I-394 HOV lanes is needed to address this problem.

OPERATION OF HIGH-OCCUPANCY-VEHICLE FACILITY WITHIN CONTEXT OF MANAGED FREEWAY

The most difficult issue associated with evaluating the operation of the I-394 HOV lane is related not only to the change of the highway from a signalized arterial to a limited access freeway, but also to the change from an unmanaged highway to a fully automated, managed system. The ability to control the flow and, therefore, the speed and volume of traffic in the mixed traffic lanes has a direct impact on the performance of the system as a whole and the differences in performance between the HOV lane and the mixed traffic lanes. Typically, HOV lanes operating in the context of an automated, managed freeway system will have the following characteristics:

- There may be a lower mainline speed differential between the HOV lane and the mixed traffic lanes because of the ability to control mainline speed in the mixed traffic lanes.
- HOV bypass lanes at metered ramps are very important because the time savings for HOVs may be more significant on the bypass lanes than on the mainline.
- The overall vehicle occupancy rate and the percentage of car pools in the total traffic mix may be lower because the peak-hour capacity of the mixed traffic lanes can be increased by the traffic management system.

These potential impacts need to be addressed in a well-defined philosophy for traffic management in an HOV system corridor, and specifically for the management of traffic on I-394. These factors also need to be reflected adequately in the interpretation of traffic data relative to system performance.

Publication of this paper sponsored by Committee on High-Occupancy Vehicle Systems.

Design of Incident Detection Algorithms Using Vehicle-to-Roadside Communication Sensors

EMILY PARKANY AND DAVID BERNSTEIN

Incident detection methods for the automatic recognition of accidents and other freeway events requiring emergency responses have existed for over twenty years. Most of the developed and implemented algorithms rely on inductive loop data. Inductive loops are the most commonly used traffic sensor and collect data such as volume and occupancy at a point. However, the implemented algorithms using inductive loop data work with mixed success. Recently, there has been renewed interest in incident detection algorithms partly because of new sensors for obtaining traffic information. One of these new sensors is vehicle-to-roadside communications (VRC), which consists of electronic "tags" on the vehicles and readers along the roadway. These obtain counts, headways, travel times, lane switches, and other information about vehicles between subsequent readers. This paper explores the use of VRC data for incident detection. After a discussion of the use of VRC as a surveillance tool for incident detection, a few example pattern-based algorithms are described. Preliminary results of these algorithms suggest that VRC is a viable sensor to use for incident detection. The final section discusses further directions for this type of research.

A recent study suggests that incident-related congestion accounts for 64 percent of the total delay due to congestion and that this incident-related congestion could increase to 72 percent of total congestion by the year 2005 (*1*). Additionally, wasted fuel and lost productivity caused by congestion delays have great societal costs [one article suggests congestion delays cost $34 billion a year (*2*)]. The congestion-impact numbers are staggering and show that incident management that reduces these delays could provide a significant contribution toward the goals of increasing the capacity of existing roadways, enhancing air quality, reducing driver frustration, and increasing safety. Incident detection, the first step of incident management, is determining that an accident, stall, or something else that requires a response has occurred.

Automatic incident detection algorithms for freeways have existed and have been implemented since the early 1970s. Few algorithms have been developed for arterials because that is a much more complicated problem. For some researchers, new sensors and new data have led recently to renewed interest in incident detection. The goal of this paper is to show that the data obtained with vehicle-to-roadside communication devices (VRC), also called automatic vehicle identification (AVI) equipment, may lead to better-performing incident detection algorithms. Other work has been conducted recently in this area (*3–5*), but the approach and algorithms presented in this paper are unique.

E. Parkany, University of California at Irvine, Institute of Transportation Studies, 401 Berkeley Place, Suite 200, Irvine, Calif. 92717. D. Bernstein, Princeton University, Department of Civil Engineering and Operations Research, E-Quad, Room E-408, Princeton, N.J. 08544.

VRC has received considerable attention as the enabling technology for electronic toll collection and related applications such as congestion pricing. However, the data that can be obtained from VRC are also valuable for traffic monitoring and as inputs to traffic control algorithms. This multifunctionality helps to set VRC apart from other traffic sensors. An integrated congestion pricing, incident detection, and route guidance system is described further in Bernstein et al. (*6*). With such integrated systems, operating agencies are getting "double-duty" from their technology investment. Additionally, system operators may find it easier to get the public to accept a controversial component, for example, congestion pricing, when the public believes that the system is providing additional benefits such as incident detection and route guidance.

This research is concerned with detecting the beginning of accidents and stalls and other incidents that cause traffic disruptions on freeways and require the emergency response of an ambulance, police, and/or tow truck. The basic premise of this research is that it may be possible to replace the inductive loop data (or loop-emulated data) currently used for incident detection with data obtained from a VRC system.

There are two ways that this can be done. In the first the data obtained from the VRC system would be used with existing algorithms. However, this may not be effective for two reasons: existing inductive loop-based algorithms do not work very well [see, e.g., the review articles by Stephanedes et al. (*7*) and Chen and Chang (*8*)], and VRC data represent only a percentage of the vehicles and vehicle types on the freeway and, hence, to use these data in existing algorithms may require processing or manipulation to be representative of all traffic. The second way to use VRC for incident detection is to develop new algorithms that take advantage of the different attributes of VRC data. In this paper we ex-plore some of the properties of VRC data and take some initial steps toward the development of VRC-based incident detection algorithms.

The VRC-based algorithms we develop and describe in this paper incorporate several approaches to incident detection. The algorithms consider temporal and spatial patterns in the data. The algorithms detect in all traffic flow levels and require little persistence checking. It is expected that these algorithms can either stand alone or be combined with other sensor data and other incident detection methods for an incident detection system.

Vehicle-to-roadside communication is described here as a surveillance tool for incident detection along with a few possible incident detection algorithms that use VRC data, preliminary results for these algorithms, and future directions for this type of research.

VRC-BASED SURVEILLANCE FOR INCIDENT DETECTION

In our opinion there are three broad categories of data that can be used for incident detection. Point data are collected at a single, specific location and include occupancy, instantaneous speed, and flow. Area data are collected over a segment of roadway and include density. Finally, point-to-point data are collected between pairs of specific locations and include travel time.

Different types of sensors can (and should) be used to collect different types of data. The most common sensor in use today, inductive loops (and loop emulators), collect point data only. The potential advantage of a VRC-based surveillance system is that it can be used to collect both point data and point-to-point data. Very briefly, VRC consists of a transponder or "tag" on the vehicle and a reader along or over the road and the communications link between the two. In all VRC systems, the VRC readers obtain at least the individual vehicle identification number from each transponder-equipped vehicle that passes. With the identification numbers, the system knows, for example, what time car Number 123 passed Reader A and what time it passed Reader B and can calculate that car's travel time between the readers. Additionally, a VRC system can obtain lane-specific and station-specific headways (time between transponder-equipped, "tagged" vehicles), the volume of tagged vehicles on a section at any point in time, the number of tagged vehicles passing in each lane at a reader station, and the number of tagged vehicles that switch lanes between stations. More information is obtained with read-write capabilities as described below.

Even the more easily obtained data items—vehicle-specific travel times, lane- and station-specific headways, section volumes, and lane switches—are all parameters that can only best be obtained with a sensor that obtains data between two or more points such as VRC. This point-to-point data should better represent traffic compared with data collected at a certain point. Of course, VRC is not the only technology that can be used to collect point-to-point data. In fact, as discussed in the review paper by Bernstein and Kanaan (*9*), any automatic vehicle identification (AVI) technology can be used for this purpose. Hence, before moving to a discussion of the algorithms themselves, some of the issues regarding the development of algorithms with VRC are briefly discussed here. These issues include the number of sensors used, read-only versus read-write capabilities, and penetration rates (percentage of vehicles accurately detected) of the sensors.

Numbers of Sensors

Some people may presume that the only VRC sensors available will be those used for another purpose (such as electronic toll collection). However, it is possible to install "extra" readers. In fact, using VRC systems for incident detection likely will involve readers in additional locations to those required for electronic toll collection, which raises at least three institutional issues. First, although it is most likely that the cost of installation of extra readers and software for incident detection will be less than the benefits provided by having an incident detection system, extra readers are an extra expense, making it impossible to use the "toll collection and incident detection for the price of one" argument. Second, users need to understand that the extra readers will not deduct yet another toll but are there to aid in incident detection and thus reduce congestion. Third, on proposed systems where conventional toll payers (those using manual toll booths) must exit the roadway to pay a toll, the extra readers may be additionally confusing and a safety hazard.

Read-Only Versus Read-Write Capabilities

Read-write means the reader obtains information and also writes information on the "tag." In addition to the vehicle identification being passed, the transponder can pass information written to the tag such as initial reader passed (origin information), time passed last reader (making for easy travel time calculations), even processed data such as average volumes or headways passed from one reader to the next. A more sophisticated system may also pass the vehicle type, driver-input origin, and destination information, even the traveling speed from the vehicle's speedometer to the reader and, subsequently, an incident detection system. With a read-write system, VRC, unlike other surveillance technologies, allow for the passing of required information from one point to the next and traffic control such as incident detection to be performed locally.

Incident detection can be performed with a read-only system, but there would be advantages to using a read-write system. Most obviously and at the simplest level, the time passed the last reader and processed data may aid the calculations required at the downstream readers and may speed up the effort required for incident detection. The algorithms presented in this paper work with read-only technology but could be enhanced with read-write technology.

Penetration Rates

Although VRC-based systems have the advantage of being read-write-capable, they do have one important disadvantage. VRC data is incomplete in that not all vehicles and types of vehicles are represented with the data. This is because generally only a certain percentage of vehicles are equipped with transponders. Additionally, VRC may be used by a specific group of vehicles such as heavy trucks. For example, there are already two multi-state heavy truck implementations of VRC. All of this information is valuable, but it should be explored whether the partial information is sufficient to represent all traffic for incident detection algorithms. It is our guess that even a "small" (30%) portion of vehicles if they were autos or regular commuters would be sufficient data. However, if only a small percentage of vehicles are tagged, then the data obtained have greater variance and smaller reliability. For example, headway information would be subject to randomly distributed fluctuations, especially in light traffic, if a small number of vehicles are transponder-equipped. This variability may lead to using different algorithms for different percentages of tagged vehicles. Additional consideration should be given if only trucks or buses were tagged. This surely would add a bias to the data because trucks and other heavy vehicles do not travel as quickly or maneuver as smoothly as typical traffic. Yet, this bias may also work in favor of an algorithm. For example, an incident detection algorithm based on lane changes may be more powerful if it detects that a truck or several trucks have moved from the typically used right lane.

EXAMPLE PATTERN-BASED ALGORITHMS

There are several different approaches that may be used for incident detection algorithms. Recent research has investigated the use of

processed video (*10*), catastrophe theory (*11*), and artificial intelligence (*12*) as methodologies for incident detection algorithms. More typical designs are either statistical (generally after a time series until an anomaly occurs) (*13,14*) or pattern-based where the data are typically compared with numerical thresholds to determine that a stoppage has occurred (*15,16*). Most of the implemented algorithms follow a pattern-based approach, probably for several reasons. First, the logic is simple and easily understood by traffic operators who must trust the results of the algorithms. Second, they use less computer time and hardware than other methods. Although it is considered by many current researchers that other methods will perform better and that computer requirements and other technology advances are less constraining now than fifteen or more years ago when different methods were first proposed, this research focuses on pattern-based approaches to show that VRC-based pattern algorithms work just as well as currently implemented algorithms using other sensors.

Previous research (*17*) and a previous paper (*18*) describe various VRC data "incident indicators" and four possible pattern-based algorithms. The following paragraphs describe the best-performing of these algorithms, named the Headway Algorithm, and two additional algorithms that we have named the Lane Switches Algorithm and Lane-Monitoring Algorithm. The algorithms are designed to be used with vehicle to roadside communication sensors (less than 100% of vehicles tagged, accurate identification of all tagged vehicles) but can also be used with other types of automatic vehicle identification sensors (e.g., processed video license plate readers). The motivation and logic behind these algorithms is given along with a flow chart of each algorithm. All three of these algorithms represent new logical approaches to incident detection. These ideas have not yet been described or implemented for any sensor.

Headways Algorithm

VRC data can be used in two major ways in an incident detection algorithm. First, it can be used to observe temporal differences. In general, increased travel times from one period to the next or any large difference in travel times from one period to the next strongly indicates unstable conditions and, possibly, incidents. Second, spatial comparisons can be made by either comparing headways (time between subsequent tagged vehicles) at two different readers or the volumes (number of tagged vehicles) on two different sections. Longer headways at downstream readers compared with upstream readers or smaller volumes on downstream sections compared with upstream sections may indicate an incident.

Both temporal and spatial comparisons of travel times and headways are used in the Headways Algorithm. The algorithm consists of three sequential tests; if the three tests are satisfied during a particular time interval, then an incident is declared. The first test looks for a significant difference in travel time from one time interval to the next. It is thought that slower travel times may be indicative of an incident. The second test, another temporal test, considers the differences in headways at the downstream reader for the current time interval and the previous time interval. An incident is likely to cause longer and longer headways as vehicles are queued and then have to maneuver around the incident. The third test makes the spatial comparison of whether headways are different at different reader locations. Again, headways may be longer in the vicinity of the incident and then decrease downstream of the incident. These different tests can be described mathematically as follows.

The following text describes the Headways Algorithm, which is illustrated in Figure 1. In the Figures, *int* denotes the length of the selected time interval. Let $Nv(t_{cur})$ denote the number of vehicles that passed the downstream reader location r_{down} during the current time interval, t_{cur}. Then the first test compares the average travel time during the current interval, t_{cur}, with the average travel time during the previous interval, t_{rev}. The average travel time, $ATT(t_{cur}, r_{down})$, between the upstream reader, r_{up}, and the downstream reader, r_{down}, during the current interval is given by

$$ATT(t_{cur}, r_{down}) = \frac{1}{Nv(t_{cur})} \sum_{j=1}^{Nv(t_{cur})} TT_j \quad (1)$$

where TT_j is the travel time between readers r_{up} (upstream) and r_{down} (downstream) of the jth vehicle to pass reader r_{down} during the current interval, t_{cur}. If $|ATT(t_{cur}, r_{down}) - ATT(t_{prev}, r_{down})|$ is greater than a prespecified, possibly flow-dependent threshold, *HD_TH1*, then the next test is conducted.

The second test compares the headways (time between vehicles) at the downstream reader for two different time periods (current time interval and previous time interval) against a second threshold,

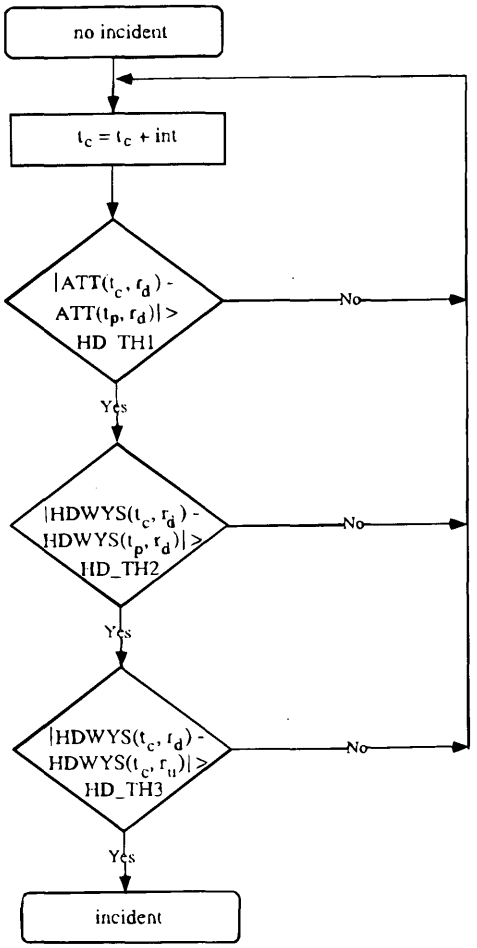


FIGURE 1 Headways algorithm.

HD_TH2. Our parameter for *average headways*, $HDWYS(t_{cur}, r_{down})$, is defined as follows:

$$HDWYS(t_{cur}, r_{down}) = \frac{1}{Nv(t_{cur})} \sum_{j=2}^{Nv(t_{cur})} t_j - t_{j-1} \quad (2)$$

where t_j is the actual time of the jth vehicle to pass reader r_{down} during the t_{cur} interval.

The final test before an incident is declared compares the headways at the downstream reader and the upstream reader during the current time interval against a third threshold, *HD_TH3*. We define r_{up} as the upstream reader. If $|HDWYS(t_{cur}, r_{down}) - HDWYS(t_{cur}, r_{up})|$ is greater than *HD_TH3*, then an incident is declared.

Lane Switches Algorithm

In addition to travel time and headway comparisons, there are other ways that VRC data may indicate that an incident has occurred. For example, VRC is able to provide vehicle-specific data such as lane change information. A large number of lane switches noted from one reader to the next likely indicates unstable traffic conditions.

Our Lane Switches Algorithm is depicted in Figure 2. Basically, the system determines the number of vehicles that have switched lanes between readers, $SWITCH(t_{cur}, r_{down})$, using the lane-specific, vehicle-specific data obtained at the reader r_{down} during the current time period, t_{cur}. This number is normalized by the number of tagged vehicles that pass during the time interval $Nv(t_{cur})$ to get $NM\_SW(t_{cur}, r_{down})$, the *normalized number of switches*:

$$NM\_SW(t_{cur}, r_{down}) = \frac{1}{Nv(t_{cur})} SWITCH\,(t_{cur}, r_{down}) \quad (3)$$

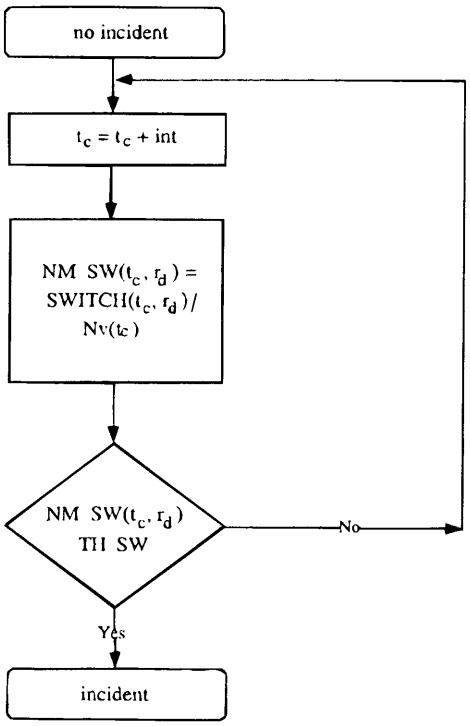


FIGURE 2 Lane switches algorithm.

If the result exceeds a certain threshold corresponding to the percentage of vehicles that have switched lanes (*TH_SW*), an incident is declared. The tested algorithm counted a lane switch from the right lane to the left lane (through the middle lane) as one switch. Perhaps counting such a maneuver as two lane switches would provide even more information about traffic conditions.

Lane-Monitoring Algorithm

The idea behind the Lane-Monitoring Algorithm is to track over two or more time intervals the vehicles that pass in each lane at a reader location. If fewer vehicles pass in a certain lane than expected, then the other lanes are checked to see if more vehicles than usual have passed. Rerouted vehicles may be indicative of an incident. Each interval, each lane in turn is compared against the low threshold. If the low threshold is met, then the other lanes are checked against a high threshold.

Figure 3 shows our Lane-Monitoring algorithm. To smooth out the data (and prevent false alarms), this algorithm uses the average number of vehicles that pass the reader in each lane over a prespecified number of intervals, say, two or three intervals. For example,

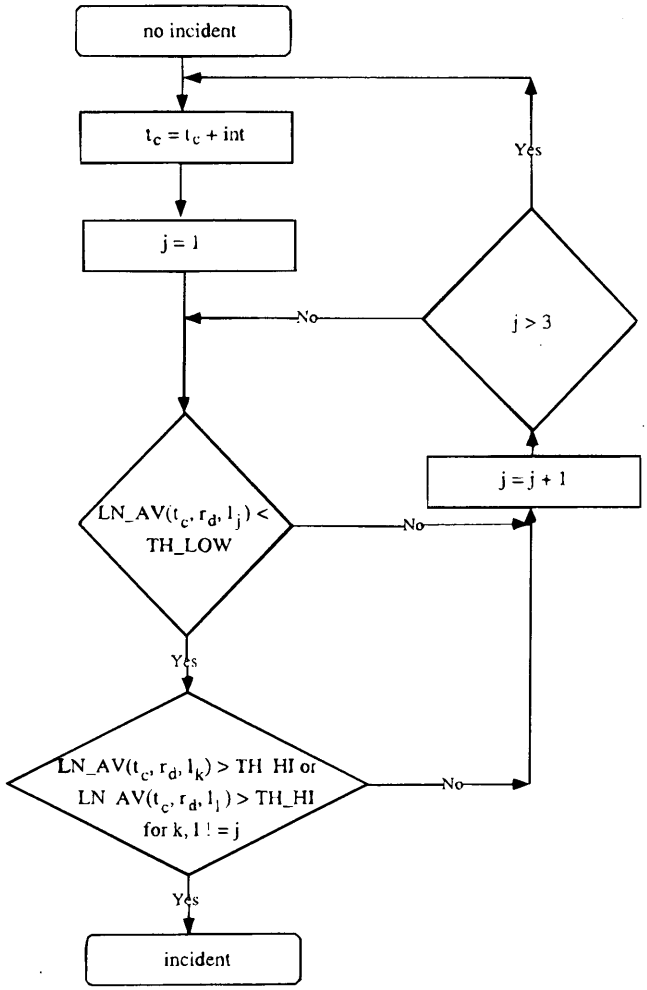


FIGURE 3 Lane monitoring algorithm.

the *lane average* for two time intervals for vehicles in the right lane (lane 1) is given by

$$LN\_AVG(t_{cur}, r_{down}, l_1) = \frac{Nv(l_1, t_{prev}) + Nv(l_1, t_{cur})}{2} \quad (4)$$

where t_{cur} is the final (current) time interval, r_{down} is the reader location, l_1 indicates the right lane, and $Nv(l_1, t_{cur})$ is the number of vehicles passing reader r_{down} in lane 1 (the right lane) during the current interval. The average number of vehicles in each lane is compared with a low threshold (*TH_LOW*). If the average number of vehicles is less than this threshold then the average number of vehicles at that reader passing during those time periods in the other lanes is compared with a high threshold (*TH_HIGH*). If any of the other lane volumes exceed the high threshold, then an incident is declared. For example, if few cars are in lane 3 but a higher than usual number of vehicles are in lanes 1 and 2, then probably an incident occurred in lane 3.

TESTING THE ALGORITHMS USING SIMULATION

Ideally, one would like to use field data to test and evaluate incident detection algorithms. In this case, this was not possible. One of the objectives of this research was to evaluate incident detection algorithms that use VRC data against an algorithm that uses inductive loop data. For this the VRC and loop data must be obtained from the same vehicles during the same time periods at the same points. Although field inductive loop and VRC data exist separately, both sensor types are not available at the same locations.

Instead of field data, the algorithms were tested on data generated by a microscopic traffic simulator described previously (*19*, *20*). The simulator has been tested with the San Diego data set used previously to illustrate the performance of INTRAS, another microsimulator, and was found to perform slightly better than INTRAS in replicating the data collected from the field (*19*). The tested incident detection algorithms were run on simulated inductive loop and VRC data. Although simulated data are not perfect, there are several advantages to using this simulator. First, the consistent (VRC and loop data at the same locations) data can be obtained. Second, it is easy to test different incident scenarios under different flow conditions. Third, it is possible to change detector spacings and configurations and roadway geometries to determine how the algorithms perform most optimally.

The network used to obtain these preliminary results is a 19.31 km (12.0 mi) long three-lane highway. The first half of the freeway is used for warm-up, and then the detectors are placed at typical 1.21 km (0.75 mi) intervals. Detectors are located in each lane at the 9.66, 10.86, 12.07, 13.28 km (6.0, 6.75, 7.5, and 8.25 mi) points. The first half [9.66 km (6.0 mi)] of the network is used for warm-up because vehicles are loaded gradually onto the network and reach the desired flows for the second half of the simulation.

Several different incident scenarios were designed to test algorithm performance with different incident types, different incident locations with respect to the sensors, and different traffic flow levels. The forty different incident data sets cover short-duration incidents (e.g., a single-vehicle stall on the mainline for 5 minutes), more serious accidents (three cars blocking traffic for 20 minutes), and incidents in between. To test the performance of the algorithms in relation to incidents' locations with respect to the sensor (reader or loop) locations, incidents were staged 0.40 km (0.25 mi) before the sensors, 0.81 km (0.50 mi) before the sensors, and at the sensors. To test the algorithms against false alarms, nonincident data sets included normal flowing traffic and sets where two vehicles traveled in adjacent lanes at speeds slower than the rest of the traffic flow. The tested flows include 1,000 vehicles per lane per hr, 1,200 vehicles per lane per hr, and 1,400 vehicles per lane per hr. These less-than-heavy traffic flows were dictated by hardware constraints, but it is assumed that if the algorithms perform adequately in medium flows they will perform as well or better with heavier flows.

A 40-min simulation is used that includes a 15-min warm-up time, incident occurrence, and clearance. The warm-up time and warm-up distance are necessary for traffic to follow normal behavior at the desired flow level by the time the sensor locations are reached. Thus, 25 effective minutes of simulation are obtained from each data set. The same input file is used for each flow level. Vehicles enter one end of the freeway according to specified rates and exit if they reach the other end of the freeway during the 40 simulated minutes.

All of the algorithms were tested with the same data. It was assumed that 50 percent of the vehicles were equipped with VRC transponders. This percentage was chosen because current toll collection systems have participation rates of 30 to 80 percent. Thirty-second time intervals were used. Data were averaged across all lanes at a sensor or considered separately according to the algorithm used.

PRELIMINARY RESULTS

Although we are not yet ready to draw any final conclusions about the performance of these algorithms, there are some important insights that can be drawn from these simulation results. The intent of these results is to show that even the simple VRC-based algorithms perform at least as well as implemented algorithms using other sensors. Additionally, compared with other simple VRC-based algorithms developed, implemented, and tested during the course of this research, these specific algorithms and their corresponding logics seem to give the most promising results.

Quantitative Results

The typical incident detection performance measures—detection rate, false alarm rate, and mean time to detect—were used here. These measures are defined as follows. First, the detection rate:

$$DETECTION\ RATE = \frac{\#inc_{det}}{\#inc_{true}} \quad (5)$$

where $\#inc_{det}$ is the number of detected incidents and $\#inc_{true}$ is the number of actual (simulated) incidents. There are two ways to define false alarm rates. First:

$$FALSE\ ALARM\ RATE\ (\text{def } 1) = \frac{\#inc_{false}}{hours\ of\ simulated\ time} \quad (6)$$

This provides a false alarm rate per hr. False alarms are counted when the algorithm detects that an incident has occurred and yet no incident has occurred at that time. Alarm triggers related to an initial false alarm are counted only once. This is similar to the case of

detecting true incidents that cause many triggers—an incident can only be detected once. For this calculation the warm-up times were deleted from the hours of simulated time. For the second definition false alarm rates can be expressed as a percentage of false alarms over the number of times the algorithm is repeated:

$$FALSE\ ALARM\ RATE\ (def\,2) = \frac{\#inc_{false}}{\#algorithm\ repetitions} \quad (7)$$

The algorithm is used after an appropriate warm-up time. Finally, the average time to detect is calculated by

$$AVERAGE\ TIME\ TO\ DETECT = \frac{\sum time_{det}}{\#inc_{det}} \quad (8)$$

where $time_{det}$ is the time until the algorithms first declare a true incident from the time an incident is simulated to begin. Incidents that are not detected within 6 min are considered undetected (affecting the detection rate), and subsequent detections are considered false alarms. Thus, this statistic includes only those true incidents detected in less than 6 min.

Table 1 shows quantitative results of the VRC-based algorithms described here compared with our implementation of California Algorithm #7, a typically implemented pattern-based inductive loop-based algorithm. All three algorithms seem to be successful compared with the California Algorithm. It is important to consider that a different simulator, different flows, different detector spacings, and/or different percentages of vehicles used likely would lead to different numbers—what is important is the relative performances of the values.

These results have been obtained for all four algorithms (Headways Algorithm, Lane Switches Algorithm, Lane-Monitoring Algorithm, and California Algorithm) with the thresholds that jointly maximize the detection rate and minimize the false alarm rate and detection time. One expects a linear relationship or high positive correlation between detection rate and false alarm rate and between detection rate and time-to-detect. Previous research shows graphs with such relationships (*7*). However, this does not necessarily seem to be the case with the VRC-based algorithms. Different combinations of thresholds often led to one performance measure remaining fairly constant while the others fluctuated.

Several results may be ascertained. The table shows that although the false alarm rates for the Headways Algorithm and the California Algorithm are similar, the Headways Algorithm clearly performs better in terms of detection rate and average time to detect. If average time to detect is not of great concern (the difference between 2 and 3 min may not be large when an incident impacts traffic for 40 min or more), then the Lane Switches Algorithm clearly performs well. Finally, the Lane-Monitoring Algorithm works quickly with a high detection rate but a relatively high false alarm rate.

TABLE 1 Most Promising Evaluation Results

Algorithm Name	Detection Rate (#incidents detected/true incidents)	False Alarm Rate (def 1) (false alarms/ hour)	False Alarm Rate (def 2) (false alarms/ algorithm repetitions)	Average Time To Detect (minutes)
Headways Algorithm	0.75	1.30	0.0195	2.00
Lane Switches Algorithm	0.94	0.65	0.0098	2.94
Lane Monitoring Algorithm	0.92	2.20	0.0330	1.73
California Algorithm #7	0.53	1.05	0.0158	2.19

Qualitative Results

Although the numerical results seem promising, it is also valuable to describe the performance of the algorithms with respect to incident types, traffic flows, and incident location with respect to the detectors. It is helpful to note that slight modifications may improve results.

Headways Algorithm

The detection rate (~75%) was constant for the different flows tested. This algorithm is successful in detecting a short-duration, single-vehicle stall type incident. The algorithm seems to be insensitive to incident location with respect to the detectors. As expected, detection times are reduced as flow increases. False alarms seem to be uncorrelated to flow levels or incident types. This algorithm may work better if the spatial comparison is made between the reader and the reader downstream from it rather than upstream—at the downstream reader traffic would be flowing more normally.

Lane Switches Algorithm

This simple algorithm performs remarkably well. Although the time-to-detect is slow, the time-to-detect is lower for lower flow levels, which is the inverse of the performance of typical algorithms. As expected, detecting stalls requires the longest detection time. Also as expected, the closer the downstream detector to the incident location, in general, the quicker the detection. A simple modification such as counting switches across two lanes as two switches rather than one switch may improve the results significantly.

Lane-Monitoring Algorithm

Despite the high false alarm rate, this algorithm shows promise. In general, the detection rate, false alarm rate, and time-to-detect values are not correlated with the type of incident, traffic flow level, or location of the incident with respect to the readers. This robustness is very attractive in an incident detection algorithm. The algorithm is then appropriate for many different situations. This algorithm may have potential as a back-up or secondary algorithm in an incident detection system.

CONCLUSIONS

All three of these algorithms perform reasonably. They show that VRC has great potential as a stand-alone sensor for incident detection. The algorithms perform better or as well as expected and decisively better than the typically used California Algorithm. The algorithms' robustness to various situations make them additionally appropriate. Future research as described in the next section may enhance the performance of these and other VRC-based algorithms.

FUTURE RESEARCH

The following paragraphs introduce several ideas to spark future research. These include extensions to the current research, combining VRC data with data from other traffic sources, considering other algorithm methodologies, and performing a cost-benefit analysis of algorithms that use VRC compared with algorithms that use other sensors.

There are several obvious extensions to this research. These include testing the described algorithms with field data. A complete field test would provide the best indication of the algorithms' performance. Another extension proposed by this research but that has not been explored fully is to use thresholds that are functions of the flow. Although calibrating such equations would not be easy, threshold-flow functional relationships would reduce dramatically the overall calibration effort required. The threshold function may also have the percentage of vehicles tagged and/or types of vehicles tagged as parameter(s). Also, additional work may be performed in calibrating algorithm parameters considering detector spacings, configurations, and roadway design with both simulated and field data. There is assumed to be a relationship between sensor spacing and mean time to detect [closer spacing, lower time to detect (*21*)], thus spacing as close as financial and human factors constraints allow should be best. But it is possible that spacing too close together causes an intolerable false alarm level. Similarly, it is important to investigate how different percentages of tagged vehicles and different types of tagged vehicles will change the algorithms' parameters and performance. For example, in general, higher percentages of vehicles would result in better results. Additionally, some systems may have only heavy vehicles tagged, which would introduce a bias into the data and should be considered and weighted as such.

One of the more exciting future research topics is combining VRC data with data from other detector types. It is likely that VRC systems will be installed on systems already outfitted with inductive loops. The spatial, microscopic data from VRC can be combined with inductive loop data, or other point data sources such as infrared and ultrasonic detectors, representing all vehicles (VRC-tagged and nontagged) to obtain better parameters for use in incident detection algorithms. It is likely that even better parameters would result if VRC data were combined with video or radar images that produce density and other spatial information.

Currently, implemented incident detection algorithms are all pattern-based. However, there are several other proposed methodologies for incident detection. These include statistical methods including times series and filtering, the application of catastrophe theory or artificial neural networks, and the use of a traffic flow model. Some of the proposed VRC algorithms are statistically based. VRC data may require less computational power to obtain the space mean speed and density needed in traffic flow models. Model-based incident detection algorithms are expected to work better than other incident detection algorithms because the traffic flow model more accurately represents traffic and thus can determine better whether the traffic flow is non-normal, hence, that an incident has occurred. VRC can be used with any of these methodologies.

It would be beneficial to do a cost-benefit analysis comparing algorithms that use VRC with algorithms that use other sensors and other incident detection methods. Implemented inductive loop-based algorithms, proposed inductive loop algorithms, the VRC algorithms described here, other VRC algorithms, processed video algorithms, CCTV scanning by traffic management center operators, and cellular phone calls from drivers should all be compared. Such an analysis would include benefits described by the performance measures of detection rate, false alarm rate, time-to-detect, public awareness and acceptance, and ease of operation of the algorithm or system by the traffic management center. The costs should include the costs of the sensors, software development, maintenance, and the personnel required to run the system.

This paper has provided an initial contribution to the use of vehicle to roadside communication sensors for incident detection. After discussing the use of vehicle to roadside communication sensors as the surveillance sensor for incident detection, the paper has concentrated on three pattern-based algorithms for detection that show great promise. The above paragraphs suggest some of the many directions to follow to continue the research.

ACKNOWLEDGMENT

This work was performed when both authors were at the Massachusetts Institute of Technology.

REFERENCES

1. Lindley, J. A. Urban Freeway Congestion: Quantification of the Problem and Effectiveness of Potential Solutions. *ITE Journal*, January 1987, pp. 27–32.
2. Drummond, J. T. A New Era in Road Policy. *Nation's Business*, September 1991.
3. TRANSCOM. TRANSCOM ETTM IVHS Operational Field Test: Final Feasibility Report, TRANSCOM Contract No. XCM 92-090.01, 1993.
4. Hallenbeck, M. E., T. F. Boyle, and J. Ring. Use of Automatic Vehicle Identification Techniques for Measuring Traffic Performance and Performing Incident Detection. Washington State Transportation Commission and Transportation Northwest, Olympia, Wash., June 1992.
5. Ivan, J. N., J. L. Schofer, C. R. Bhat, P.-C. Liu, F. S. Koppleman, and A. Rodriquez. Arterial Street Incident Detection Using Multiple Data Sources: Plans for ADVANCE. *Proc. Pacific Rim TransTech Conference/3rd ASCE Applications of Advanced Technologies in Transportation Engineering Conference: Volume 1*, 1993, pp. 429–435.
6. Bernstein, D., I. El Sanhouri, and E. Parkany. An Integrated System for Peak-Period Pricing, Incident Detection and Route Guidance Using Automatic Vehicle Identification. *Proc. 2nd IBTTA International Symposium on ETTM Technology*, 1993.
7. Stephanedes, Y., A. P. Chassiokos, and P. Michalopoulos. Comparative Performance Evaluation of Incident Detection Algorithms. In *Transportation Research Record 1360*, TRB, National Research Council, Washington, D.C., 1992, pp. 50–57.
8. Chen, C.-H, and G.-L. Chang. A Review of Recent Freeway Incident Detection Algorithms. DTFH61-92-R-00122 (Federal Highway Administration Incident Detection Request for Proposals) Attachment No. 9, 1992.
9. Bernstein, D. and A. Kanaan. Automatic Vehicle Identification: Technologies and Functionalities. *IVHS Journal,* Volume 1, No. 2, 1993.
10. Michalopoulos, P. Automatic Incident Detection through Video Image Processing. *Traffic Engineering and Control*, Volume 34, No. 2, February 1993, pp. 66–75.
11. Autlman-Hall, L., F. L. Hall, Y. Shi, and B. Lyall. A Catastrophe Theory Approach to Freeway Incident Detection: Applications of Advanced Technologies in Transportation Engineering. *Proc. 2nd International Conference*, sponsored by ASCE, Minneapolis, Minn., August 1991, pp. 373–377.
12. Ritchie, S. G. and R. L. Cheu. Simulation of Freeway Incident Detection Using Artificial Neural Networks. *Transportation Research C*, September 1993, pp. 203–218.
13. Ahmed, S. A. and A. R. Cook. Time Series Models for Freeway Incident Detection. *ASCE Journal of Transportation Engineering,* Vol. 106, No. 6, November 1980, pp. 731–745.

14. CSST. Development of an Incident Detection Algorithm. Report for DRIVE Project, 1991.
15. Dudek, C. L., C. J. Messer, and N. B. Nuckles. Incident Detection on Urban Freeways. In *Transportation Research Record 495*, TRB, National Research Council, Washington, D.C., 1974, pp. 12–24.
16. Payne, H. J., E. D. Helfenbein, and H. C. Knobel. Development and Testing of Incident Detection Algorithms. Vol. 2. Research Methodology and Detailed Results. Report No. FHWA-RD-76-20, FWHA, U.S. Department of Transportation, 1976.
17. Parkany, A. E. *Using Vehicle to Roadside Communication Data for Incident Detection*. M.S. thesis. Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, 1993.
18. Parkany, A. E. and D. Bernstein. Using VRC for Incident Detection. *Proc. Pacific Rim TransTech Conference/3rd ASCE Applications of Advanced Technologies in Transportation Engineering Conference: Volume 1*, 1993, pp. 63–68.
19. Yang, Q. A Microscopic Traffic Simulation Model for IVHS Applications. M.S. thesis. Massachusetts Institute of Technology, Cambridge, 1993.
20. Hotz, A. F., M. Ben-Akiva, D. Bernstein, A. Chachich, R. Mishalani, N. V. Jonnalagadda, Q. Yang, H. Koutsopoulos, R. W. Brindley, D. Krechmer, and C. T. Marcus. Evaluating Traffic Control Systems Using Microsimulation: The CA/T IPCS. *Proc. IVHS America Conference*, 1994.
21. Klein, L. A., N. A. Rantowich, C. C. Jacoby, and J. Mingrone. IVHS Architecture Development and Evaluation Process, *IVHS Journal,* Volume 1, No. 1, 1993.

Publication of this paper sponsored by Committee on Transportation System Management.

Examining the Potential of Using Ramp Metering as a Component of an ATMS

BRUCE HELLINGA AND MICHEL VAN AERDE

The current emphasis on utilizing existing transportation infrastructure more efficiently has added impetus to the recent focus on advanced traffic management systems. An advanced traffic management system typically combines existing hardware, software, and traffic engineering expertise to observe and manage transportation systems more effectively. The potential of ramp metering to provide reductions in system delay has been recognized for some time. Simple analytical techniques have been used to demonstrate the magnitude of these benefits. However, these analytical methods can rarely reflect fully all the spatial and temporal dynamics that may exist within integrated freeway and arterial networks. In this paper a network traffic simulation model is used to examine the potential benefits of implementing ramp metering strategies and to quantify how sensitive these benefits are to a number of factors including the metering rate, the timing of the implementation of the metering, and various assumptions regarding driver rerouting behavior. The results of this investigation indicate that, as expected, ramp metering can result in reductions in total travel time, but it may also yield increased net delays if it is not implemented correctly. This investigation indicated that the temporal window of opportunity during which ramp metering can be implemented and be of benefit is surprisingly small. Results for a simple network indicate that under ideal conditions, in which drivers are able to divert their routes, a benefit of as much as a 14 percent reduction in total travel time may be possible. If it is assumed that a capacity loss of 5 percent occurs once the freeway becomes congested, then the benefit of metering may be as great as a 26 percent reduction in total travel time.

With the recent emphasis on advanced traffic management systems, ramp metering is receiving considerable attention as a traffic management strategy. This consideration is not new. Ramp metering has been considered for more than four decades as a traffic management technique. During this time, many forms of ramp metering have been examined, including simple fixed time metering (*1*) and real time responsive metering (*2*). Much effort has also been expended on researching methods of optimizing the metering rates of isolated meters (*3*) and systems of coordinated meters (*4*).

The earliest works utilized linear programming techniques to determine optimal time-of-day metering rates (*5*). Most of these optimization strategies assume freeway throughput as the objective function. However, few of the evaluation methods explicitly consider the possibility that route diversion may take place. One recent notable exception is the work carried out by Nsour et al., (*6*), who utilized the INTRAS simulation model to examine the impacts of ramp metering with and without diversion. Based on their simulation of an 11.2-km section of freeway in California, Nsour et al. concluded that a 10.5 percent reduction in system delay could be obtained under ideal metering and diversion conditions. However, as diversion rates were prespecified, drivers of vehicles did not have the ability to make routing decisions based on currently available estimates of alternative route travel times. Little research has been conducted to identify and quantify the net benefits of ramp metering when realistic route diversion is considered. More importantly, the sensitivity of these benefits to various control parameters has typically not been examined.

Department of Civil Engineering, Queen's University, Kingston, Ontario K7L 3N6, Canada.

PURPOSE OF RAMP METERING

In most ramp metering analyses the intended purpose of utilizing ramp metering is the avoidance of flow breakdown on the freeway. To meet this goal, the capacity of each freeway segment is determined, demands are estimated, and metering rates are imposed such that the freeway operates without congestion. This process is rather straightforward and is described elsewhere in the literature (*3*). However, the process of quantifying the net benefits of ramp metering is more difficult. Reduced delay is often considered to be the primary benefit of ramp control; nevertheless, impacts on fuel consumption, emissions, and safety may also be components of the net benefit.

POTENTIAL BENEFITS OF RAMP CONTROL

To achieve the goal of ramp control, that of reducing total network travel time, it is necessary to identify, and then seek to satisfy, a number of specific objectives. One such objective might be the reduction of the size of queues on the freeway by controlling ramp access. Another objective might be the improvement of average freeway speed by ensuring that the freeway operates in an uncongested mode. One might even desire a freeway speed that is higher than the speed at capacity, thereby requiring a more restrictive metering rate.

For safety considerations, a potential objective might be the reduction in the variability of vehicle speeds. For delay and throughput considerations, a potential objective might be the avoidance of freeway queues spilling upstream and blocking access to some heavily utilized exit ramps. Ramp control can also be used to avoid capacity reduction effects that occur when flow breaks down or to encourage spatial, temporal, and modal diversions to other roads, times, and modes having lower marginal system costs.

Each of these potential ramp metering objectives may have a different impact on net system benefits. An evaluation of the impact on benefits of many of these potential objectives is usually too complex to be carried out adequately by standard analytical techniques. In this paper we examine a number of these potential objectives and evaluate their impacts on network travel time, using the INTEGRATION simulation model version 1.5c and, where possible, analytical techniques.

BASE CASE: NO METERING

In this section we present the example network used to demonstrate the relative benefits and drawbacks of ramp metering. The network characteristics, origin–destination demands, and speed-flow relationships are provided. An analytical analysis of these base case traffic conditions is conducted. Simulation results reflecting network traffic conditions are examined. Finally, analytical and simulation results are compared.

Example Network

The example network, illustrated in Figure 1, consists of a freeway section that has two identical junctions and a parallel arterial. There are 6 origin–destination zones, 47 nodes, and 64 directional links. All links are 0.5 km in length, except for arterial link 44, which is 5.15 km long. Each freeway link consists of two lanes, whereas all other links consist of a single directional lane.

The network was intentionally made to be simple for two reasons. First, to permit analytical analyses of traffic conditions, the network could not be designed to be too complex. Second, the intent of this study is to examine the impacts of a number of factors on total delay with the express purpose of identifying and illustrating the relative impacts of factors that affect ramp metering. The intent is to quantify the relative impacts, not the absolute ones. Furthermore, this study serves as an initial effort, and, as described in the Conclusions section, further research should be conducted.

In this paper ramp metering at only the second ramp junction (links 53 and 54) is examined. Subsequent research will examine the impacts and implications of metering at both ramp junctions. The ability to use the same network configuration is of benefit, as it will permit results to be compared directly with those described in this paper.

The origin–destination demands initially imposed on the network were the following: 3400 vehicles per hour (vph) from zone 1 to zone 4, 500 vph from zone 1 to zone 3, and 800 vph from zone 6 to zone 4. Initially there are no other demands on the network.

To determine the progression of traffic through the network and to quantify travel time, a single regime speed-flow–density relationship (*7*) was utilized. The freeway speed-flow relationship is nonparabolic and is characterized by a free speed of 105 km/hr, a capacity of 2000 vph/lane, a speed at capacity of 80 km/hr, and a jam density of 100 vehicles/km/lane.

Analytical Evaluation: No Metering

On the basis of the network characteristics, traffic demands, and specified speed-flow relationships, it is possible to carry out an analytical evaluation of expected traffic conditions by standard shock wave analysis. Figure 2 provides the graphical results of this analysis.

At time 0, traffic begins to enter the network from zone 1 at a rate of 3900 vph and with a speed of 90 km/hr. Approximately 7 min are required for the leading edge of this platoon to reach the on-ramp, 10 km downstream of zone 1.

Because the on-ramp already contributes a flow of 800 vph, there is only 3200 vph of remaining freeway capacity. As there is a demand of 3400 vph, a queue begins to form at an initial rate of 200 vph (200 = 800 + 3400 − 4000).

It must be determined whether the queue forms on the freeway, on the on-ramp, or on both. It is assumed that downstream capacity is apportioned to upstream demand in proportion to the upstream capacity. The on-ramp consists of a single lane with a capacity of 1600 vph. The freeway consists of two lanes, each with a capacity of 2000 vph. Therefore it is assumed that, of the 4000-vph downstream capacity on the freeway, 1143 vph (4000 × 1600/[1600 + 4000]) is apportioned to ramp demand, and the remaining capacity (4000 − 1143 = 2857 vph) is apportioned to upstream freeway demand. In this example, because the ramp demand of 800 vph is less than the ramp's share of the downstream freeway capacity (1143 vph), none of the 200 vph excess demand is considered to queue on the on-ramp.

When the queue spills back upstream 0.5 km, direct access to the off-ramp is blocked. The flow that can be accommodated upstream of the off-ramp is a function of the downstream capacity flow (3200 vph) and of the number of vehicles that will flow onto the off-ramp. The 500-vph demand attempting access to the off-ramp constitutes 12.8 percent (500/3900 × 100) of the total freeway flow. Therefore it can be expected that a flow of 3670 vph (3200/[1 − 0.128]) can be accommodated upstream of the off-ramp. From this point the queue grows at an accelerated rate of 230 vph (3900 − 3670). The queue continues to grow until the demand is stopped after 1 hr. The maximum length of the queue is computed to be 7.5 km, and the total system travel time is estimated to be approximately 855 vehicle-hr.

Simulation Results: No Metering

The INTEGRATION simulation model is a microscopic traffic simulation model capable of modeling integrated networks, various traffic control devices, and advanced route guidance systems (*8,9*). The INTEGRATION model has been used to model a number of hypothetical and real networks (*10,11*) and is suited for use in evaluating the effectiveness of traffic control devices, including ramp meters. For the base case, no traffic control devices were modeled, and routes were prespecified such that all traffic utilized the freeway.

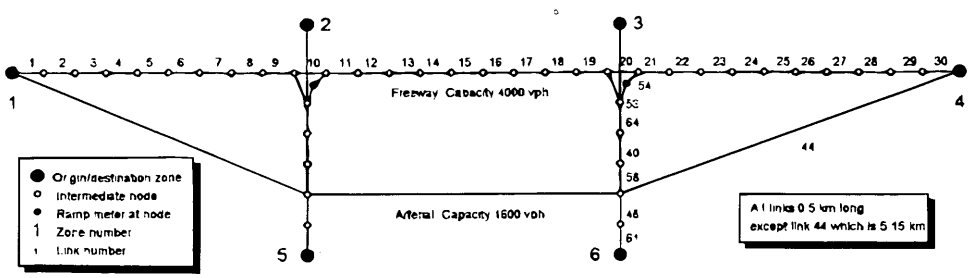


FIGURE 1 Example network structure.

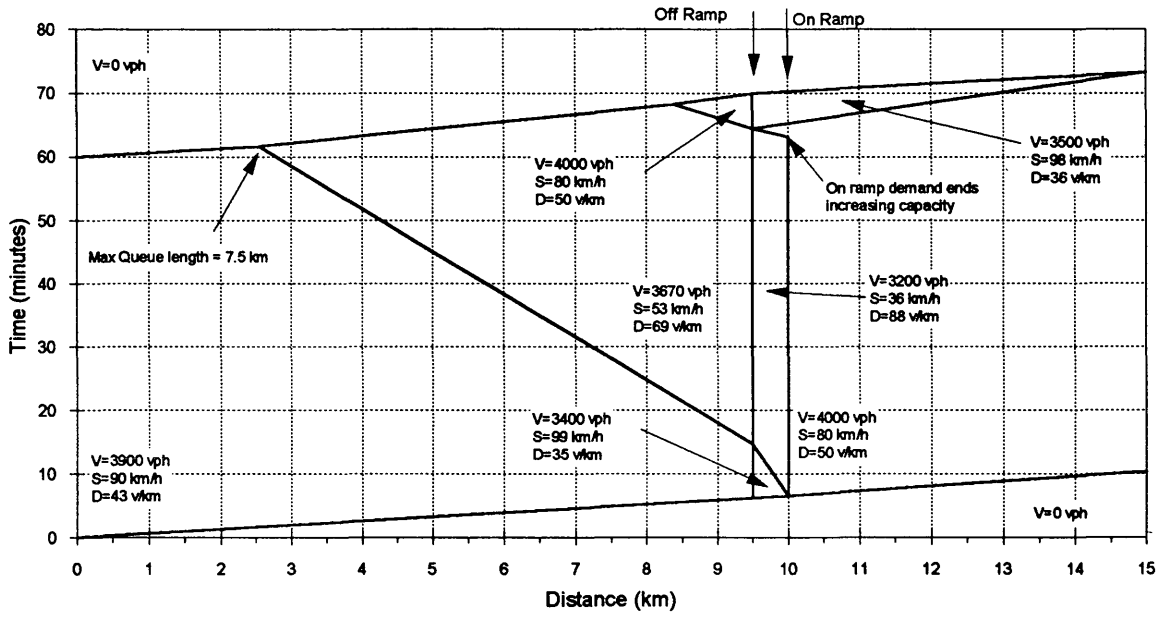


FIGURE 2 Shock wave analysis of nonmetering traffic conditions.

The model initiated simulation at time 0 with an empty network. Although demands were active for only 1 hr, the network was simulated for 80 min to permit all vehicles to reach their destinations. The model was configured to output average speed and volume every 5 min for each link in the network. At the end of each completed simulation run the total network travel time was computed.

Figure 3 shows the temporal and spatial variation in average 5-min freeway speed estimated by the simulation model. The significant region of speeds less than the speed at capacity (80 km/hr) is indicative of the congestion that occurs upstream of the on-ramp. It is also evident that after period 12 (1 hr into the simulation) the estimated speeds at the upstream end of the freeway section (low link numbers) return to the free speed value. This recovery occurs because the inflow of new demand has ceased and the network is simply emptying any vehicles that are already on the network.

Total network travel time incurred by the 4700 vehicles was 817.6 hr, or an average of 10.4 min per vehicle.

Comparison of Analytical and Simulation Results

Before using the simulation model to carry out sensitivity analyses, it is instructive to compare model results with analytical solutions.

Temporal and spatial variation in speed estimated by the simulation model (Figure 3) and analytically computed (Figure 2) can be qualitatively and quantitatively compared. Some interpolation of simulation results is necessary, as simulation results are requested from the model only at 5-min intervals. Both the analytical and simulation results indicate a triangular region of congestion. The analytical solution indicates that congestion begins at approximately 7 min and ends at time 70 min. The maximum length of queue is 7.5 km. The simulation results indicate that congestion exists after 15 min but does not yet exist after 10 min. Inasmuch as simulation results are produced at 5-min intervals, it is necessary to interpolate to estimate more precisely when congestion occurred. On the basis of Figure 3, the time at which congestion occurs is estimated to be 14 min. Congestion ends at time 67 min, and the maximum length of queue is approximately 5.7 km.

These results indicate that the simulation model predicts a shorter period during which traffic flow is congested than does the analytical solution. The main cause for this discrepancy is the different manner in which the initial flow is assumed to traverse the empty network.

The analytical solution assumes that a platoon of traffic, having some constant volume and remaining within either the congested or the uncongested flow regime, travels as a homogeneous unit at some constant speed. Figure 2 indicates that the shock wave begins at time 0 and distance 0. This shock wave has a constant speed and represents the boundary between a regime that has a flow of 3900 vph (density of 43 vehicles/km) and one that has no flow.

In reality, the speed of a vehicle is determined more microscopically by the level of freedom that the driver of the vehicle has to maneuver. The presence of upstream vehicles generally does not affect the speed of downstream vehicles. Therefore it would be expected that the first vehicle to depart zone 1 would do so at approximately free speed. Subsequent vehicles would travel marginally slower as each additional vehicle increased the impedance of upstream vehicles. The result, then, is platoon dispersion, as vehicles that are first to enter the empty network have a higher speed than those entering later, even though all vehicles enter at a constant rate of 3900 vph. As this platoon travels the 10 km to the on-ramp, this dispersion effect is magnified such that the flow rate of the downstream end of the platoon is much lower than 3900 vph. In fact, during the simulation, the flow on link 19, upstream of the off-ramp, did not reach 3900 vph until period four, 20 min into the simulation.

Certainly, the assumption within the analytical approach that vehicle speeds are based on the macroscopic flow rate of upstream vehicles is less realistic. However, because of the additional complexity, it is very difficult to capture this effect of dispersion in an analytical methodology.

On the basis of this comparison, suitable explanations exist for the discrepancies between the model and analytical results. These

Link #	Time Interval (each of 5 minute duration)															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	91	91	91	90	91	91	90	91	91	90	91	91	104	104	104	104
2	91	89	89	88	89	89	88	89	89	88	89	89	104	104	104	104
3	93	89	89	89	88	89	89	88	89	89	88	89	104	104	104	104
4	94	90	90	90	89	90	90	89	90	90	89	90	104	104	104	104
5	94	90	90	90	89	90	90	89	90	90	89	90	104	104	104	104
6	95	90	90	90	89	90	90	89	90	90	89	90	104	104	104	104
7	96	91	89	89	90	89	89	90	89	89	90	89	104	104	104	104
8	97	92	90	90	90	90	90	90	90	90	90	90	104	104	104	104
9	98	92	89	90	90	89	90	90	89	90	90	75	104	104	104	104
10	99	93	89	89	90	89	89	90	89	89	90	54	104	104	104	104
11	100	93	90	89	89	89	90	89	89	90	66	51	104	104	104	104
12	100	93	91	90	90	90	90	90	90	71	53	51	95	104	104	104
13	101	94	91	89	89	89	89	89	80	52	52	49	54	104	104	104
14	102	95	92	89	90	89	89	86	52	52	52	49	54	104	104	104
15	103	95	92	90	89	89	90	55	50	52	52	49	54	104	104	104
16	104	95	92	90	90	89	65	55	50	52	52	49	54	104	104	104
17	104	96	92	91	90	83	53	53	49	51	52	49	54	104	104	104
18	105	96	93	91	89	56	52	53	49	51	52	49	54	104	104	104
19	105	97	93	91	65	53	51	51	46	51	51	48	58	104	104	104
20	105	97	76	47	35	36	35	36	35	36	38	36	51	104	104	104
21	104	92	83	82	82	81	80	81	80	80	80	80	82	104	104	104
22	104	94	88	82	81	81	80	80	81	80	80	80	80	104	104	104
23	104	95	90	83	81	80	83	80	81	81	81	82	81	104	104	104
24	104	96	89	85	81	81	81	82	81	80	81	81	80	92	104	104
25	104	97	90	87	84	83	81	81	80	80	81	81	80	81	104	104
26	104	98	91	86	87	83	82	82	81	82	82	81	82	83	104	104
27	105	99	92	89	85	82	82	82	82	81	81	81	81	81	104	104
28	105	100	92	89	87	86	84	81	81	82	82	82	81	81	104	104
29	105	100	93	89	87	86	84	83	81	81	82	81	81	81	104	104
30	105	101	93	90	88	85	82	82	81	81	81	82	81	82	104	104

FIGURE 3 Simulation model speed estimates (km/hr) for all freeway links for nonmetered traffic conditions.

discrepancies appear to arise from simplifying assumptions that are made in the analytical approach but that are not made by the simulation model.

IMPACTS OF RAMP METERING

Having determined expected traffic conditions when no ramp control is in place, we are interested in determining the impact that ramp controls may have. Operational advanced traffic management systems utilize a wide range of ramp metering control strategies, from fixed metering rates to more complex ramp metering control strategies in which metering rates are determined on-line as a function of the freeway traffic conditions, the minimum and maximum metering rates, and queue spillback constraints. However, in this paper it is assumed that a fixed-rate time-of-day metering control is to be used. Because rates are fixed, no consideration is given to rate modification as the result of queue spillback. Before evaluation, the metering rate and the time period during which metering should take place must be determined.

Typically, ramp metering rates are set such that maximum utilization of the freeway is achieved without the occurrence of congestion. Analytically, it is quite clear that the ramp demand of 800 vph exceeds the available freeway capacity by 200 vph. Therefore a metering rate of 600 vph, or one vehicle every 6 sec, could be used. To avoid incurring unnecessary delay, metering should not begin until the traffic flow on the freeway at the on-ramp reaches 3400 vph. Furthermore, metering should continue only as long as the freeway flow remains at this level. For this analysis it is assumed that no spatial, temporal, or modal diversion occurs.

The INTEGRATION model's ability to represent traffic signals was utilized to emulate fixed-time ramp meters. On the basis of results for the premetering case, ramp metering controls were initiated at 800 sec, as this is the time at which flow on the freeway (at the on-ramp) reaches 3400 vph. Metering continued until time 4000 sec, at which time the flow on the freeway (at the on-ramp) dropped to zero.

Figure 4 illustrates the temporal variation in average 5-min speed for three freeway links adjacent to the on- and off-ramps. It is clear from this figure that speeds never fall below 80 km/hr, the speed at

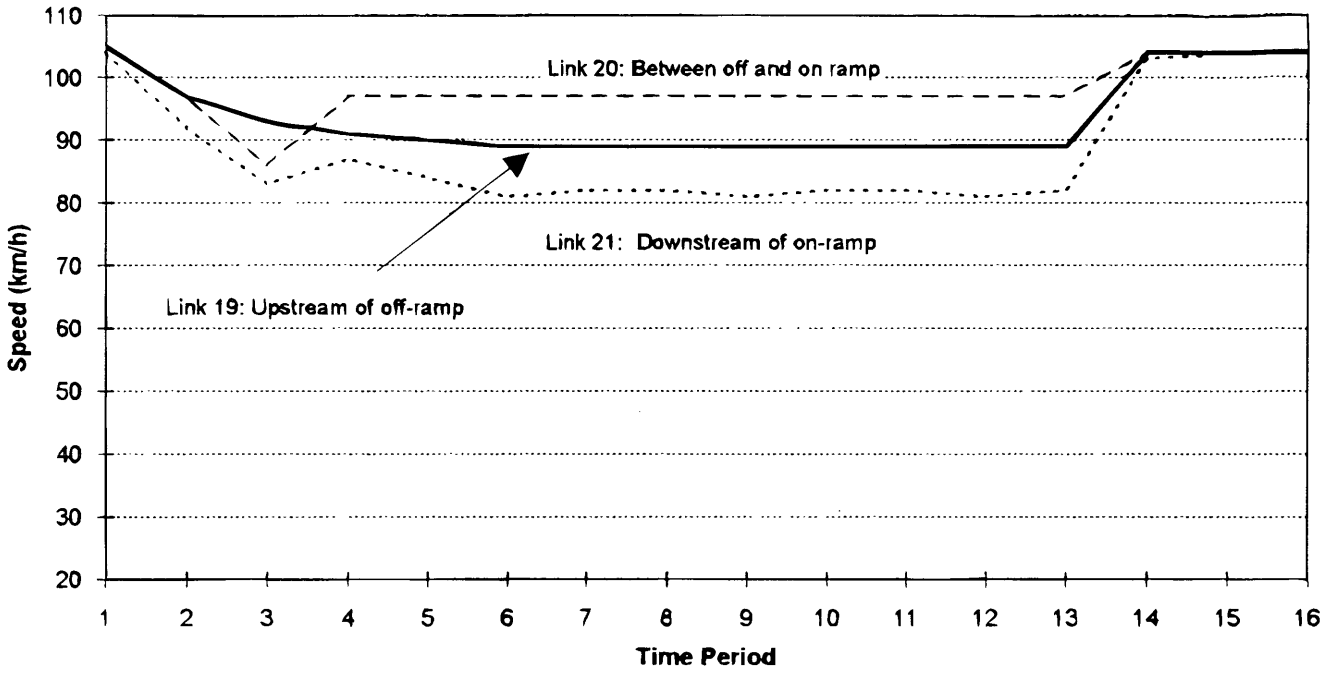


FIGURE 4 Simulation model estimates of temporal variation in average 5-min speeds for freeway links adjacent to the on-ramp when on-ramp flows are metered at 600 vph.

capacity. This indicates that this section of freeway never operates in the congested regime of the speed–flow relationship. The effect of ramp control can clearly be seen when the link speeds depicted in Figure 4 are compared with those provided in Figure 3.

Total network travel time was found to be 814.4 vehicle-hr, a travel time reduction of only 0.39 percent over that in the premetering case.

Because the computed benefit was rather small compared with benefits reported in the literature, it was decided to investigate those factors that might affect these benefits. This investigation is described in the next two sections.

SENSITIVITY ANALYSIS: NO DIVERSION

To identify those factors that might have a significant impact on the benefits of ramp metering, a series of sensitivity analyses was performed. In each case the base condition is the metering condition discussed in the previous section.

Four factors that affect ramp metering benefits were examined: the timing of the ramp control, the metering rate, the capacity drop effects, and the origin-destination (O-D) demands. Each of these is discussed in turn in the following sub-sections.

Effect of Timing of Implementation

The time of implementation of ramp control was found to have a significant impact on the estimated benefits of ramp metering. Figure 5 illustrates the variation in total network travel time with changes in the implementation time of the ramp metering. For this evaluation only the time at which metering began was altered; all other conditions remained unchanged from those discussed in the previous section, including the duration of the period for which metering was in effect. Figure 5 indicates that, for the effective conditions here, initiating ramp metering just 2 min earlier than optimal can negate any metering benefits.

Beginning metering later than optimal does not have such a significant effect. In fact, one would expect that if metering were begun approximately 1 hr after the optimal time, then all demands would have already passed the ramp and the result would be the the same as for the premetering situation.

The implication of Figure 5 is that fixed metering plans, which invoke metering at prespecified times of day independently of actual main-line flows, may cause a net increase in total delay if metering begins before the freeway flows reach capacity.

Effect of Metering Rate

We examined the effect of the actual metering rate by varying the ramp signal cycle length within consecutive runs of the simulation model. Six metering rates, ranging from one vehicle/8 sec (450 vph) to one vehicle/3 sec (1200 vph), were evaluated. Figure 6 illustrates the impact of metering rate on total network travel time. Clearly, metering rates that are more restrictive than necessary to prevent flow breakdown result in rather significant increases in total travel time. Under these conditions the additional delay incurred by traffic utilizing the on-ramp far outweighs the travel time savings experienced by freeway users as the result of the slightly higher freeway speed.

Interestingly, a metering rate of 720 vph results in a marginally lower total travel time than for all other rates. It must be remembered that capacity drop effects are not considered in this analysis. Inasmuch as there is no additional penalty incurred when flow breakdown occurs, total travel time is minimized when queuing occurs on both the freeway and the arterial.

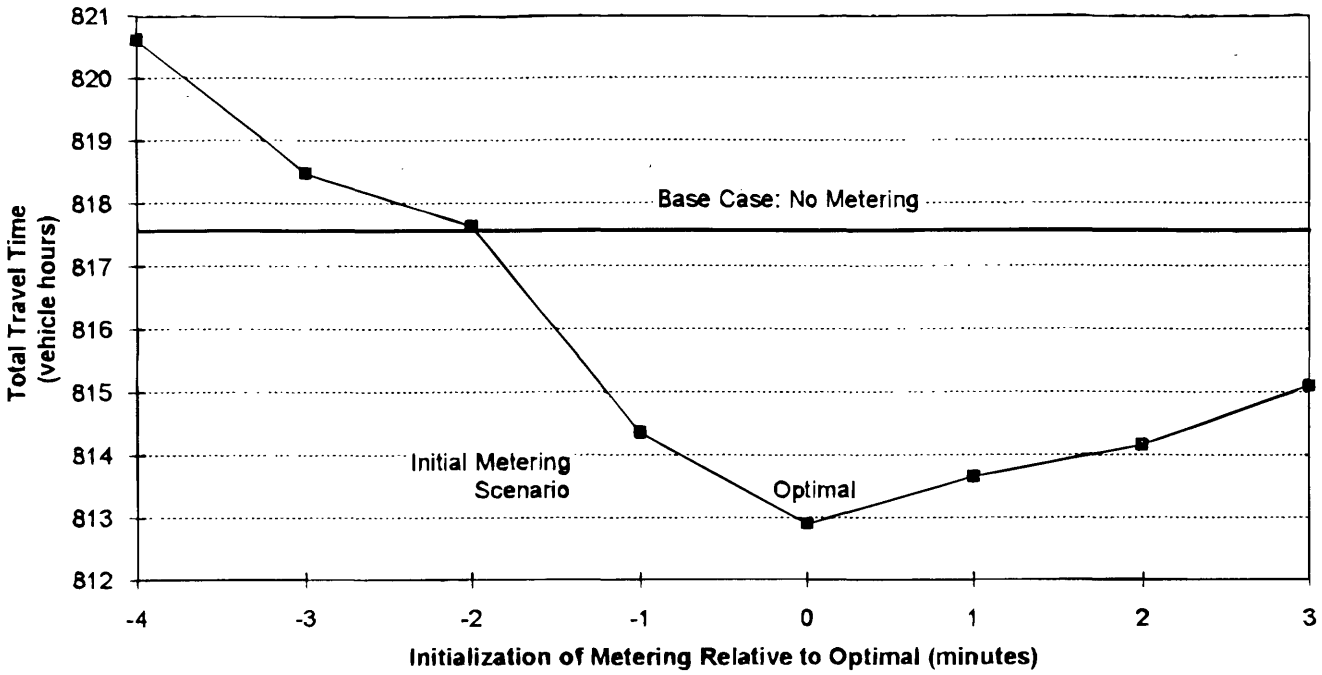


FIGURE 5 Effect of ramp control initialization time on total network travel time when diversion is not considered.

Effect of Capacity Drop

There has been some debate over whether freeway capacity is reduced after flow breakdown occurs (*12,13*). It is not our intent in this paper to add to this debate. Rather, our intent is to examine the impact that this phenomenon might have on modifying the potential benefits of ramp metering.

Proponents of the capacity drop concept indicate that, once flow breakdown occurs on the freeway, the capacity is reduced from precongested conditions to a lower congested value and is not restored to the precongested conditions until the freeway flow is again uncongested. It is not clear, however, what the potential magnitude of this capacity loss is. It has been stated that capacity reductions of as much as 25 percent may be possible (*14*).

This capacity loss was replicated in the INTEGRATION model by introduction of an incident to reduce the capacity of the freeway immediately upstream of the on-ramp. This capacity reduction was implemented at the onset of congestion and remained in effect until the freeway became uncongested.

To explore the potential effects of capacity reduction, we selected a modest reduction of 5 percent. Without metering, the effective capacity of the freeway immediately upstream of the on-ramp is

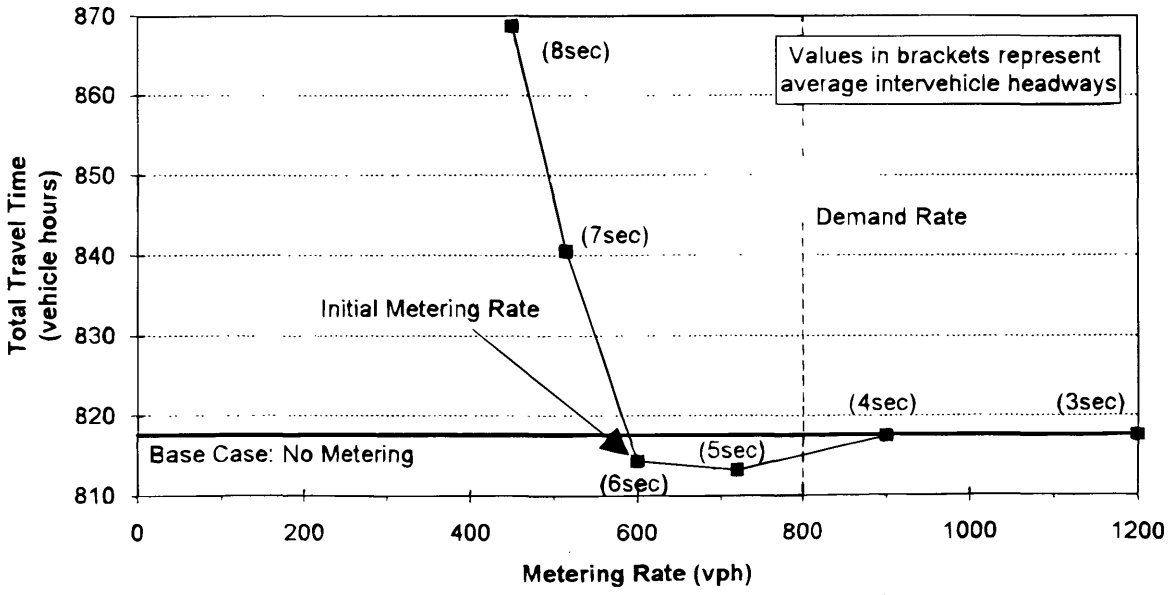


FIGURE 6 Effect of ramp metering rate on total network travel time when diversion is not considered.

3200 vph, as the remaining 800 vph is utilized by on-ramp flow. A 5-percent capacity reduction of 3200 vph is 160 vph. After the on-ramp flow ceases, the capacity of the freeway returns to 95 percent of 4000 vph, or 3800 vph. When the remaining queue on the freeway is served and flow becomes uncongested, capacity returns to the full 4000 vph.

The total network travel time associated with this model was simulated to be 951.6 vehicle-hr. This delay can be compared with the initial premetering model presented above that resulted in a total travel time of 817.7 vehicle-hr. Thus, the occurrence of a 5-percent reduction in capacity during congestion is estimated to cause a 16.4-percent increase in total travel time for this example network.

If the capacity drop phenomenon is to be considered part of the base case model, the ramp metering benefits will suddenly have the potential to be much larger. Specifically, a modest travel time reduction from 817.6 to 814.4 vehicle-hr (3.9 percent) would suddenly become a reduction from 951.6 to 817.7 vehicle-hr (14.1 percent).

Effect of O-D Demand

The absolute magnitude of metering benefits is also known to depend on the characteristics of the network in question, the operation of the ramp controls, the O-D demands on the network, and the availability and the quality of alternative routes.

O-D demands can have significant effects, particularly on ramp metering benefits, when queues that form on the freeway when metering does not exist spill back upstream and block access to upstream off-ramps.

To illustrate this, we consider that the capacity of the example freeway section has been increased from 4000 to 5000 vph upstream of the off-ramp. We consider also that demands from zone 1 to 3 have increased from 500 to 1500 vph, while all other characteristics of the freeway remain unchanged. The simulation of these conditions without ramp control results in a total travel time of 967.6 vehicle-hr. When these conditions are simulated again with the ramp flow metered at a rate of 600 vph, in the same manner as described in the initial metering model, the total travel time is estimated to be 958.3 vehicle-hr, which represents a reduction in travel time of 0.94 percent from the nonmetering case. This reduction of 0.94 percent can be compared with the 0.39-percent reduction obtained earlier. Clearly, inasmuch as the two models were identical except for the flow from zone 1 to 3, the additional benefits result from the fact that, with metering, the flow utilizing the off-ramp is not impeded. As in this model this flow is three times as large, the benefits are also much larger.

SENSITIVITY ANALYSIS: DIVERSION

The discussion so far has not considered the diversion of vehicles. In reality, if alternative routes are available a nontrivial diversion may occur. In this section we investigate the impact of two alternative diversion strategies, namely, a user optimal and a system optimal diversion.

Effect of User Optimal Diversion

In general, it is considered that individual drivers tend to choose routes that minimize their own travel times (*15*). In accordance with this behavior, it can be expected that drivers faced with extensive delays caused by the metering of a ramp will seek alternative routes that will result in a lower travel time cost.

There are, of course, numerous issues regarding perceived versus real costs, quality of available information, and bias toward certain roadway types. These concerns, though they are sometimes important, are not examined here.

For this model, vehicles received traffic information every 2 sec. Because, in practice, perfect information is rarely available, a 5-percent error was introduced into the information before it was provided to drivers.

The initial ramp metering model presented above was simulated with all drivers traveling from zone 1 to 4 receiving network information and able to divert. The resulting total travel time was 718.1 vehicle-hr, a reduction of 12.17 percent of that for the base premetering case.

To check that drivers routed themselves according to user optimal criteria, the average travel times for the two alternative routes from zone 1 to 4 were computed. Traversal of the ramp route required, on average, 7.1 min, whereas the arterial route required 6.9 min. As both routes have approximately the same average travel time, it can be concluded that the vehicles were diverted in accordance with user optimal behavior.

Effect of System Optimal Diversion

In the previous subsection we examined the effect of user optimal diversion, which is the way in which individuals are considered to behave at present. If, however, drivers were routed such that system optimal routings could be achieved, total system travel time would be further minimized.

Figure 7 illustrates the proportion of vehicles from zone 1 to 4 that use the arterial route and the associated system cost in terms of total travel time. As indicated, the system optimal diversion rate indicates that 100 percent of the vehicles that would normally use the controlled ramp should divert to the arterial. In this case the total travel time is only 701.4 vehicle-hr, representing a 14.21-percent reduction in system travel time compared with that for the base premetering case.

Drivers do not select system optimal routes unless they are forced to do so, so this analysis would be difficult to implement in practice. However, it serves as a convenient estimate of the upper limit on the benefits that one could achieve through the implementation of ramp metering in this example network.

CONCLUSIONS

A number of factors were shown to have a significant impact on the net benefits of ramp metering. Specifically, the effects of several of these factors, including O-D demands, metering rates, initiation time of metering, capacity drop, and diversion strategies, were examined.

These effects were quantified through the application of a simulation model for a small example network. Figure 8 provides a summary of these results. This analysis indicated that benefits of as much as a 26-percent reduction in total travel time may be obtained if metering is carried out efficiently while drivers are routed in a system optimal manner and that a 5-percent reduction in capacity occurs when the freeway becomes congested.

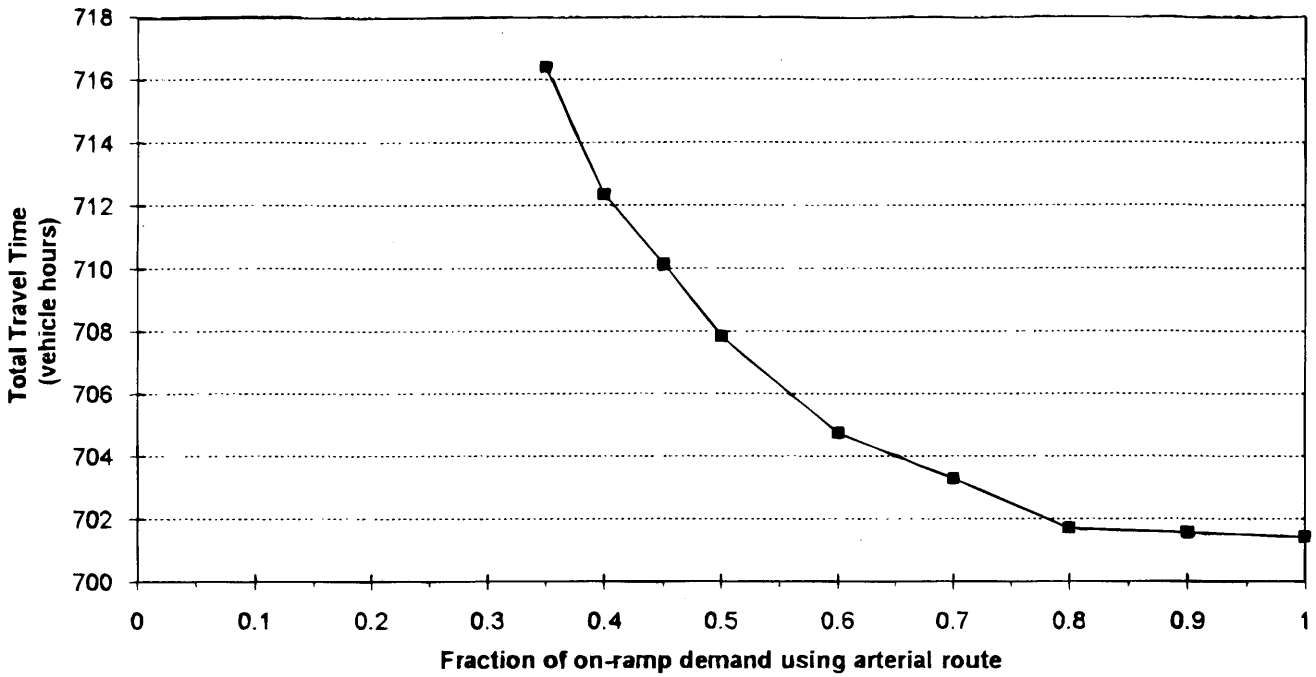


FIGURE 7 Effect of diversion rate on total network travel time.

In the absence of any capacity reduction during congestion, benefits in the range of 12–14 percent can be obtained if drivers are permitted to divert to alternative routes.

In the absence of alternative diversion routes and a reduction in capacity during congestion, ramp metering was shown to be a potentially inefficient means of reducing total travel time.

It must be noted that travel time reductions as the result of the implementation of ramp metering strategies are highly network dependent. The presence and quality of potential diversion routes, the prevailing origin–destination patterns, and the physical locations of alternative routes all affect the premetering traffic conditions and dictate the benefits that might be obtained through the use of ramp metering.

In this paper we have examined the relative benefits of several control parameters through the use of a simple example network. The net benefits computed from this examination may not be applicable to more general networks.

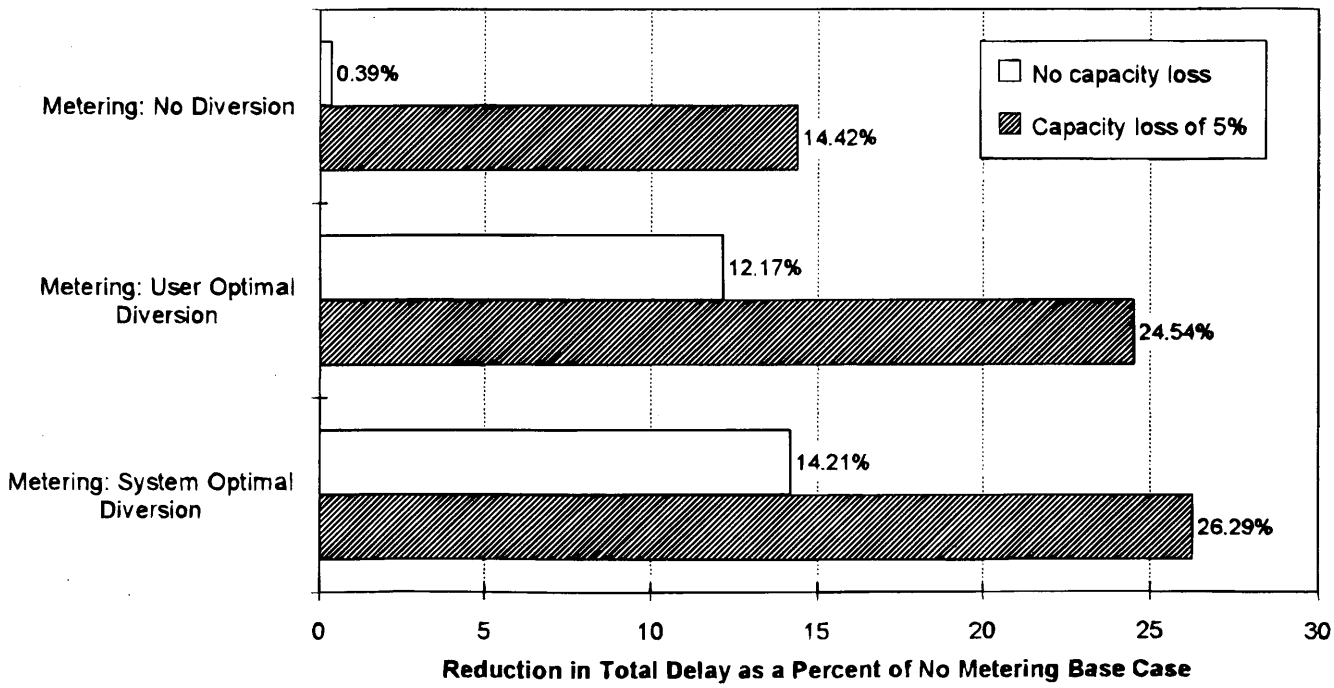


FIGURE 8 Comparison of effects of ramp metering with and without diversion on total network travel time.

Examining the impact of ramp metering by using analytical techniques for even simple networks can be rather difficult. This level of difficulty rises rapidly when traffic conditions and control strategies become more representative of actual field conditions. Furthermore, analytical techniques rely on simplifying assumptions regarding traffic behavior that limit their range of applicability.

The INTEGRATION model was found to be a robust evaluation tool that can be used objectively to quantify expected benefits of different ramp metering models under a variety of routing and controlled conditions, something that is difficult to do by using analytical techniques.

In this paper we have evaluated factors that affect ramp metering strictly in terms of reductions in total travel time. Because fuel consumption, emissions, and safety are also significant attributes of net benefits, effort should be undertaken to incorporate these factors into the evaluation.

Having shown that the INTEGRATION simulation model is able adequately to reflect traffic behavior and network control devices, we can then use it to evaluate the impact of ramp metering for various conditions on a more representative network.

REFERENCES

1. Salem, H. H., J. M. Blosseville, and M. Papageorgiou. On Ramp Control: Local. In *Concise Encyclopedia of Traffic and Transportation Systems* (M. Papageorgiou, ed.) Pergamon, London, 1991, pp. 289–294.
2. Jacobson, L. N., K. C. Henry, and O. Mehyar. Real-Time Metering Algorithm for Centralized Control. In *Transportation Research Record 1232,* TRB, National Research Council, Washington, D.C., 1989, pp. 17–26.
3. May, A. D. *Traffic Flow Fundamentals.* Prentice Hall, Englewood Cliffs, N.J., 1990, pp. 238–245.
4. Papageorgiou, M. On Ramp Control: Coordinated Traffic-Responsive Strategies. In *Concise Encyclopedia of Traffic and Transportation Systems* (M. Papageorgiou, ed.) Pergamon, London, 1991, pp. 289–294.
5. Wattleworth, J. A., and D. S. Berry. Peak period control of a freeway system—Some theoretical investigations. In *Highway Research Record 89,* HRB, National Research Council, Washington, D.C., 1965, pp. 1–25.
6. Nsour, S. A., S. L. Cohen, J. E. Clark, and A. J. Santiago. Investigation of the Impacts of Ramp Metering on Traffic Flow with and without Diversion. In *Transportation Research Record 1365,* TRB, National Research Council, Washington, D.C., 1992, pp. 116–124.
7. Van Aerde, M. A. Single Regime Speed-Flow-Density Relationship for Congested and Uncongested Highways. Presented at 74th Annual Meeting of the Transportation Research Board, Washington, D.C., January 22–28, 1995.
8. Rilett, L., M. Van Aerde, G. MacKinnon, and M. Krage. Simulating the TravTek Route Guidance Logic Using the INTEGRATION Traffic Model. *Proc. VNIS-91 Conference,* Dearborn, Mich., 1991, pp. 775–787.
9. Van Aerde, M., and S. Yagar. Dynamic Integrated Freeway/Traffic Signal Networks: A Routing-Based Modeling Approach. *Transportation Research A,* Vol. 22A, 1988, pp. 445–453.
10. Hellinga, B., and M. Van Aerde. An Overview of a Simulation Study of the Highway 401 Freeway Traffic Management System. *Canadian Journal of Civil Engineering,* Vol. 21, 1994, pp. 439–454.
11. Bacon, V., D. Lovel, A. D. May, and M. Van Aerde. Using the INTEGRATION Model to Study High Occupancy Vehicle Facilities. Presented at 73rd Annual Meeting of the Transportation Research Board Meeting, Washington, D.C., January 9–13, 1993, paper 940472.
12. Hall, F. L., and L. M. Hall. Capacity and Speed-Flow Analysis of the Queen Elizabeth Way in Ontario. In *Transportation Research Record 1287,* TRB, National Research Council, Washington, D.C., 1990, pp. 108–118.
13. Hurdle, V. F., 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.
14. Transportation Research Board. *Highway Capacity Manual—Special Report 209.* National Research Council, Washington, D.C., 1985, pp. 5–16 and 5–23.
15. Wardrop, J. G. Some Theoretical Aspects of Road Traffic Research. In *Proceedings, Institution of Civil Engineering II* (I), 1952.

Publication of this paper sponsored by Committee on Transportation System Management.

Incident Management via Courtesy Patrol: Evaluation of a Pilot Program in Colorado

PEGGY CUCITI AND BRUCE JANSON

A courtesy patrol program was operated by the Colorado Department of Transportation on urban freeways during peak periods to reduce congestion attributable to incidents. In this article are described the program's implementation using two approaches to service delivery, the types of incidents encountered, services provided, and impacts on traffic flows. During the pilot program, the duration of incidents was reduced by 8.6 to 10.5 min. Using a deterministic queuing model, average delays were estimated to be reduced by 71 to 98 vehicle-hr per incident, depending on roadway position, time of day, and assumptions regarding lane blockage effects. The program's benefits far exceeded its costs.

The Colorado Department of Transportation (CDOT) initiated a courtesy patrol program on a pilot basis in the summer of 1992 to provide incident management on major roadways during rush hour periods, with the goal of reducing congestion. This article is drawn from a larger evaluation (*1*) and includes reports on program implementation, incident type, service levels, and program effectiveness.

PROGRAM APPROACH AND IMPLEMENTATION

Congestion and Incident Management

Congestion is an increasingly serious problem. Nationally, congestion on urban freeways is responsible for as much as 2 billion vehicle-hr of delay and $16 billion in costs (*2*). In addition, congestion contributes to poor air quality, wasted fuel, and accidents.

While some amount of congestion stems simply from traffic volumes exceeding roadway capacity, studies have shown that incidents—vehicle breakdowns and accidents on or along the road—account for as much as 60 percent of all congestion. Incidents include major accidents that tie up several lanes for hours; minor accidents and stalled vehicles that block only one lane for short durations; vehicles stopped in shoulders; spilled loads; construction, utility, and maintenance activities; and special events that generate heavy traffic volumes (*3*).

According to a Federal Highway Administration report (*3*), incidents blocking one lane of a three-lane road will reduce capacity by almost half. Even an incident on the shoulder that does not physically block a lane, such as a stalled vehicle or a law enforcement stop, can cause a 25 percent capacity reduction. Capacity reductions occur even when lanes are not blocked, due to the "gawking" effect, which is caused by drivers slowing to observe the incident. The faster an incident can be cleared from the roadway, the less impact it has on traffic flow. The California Department of Transportation estimates that for each minute the time to clear blocked lanes is reduced, a motorist's delay is reduced by 4 to 5 min (*3*).

How quickly vehicles are moved off the roadway depends on a number of factors, including how fast an incident is detected, how quickly help arrives, the motorist's response to an offer of service, the time it takes to provide the service, and the legal framework that governs vehicles disabled along the roadway.

Program History

The idea for the courtesy patrol came from the Colorado Incident Management Coalition (CIMC), a multidisciplinary task force convened by CDOT in 1991. The CIMC recommended implementation of a comprehensive incident management program. Continuous flows of information concerning volume, speed, accident information, and lane closures would be sent to a Traffic Operations Center, which, in turn, could direct response and relay information to motorists. Full implementation of the plan required creation of a new high-technology infrastructure involving electronic and communications equipment.

The courtesy patrol was one part of the system that could stand alone. Hence, the first of CIMC's recommendations to be implemented by CDOT was the Mile High Courtesy Patrol (MHCP).

The Program Model and Implementation

The program model is depicted in Figure 1. Colorado tried two approaches to service delivery. CDOT entered into contracts with the Colorado State Patrol (CSP) to provide service in one zone and the American Automobile Association (AAA) in another. See Table 1.

Cooperative relationships were established between CDOT and numerous other entities. Metro Traffic Control, various media organizations, and sky-based traffic observers were important partners. These organizations play a role in incident detection and in communicating to the broader public information regarding traffic conditions. Links were established with the Denver Police Department, which has responsibility for traffic law enforcement and emergency response within Denver city limits. Also, various private businesses were involved in program planning and operation. For example, businesses allowed specific parking lots to be used as "safe havens" for disabled vehicles moved by the MHCP from the interstate.

P. Cuciti, Graduate School of Public Affairs, University of Colorado at Denver, 1445 Market St., Denver, Colorado 80202. B. Janson, College of Engineering, University of Colorado at Denver, P. O. Box 173364, Denver, Colorado 80217–3364.

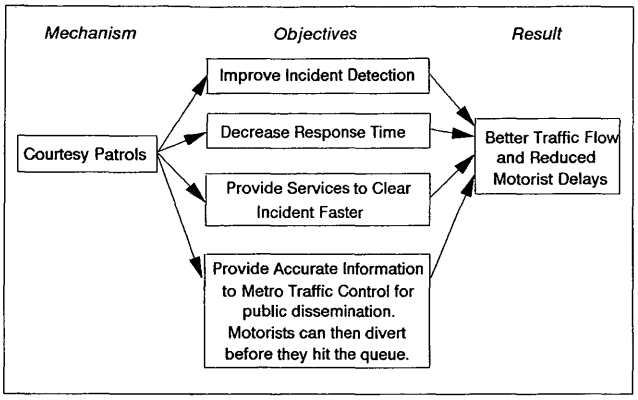


FIGURE 1 Mile High Courtesy Patrol program model.

Time and Place of Operation

Six courtesy patrols operated during rush hours on approximately 43 km along I-25 and a short stretch of I-70 near where it intersects I-25. These corridors were chosen because they have high traffic volumes, flow difficulties attributable to changes in road geometry (e.g., shift in number of lanes) or construction activities, or they lack a shoulder. Three zones were established, each patrolled by two MHCP vehicles.

Vehicles

Two types of vehicles were used by MHCP. AAA used Class A tow trucks and could tow vehicles to safe havens off the freeway. The CSP, with its four-wheel-drive vehicles equipped with heavy push bumpers, could move a more limited range of vehicles, and only for short distances.

Staffing

Each MHCP vehicle was staffed 6 hours a day, split between morning and evening periods. The CSP used 18 off-duty officers who volunteered to work on an overtime basis. They added a 3-hour MHCP shift at the beginning or end of a regular work day or worked one or more shifts on their days off.

AAA staffed the courtesy patrol with 10 regular AAA drivers who volunteered to participate in the pilot project. Their work week was structured, however, so that MHCP substituted for other work. Drivers worked three 12-hour days, splitting their time between MHCP (during rush hours) and regular AAA duties (during the middle of the day).

Patrol operators were required to study volumes detailing MHCP procedures, but other than that they received no special training. State patrol officers had received basic life support training, such as CPR and First Aid, when they first joined the CSP. When they are first hired, AAA's drivers take a 2-day training course that includes defensive driving, drivers' education, and general mechanical training (such as changing a flat tire, jump starting a vehicle, and diagnosing problems on the scene).

INCIDENT OCCURRENCE AND MANAGEMENT

Between August 28, 1992 and February 26, 1993, the courtesy patrol reported 3,393 incidents, an average of 27.6 incidents per day.

TABLE 1 Comparison of Key Features: CSP versus AAA Implementation of the Courtesy Patrol Program

	CSP	AAA
Territory	I-25 between Colfax and 84th Avenue; I-70 between Federal and Washington	I-25 between Colfax and County Line Rd.
Equipment	Four Wheel Drive vehicle equipped with push bumpers and removeable magnetic signs designating courtesy patrol on door and roof.	Standard Tow Truck, with removable magnetic signs designating courtesy patrol on door and roof.
Personnel	Off-Duty,uniformed, state patrol officers	Regular AAA tow truck drivers
Communication	Linked by stationary radio to CSP dispatcher; Linked by portable radio to state base's construction-based communications system and to Metro Traffic Control.	All communication via stationary radio to AAA dispatcher. Dispatcher communicates by phone with Metro Traffic Control.
Number of Patrol Units	2	4
Roadway Center Line Km. Lane Km.	 20 147	 25 157
Incidents (excl.abandoneds) Per Patrol Unit Per Lane Km.	 529 7	 395 10

Cars accounted for 61 percent of incidents; pickup trucks or vans accounted for another 29 percent. Larger vehicles such as trucks, vehicles with trailers, or buses, which could pose greater difficulty for the MHCP in terms of movement, accounted for just under 9 percent of incidents.

In almost three-quarters (72.7 percent) of the incidents reported, the vehicle was not in a lane of traffic. Most vehicles (63 percent) were found on the right shoulder. Six percent of incidents were in the left lane, 4 percent in the middle lane, 10 percent in the right lane, and 8 percent in an acceleration lane or on-ramp.

Abandoned vehicles—a problem that the MHCP could do little about—accounted for 22 percent of all incidents. Courtesy patrol operators reported the following causes for disabled vehicles: miscellaneous mechanical problems (34 percent), flat tire (14 percent), gas outage (11 percent), and accidents (9 percent).

MHCP Activity

Ninety percent of all incidents reported were detected by the courtesy patrol. Nine percent were reported to Metro Traffic Control or the dispatcher who relayed the information to the courtesy patrols. The courtesy patrol took 7 min, on average, to arrive at the scene of an incident reported to them by any outside source.

The courtesy patrols were capable of providing a range of services to stopped motorists. They could fix flat tires, provide a free gallon of gasoline, fill radiators with water, jump-start stalled vehicles, and fix some other minor mechanical problems. If a vehicle had more a serious or difficult-to-identify mechanical problem, the courtesy patrol could move the vehicle or call for other assistance. In addition to providing services to the stopped motorist, the courtesy patrol would protect the scene (particularly if the vehicle was in a lane of traffic). Using its vehicle's emergency lights, the courtesy patrol would alert upcoming motorists to the problem and hence avoid accidents.

The courtesy patrol obtained permission from the motorist before providing any assistance. Service was refused in 14 percent of the cases. The usual reason for rejecting service was that the situation was under control or that help was already on the way.

Table 2 shows the proportion of incidents, classified by problem type, that received different kinds of service. In 36 percent of all incidents (not including abandoned vehicles) the courtesy patrol could provide a direct service related to the presenting problem, such as fixing a flat tire or providing gasoline. In other cases, they may have moved the vehicle to a safer location, protected the scene, or called for assistance.

Vehicle Movement

Vehicles were moved by MHCP, by a push or tow, in approximately one fifth of all cases. Vehicles disabled in traffic lanes were more likely to be moved than those stopped in other positions on the roadway. Table 3 shows that the MHCP provided a tow or push to roughly half the vehicles disabled in traffic lanes.

Vehicles were often moved by private tow operators as well as the courtesy patrol. All told, 66 to 78 percent of vehicles disabled in a traffic lane received a tow or push from someone.

Incident Duration

To minimize congestion, vehicles disabled in traffic lanes (or on the shoulder within 6 ft of traffic) must be moved off the road as quickly as possible. Table 4 indicates how long it took after MHCP arrival for the vehicle to be moved. On average, vehicles disabled in the traffic lane were moved out of that lane 9.9 min after MHCP arrived on the scene.

The courtesy patrol spent longer servicing each incident than is indicated by these movement times. The longer time is required because the move itself may have taken time, particularly if the

TABLE 2 Percent of Incidents Receiving Specified Service from the Mile High Courtesy Patrol

Percent Receiving:	Total Incidents	Presenting Problem						
		Tire	Gas	Radiator	Misc Mech	Debris	Accident	Other
Service Directly Corresponding to Problem	36%	72%	81%	51%	20%	87%	na	na
Tow/Move	21%	6%	4%	13%	36%	0%	23%	7%
Protected Scene	16%	6%	5%	4%	12%	24%	66%	19%
Call for help	15%	5%	3%	12%	19%	13%	36%	8%
Other Service	13%	10%	7%	18%	12%	2%	9%	48%
Service Refused	14%	10%	7%	20%	17%	2%	4%	26%
Count of Incidents	2559	457	356	106	1122	45	280	193

Note: The same case can receive multiple services. This is why percentages add to more than 100%. Also, the count of incidents may differ from table to table due to missing data.

TABLE 3 Movement of Disabled Vehicles Based on Initial Roadway Position

Vehicle Position	Percent Moved by:		
	Courtesy Patrol	Other Tow or Push	By Anyone
Left Lane	51%	41%	78%
Middle Lanes	48%	36%	69%
Right Lane	54%	20%	66%
Accel/Decel Lane	23%	17%	38%
Exit or Entr. Ramp	27%	13%	37%
Left Shoulder	26%	31%	45%
Right Shoulder	17%	7%	25%
Ramp Shoulder	13%	10%	18%
Off Road	10%	7%	17%
All Positions	24%	14%	35%

Note: This table shows a higher percentage of incidents receiving a tow or push from the courtesy patrol than does Table 3.5. There is some internal inconsistency in reporting. When asked on the form about the type of service provided, only 21% showed a tow or move. When asked about vehicle movement and who did it, some additional forms indicated movement by the courtesy patrol.

move was to a safe site off the roadway. In addition, the courtesy patrol may have provided a second service after the initial movement. For example, a car might run out of gas while in a lane of traffic. After moving the vehicle, the courtesy patrol would fill the car with a gallon of gas, enabling it to resume travel.

Courtesy patrol operators reported that, in their judgement, 80 percent of incidents (excluding abandoned vehicles) were cleared when they departed from the scene. An incident was considered cleared if there had been an acceptable disposition of the vehicle involved and no further impact on traffic.

ANALYSIS OF TRAFFIC IMPACTS

A deterministic queuing model was used to estimate the average vehicle delay caused by incidents. Morales (*4*) found this type of queuing model to yield close estimates of accident delays on freeways. Janson and Rathi (*5*) describe the use of this approach for estimating vehicle delays due to accidents. Dynamic modelling was not feasible for this evaluation, but Janson and Robles (*6*) later performed dynamic traffic assignment simulations for the portion of I-25 discussed here. The preliminary results of accident scenarios within their framework do not contradict the magnitudes of delay estimates reported here.

All traffic incidents involve the following phases.

- Detection Phase: time from when the event first occurs to when people capable of responding are notified.
- Response Phase: time from notification to when the response team arrives at the scene.
- Service Phase: time from arrival to when the incident is sufficiently cleared to restore the highway to normal capacity.

TABLE 4 Incident Duration by Position of Disabled Vehicle on the Roadway

Incidents	Response Time	Service Time		Total Incident Duration
		Through First Vehicle Movement	Total	
All	1.1	9.6	11.2	12.0
Traffic Lanes	1.9	9.9	13.9	15.5
Left Shoulder/Ramps	1.2	10.4	12.4	10.8
Right Shoulder	0.8	9.3	10.2	13.4

• Queue Dissipation Phase: time from capacity restoration to when normal traffic flow resumes.

The total delay caused by an incident depends on the duration of each of these phases, the traffic volume on the highway approaching the incident, and the number of blocked and unblocked lanes. An incident causes queuing and vehicle delays because the vehicle arrival rate (hourly vehicle volume) exceeds the vehicle service rate (unblocked lane capacity) during the first three incident phases.

Figure 2 shows a graph of the queuing delays caused by a lane-blocking incident as estimated by a deterministic queuing model. The total travel time delay caused by an incident is equal to the shaded area in Figure 2, as described by Janson and Rathi, (5). The slopes of lines indicated by C_1 and C_2 equal the capacity of a highway during the incident clearing and during the queue dissipation phases, respectively. The incident clearing phase (sum of Phases 1 through 3) is from event time t_0 to time t_2 when all lanes are cleared. the queue dissipation phase (Phase 4) is from time t_2 to time t_3 when the queue disappears. At time t_2, when the incident is cleared from blocking any lanes, the road's capacity returns to its pre-incident level (C_2). Because C_2 exceeds the vehicle arrival rate V_2, the queue begins to dissipate. Morales (*4*) found that a highway may not return to its pre-incident service rate at one time, and that short intermediate steps or piece-wise linear segments between lines C_1 and C_2 can represent certain incident clearing processes in more detail. This additional detail was found to alter the total delay estimate by less than 10 percent in cases in which it was used.

The vehicle service rate of unblocked lanes during the incident clearing phase, denoted as C_1, depends on the number of open lanes, plus other factors such as smoke, debris, visible wreckage, and emergency equipment.

With regard to vehicle arrival rates, the delay calculation allows the arrival rate of vehicles at the rear of the queue to decrease at time t_1 because of route diversions or lessening travel demand. Increasing travel demand could actually cause the arrival rate to increase at t_1.

The queuing model is used to estimate the traffic delays associated with incidents occurring along the southern stretch of I-25 in northbound lanes. The model uses actual times and road positions associated with incidents and actual traffic volume data for the time of day that the incident occurred. The analysis is restricted to this portion of roadway because only there is the technology in place to provide accurate data on traffic volumes.

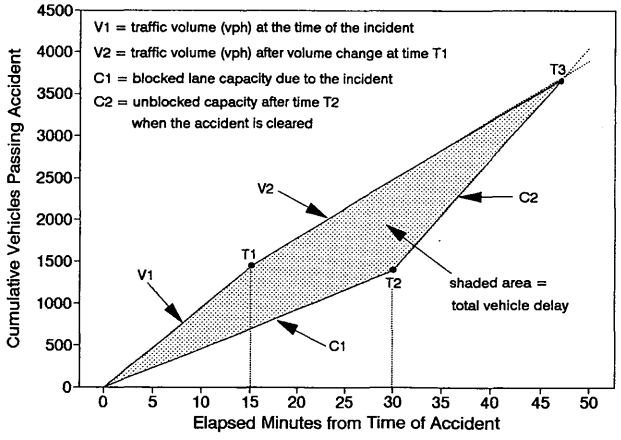


FIGURE 2 Estimation of vehicle delays due to incidents.

The model allows estimation of what traffic delays would have been, assuming different times involved in incident detection, response, and service. Hence, estimated traffic delays during the period when MHCP was operating can be compared with estimates of what occurred prior to MHCP implementation.

Core Inputs to the Impact Analysis

Time Duration

Time duration involves detection, response, service time, and queue dissipation. No direct estimates of detection time are available either before or during the period of MHCP operations. Incidents were probably detected faster with the addition of regular patrols, but since there is no proof, it has been assumed that there was no difference in detection time. Detection time is estimated at 5.5 minutes, representing how long it takes to observe any given point along the roadway, given the patrol route.

Data are available on response and service times during the MHCP pilot. Understanding of incident response prior to MHCP is somewhat limited. It is based on data for I-25 collected by Metro Traffic Control in the three months prior to MHCP implementation. Metro Traffic Control's records indicate when incidents were first observed by the sky observers (or other means) and when they reported them cleared. Both observations depend on the flight pattern of the observers. Estimates of duration are only approximate, but are the best available.

The estimates are also based on only 4.4 reported incidents per day, a fraction of the total number of incidents now known to exist based on MHCP data. Incidents attributable to accidents and involving a lane of traffic comprise a larger share of the MTC reports than of the MHCP evaluation data base.

Estimates of incident duration are compared for the period of MHCP operation and the prior period for two different sets of incidents, those blocking a traffic lane and all others. As Figure 3 shows, incident duration decreased substantially after the courtesy patrol started operations. Incident duration decreased by 10.5 min for incidents blocking a lane of traffic, and by 8.6 min for those not involving a traffic lane.

Traffic Volumes

Traffic volumes are collected by the CDOT Region 6 traffic operations office for both 5-min and 1-hr intervals at 12 counter locations on the ramps and the main traffic lanes. To ensure conservative estimates of capacity reductions, the model assumes a higher-than-standard maximum saturation flow rate of 2400 vehicles/hr for all lanes, based on data that show that flows of this magnitude regularly occur.

On a three-lane road, whenever volumes per hour exceed 5000, delays could be expected to result even from a right shoulder stall. Most of the traffic volumes observed on I-25 during the hours of MHCP operation exceeded this amount.

Lane Blockages

An important factor in estimating vehicle delays is the fraction of highway capacity lost to lane blockage and driver slowdown. The number of lanes assumed to be lost for incidents occurring in different locations on the roadway are as follows: left shoulder, 0.7;

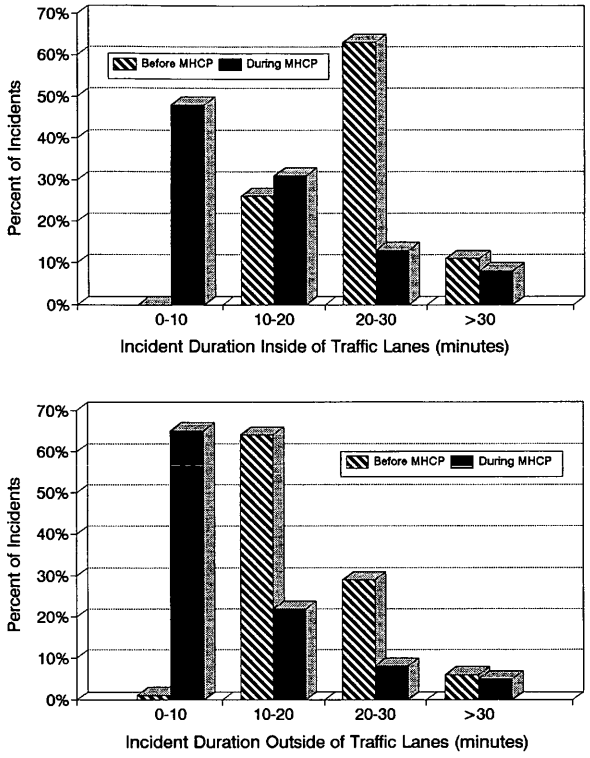


FIGURE 3 Estimated duration of incidents inside and outside traffic lanes, before and during MHCP operations.

left lane, 1.7; middle lane, 2.3; right lane, 1.7; right shoulder, 0.7; off-road 0.3; acceleration-deceleration lane or ramp shoulder, 0.0. These assumptions are rather conservative and should produce low estimates of actual vehicle delays.

Traffic Impacts: Discussion of Results

Figure 4 shows estimated average vehicle delays of all incidents (stalls and crashes) served by the MHCP in the a.m. peak period during the evaluation period. The estimated difference in average delays experienced during MHCP operations relative to the prior period is 98 vehicle-hr per a.m. incident. In the afternoon rush hour, the reduction in delay was somewhat lower, at 75 vehicle-hr per incident on average. Although traffic flows were higher during the afternoon rush hour, the mixture of accident times and locations during the a.m. peak period made its estimated average delay and before-and-during difference greater than that of the a.m. period.

To examine the sensitivity of these average delay differences to capacity reduction assumptions, an alternative estimate was performed assuming that each crash or stall on the right shoulder only reduces highway capacity by 0.1 of a lane, versus 0.7 of a lane. Average savings of 78 vehicle-hr of delay were found for a.m. incidents and 71 vehicle-hr of delay for p.m. incidents.

COST-BENEFIT ANALYSIS

Even though the courtesy patrol offers substantial benefits in terms of reduced traffic congestion, it cannot be concluded that the program is a success until costs are assessed and compared with the benefits. See Table 5.

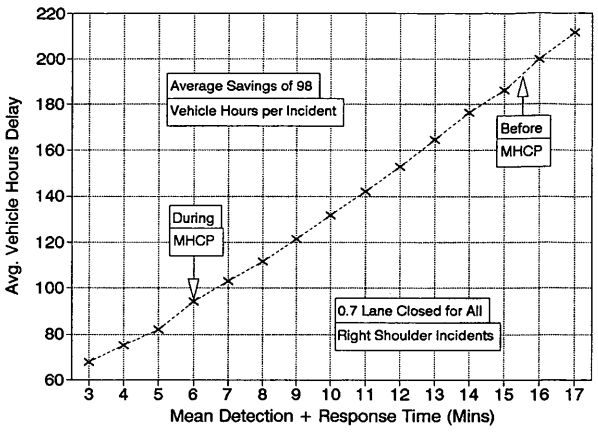


FIGURE 4 Average vehicle hours of delay per incident occurring in the morning peak period.

Assuming that the time saved is valued at $10 per vehicle hour, the courtesy patrol resulted in between $1.8 and $2 million worth of time savings over its 6 months of operation. In addition, motorists received direct services of substantial value including tire changes, minor mechanical repairs, and so forth.

The courtesy patrol program cost approximately $120,000 to operate over the same period. This figure, however, understates the true costs incurred. A comprehensive analysis showed the true cost per patrol unit per hour of operation to be $38 for CSP and $28 for AAA. CSP had lower equipment costs but higher labor costs than AAA. Using these more accurate hourly costs, the true cost of the program was estimated to be $168,000.

The contract cost during the pilot period has been used as the low-end estimate and CSP's true cost during the period (hypothetically applied to all six patrol units) as the high-end estimate of cost. Either way, the ratio of benefit to cost is very high, in the range of 10.5 or 16.9 to one.

CONCLUSION

Operating a courtesy patrol appears to be a cost-effective way of addressing congestion arising from incidents on crowded urban freeways. As a result of the evaluation, CDOT expanded patrol operations to additional corridors in the Denver metropolitan area during morning and evening rush hours.

ACKNOWLEDGMENTS

CDOT funded this research. Andrea Brett and Patut Darjadi provided research assistance. Neal Lacey and Joni Brooks of CDOT were helpful throughout the research and provided comments that improved the full report.

TABLE 5 Detail of Benefit-Cost Analysis

	AM	PM
Number of Incidents - 6 months	1095	1273
Estimated Hours of Traffic Delay Averted Per Incident		
High	98	75
Low	78	71
Estimated Dollars Savings from Reduced Traffic Delay		
High	$1,073,100	$954,750
Low	$854,100	$903,830
Hourly Cost of Operation Per Patrol Unit		
Equipment	$7	$17
Personnel	$31	$12
Estimated Costs (6 patrols)	$120,000 - $168,000	
Benefit Cost Ratio		
High	16.9	
Low	10.5	

REFERENCES

1. Cuciti, P. and B. Janson. *Courtesy Patrol Pilot Program.* Colorado Department of Transportation; Center for Public-Private Sector Cooperation, University of Colorado at Denver, August 1993.
2. Lindley, J. A. Quantification of Urban Freeway Congestion and Analysis of Remedial Measures. FHWA Staff Report RD-87/052, October 1986.
3. *Freeway Incident Management Handbook.* FHWA-SA-91-056. U.S. Department of Transportation, July 1991.
4. Morales, J. M. Analytical Procedures for Estimating Freeway Traffic Congestion. *Public Roads,* Vol. 50, Issue 2, Sept. 1986, pp. 55–61.
5. Janson, B. N. and A. Rathi. Economic Feasibility of Exclusive Vehicle Facilities. In *Transportation Research Record 1305,* TRB, National Research Council, Washington D.C., 1991, pp. 201–214.
6. Janson, B. N. and J. Robles. Dynamic Traffic Assignment with Arrival Time Costs. *Proceedings of the Twelfth International Symposium on Transportation and Traffic Theory, Berkeley, Calif., July 21–23, 1993* (C. Daganzo, ed.) Elsevier Press, Amsterdam, the Netherlands, pp. 127–146.

Publication of this paper sponsored by Committee on Transportation System Management.

Artificial Neural Networks for Freeway Incident Detection

YORGOS J. STEPHANEDES AND XIAO LIU

A freeway incident detection algorithm is developed using back propagation neural networks. Based on real-time occupancy and volume counts from pairs of adjacent loop detector stations, the network is trained with actual data, including 31 incidents from a typical freeway in the Twin Cities Metropolitan Area over the afternoon peak period during 72 days. Results indicate that the neural network, with about 1,000 connections, can learn the main characteristics of a variety of incidents. Algorithm performance, in terms of detection and false alarm rates, is superior to most of the best algorithms that have been tested with this data set.

Fast and accurate detection of incidents is vital for the successful operation of incident management systems. With incidents accounting for many of the vehicle hours lost to nonrecurring freeway congestion, prompt and reliable detection (critical for assuring effective response and clearance) can substantially contribute to improving freeway traffic flow.

Low reliability is the major shortcoming of existing automated incident detection methods for freeway operations. Because of the high number of false alarms generated by such methods, traffic engineers generally do not rely on them for automated detection of incidents.

Because incident management is critical in reducing the total delay to drivers in urban freeways, traffic planners continue to develop methods that can be used to reliably identify an incident. Interest in such methods has increased as transportation officials realize that prompt and reliable detection of incidents is critical to advanced traffic management systems (*1–2*), which seek to provide optimal control of freeway and arterial networks.

Recent research has focused on assessing how existing and new incident detection systems perform (*3*). This assessment involved developing and testing a new algorithm and comparing it with existing ones. This study represents an effort to improve incident detection systems by designing an incident detection algorithm based on artificial neural networks.

BACKGROUND

Artificial neural networks, whose structures are based on the present understanding of biological nervous systems, have been studied for many years in the hope of achieving human-like performance in various practical applications (*4*). These models comprise a large number of simple, nonlinear computational elements operating within a parallel distributed information processing architecture and arranged in patterns reminiscent of a biological neural network. Computational elements or nodes are connected to other elements via weights that are typically adapted during use to improve performance. Through each connection, each element may receive information from another, weighted by the corresponding weight of that connection. Research has demonstrated that artificial neural networks offer high computation rate, memory, learning, and fault tolerance.

Neural network classifiers are nonparametric and make weaker assumptions concerning the shapes of underlying distributions than traditional statistical classifiers. They may thus prove more robust when distributions are generated by a nonlinear process and are strongly non-Gaussian. In particular, they can identify (*a*) which class best represents an input pattern, and (*b*) where it is assumed that inputs have been corrupted by noise or some other process. Several neural network models can be used as classifiers, including the Hopfield net (*5*), the Carpenter-Grossberg classifier (*6*), Kohonen's self-organizing model (*7*), and Multi-layer network (*8*). Multi-layer networks are feedforward nets with one or more hidden layers of nodes between the input and output nodes. Since the backpropagation training algorithm was proposed, Multi-layer network has become the most popular model, particularly in the pattern recognition field. The feasibility of neural network models for freeway incident detection has been demonstrated (*9*).

Incident detection is a typical pattern recognition problem and can benefit from the application of neural network methods. In particular, in incident detection human-like performance is sought in detecting unusual events in the traffic stream and in reducing false alarms by differentiating incidents from other events, such as compression waves, traffic pulses, and equipment malfunction. Methods should be robust under the assumption that the traffic distributions are generated by nonlinear, non-Gaussian processes. Although new incident detection algorithms are promising, methods that can be adapted during use are expected to improve performance. In addition, high computation rate and fault tolerance are needed, and these are characteristics of neural networks. A neural network algorithm is developed that can be trained to recognize traffic patterns through time, and classify such patterns as having an incident or being incident-free. The sensitivity of classification performance on training methods is also investigated. Finally, the performance of the neural network is compared with recent findings from well-performing methods, with encouraging results.

Brief Description of Back Propagation Algorithm

The back propagation neural network is a feedforward, multilayer perceptron with one or more layers of nodes hidden between the input and output nodes, which can be trained using the back propagation algorithm. Although it cannot be proven that this algorithm

Y. J. Stephanedes, Department of Civil and Mineral Engineering, and X. Liu, Department of Civil and Mineral Engineering and Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, Minn. 55455.

converges, its applications have been shown to be successful in a variety of problems. A three-layer network with one layer hidden is shown in Figure 1, where

$$Y_m = f(_{k=0}\textstyle\sum_{k-1} W_{km} Z_k - \Theta_m), \qquad m = 0, 1, \ldots, M-1 \tag{1}$$

$$Z_k = f(_{i=0}\textstyle\sum_{N-1} W_{ik} X_i - \Theta_k), \qquad k = 0, 1, \ldots, K-1 \tag{2}$$

and W_{km}, W_{ik} are the connections between nodes k and m, and i and k, respectively, usually called weights.

The back propagation training algorithm is an iterative gradient algorithm designed to minimize the mean square error between the actual output and the desired output of a multilayer feedforward perceptron. Each node in the hidden and output layers adds weighted inputs from the previous layer and passes the result through a nonlinear function that must be continuously differentiable; the capabilities of this network stem from the use of this function. The following logistic nonlinearity, familiar to transportation engineers from trip demand analysis and other applications, is most commonly used.

$$f(\alpha - \Theta) = 1/(1 + e^{(\alpha-\Theta)}) \tag{3}$$

where Θ is an internal threshold or offset, and the value of f varies from 0 to 1.

The training algorithm is initialized by setting all weights at small random values from -1 to 1 and specifying an input vector X and a desired output vector D. The actual output vector Y is calculated from Equations 1, 2, and 3. Based on the error between desired and actual output, the weights are updated using a recursive algorithm that begins at the output nodes and works back to the hidden and input layers.

DATA DESCRIPTION

The neural network detection method was developed with data collected from Interstate 35W, a heavily traveled and often congested freeway in Minneapolis, Minnesota. The study was confined to the afternoon peak period (4:00 p.m. to 6:00 p.m. because incident detection under moderate-to-heavy traffic conditions is of greatest importance for advanced freeway management.

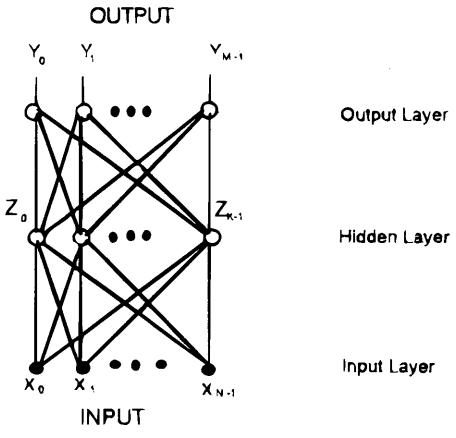


FIGURE 1 Three-layer feedforward neural network.

The selected 5.5 mi freeway segment shown in Figure 2 is fully covered by TV cameras, allowing detailed traffic information to be gathered. This segment includes most types of geometric configurations usually found on a freeway, such as entrance and exit ramps with or without exclusive lanes, bottlenecks; ramps carrying heavy volumes, etc. The data, which were obtained from 14 detector stations imbedded 0.3 to 0.7 mi apart in this segment, consist of 1-minute volume and occupancy updated every 30 sec and averaged over all lanes (see Table 1). A total of 140 hr of traffic data from 72 afternoon peak periods were used.

In the time period of this study, 31 incidents were reported by the Traffic Management Center of the Minnesota Department of Transportation. Confirmation of these incidents is made mainly through television cameras and recorded daily in incident logs by the TMC engineer. Incident logs include time and location of incident occurrence, incident type, duration, severity, impact on traffic, roadway condition, and other information.

Incidents in the data set, to be detected by the algorithm, include incidents blocking one lane, one or both shoulders, or a combination of lanes, shoulders, and freeway entrance-exit. There were no incidents blocking two or more lanes. Although the proposed algorithm is based on observable changes in the traffic flow all, incidents recorded by the traffic engineer were included, even if they had minimal or no impact on traffic. Detection of these incidents by the neural network algorithm proved to be most challenging.

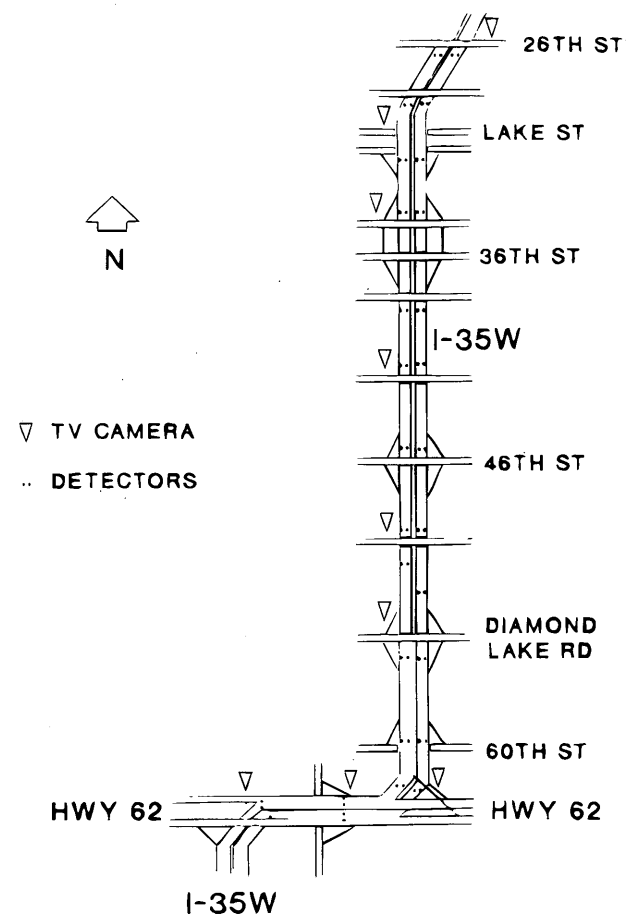


FIGURE 2 Study site in Minneapolis, I–35W.

TABLE 1 Original Volume and Occupancy Traffic Data

Date: 12/06/1989 southbound, afternoon

Station Time	042S	046S	050S	051S	055S	060S	061S	062S	063S
16:10:30	22†37‡	29 24	32 24	29 17	29 24	33 23	32 41	35 10	19 8
16:11:00	24 30	31 28	33 24	35 19	26 24	33 28	31 35	32 9	20 9
16:11:30	29 27	28 31	32 23	33 20	30 22	25 26	32 32	27 7	20 9
16:12:00	29 24	28 29	35 23	34 21	33 19	25 23	31 35	29 8	20 8
16:12:30	27 21	32 26	35 23	36 22	32 20	31 25	28 36	28 8	17 7
16:13:00	27 25	31 23	35 23	33 25	32 22	31 29	27 37	27 8	16 7
16:13:30	24 28	32 22	37 25	31 30	31 26	26 32	30 35	30 9	14 6
16:14:00	22 26	34 23	34 27	33 29	28 29	26 30	30 31	33 10	16 7
16:14:30	25 25	32 20	31 35	34 25	29 26	27 23	31 34	31 9	17 8
16:15:00	27 23	30 18	31 36	33 24	30 26	29 23	32 34	31 10	19 9
16:15:30	29 22	31 19	31 29	34 25	23 36	33 25	30 33	30 9	21 9
16:16:00	28 21	32 21	31 25	35 25	22 33	33 24	31 39	29 8	17 7
16:16:30	26 22	32 24	32 22	32 24	27 22	31 23	30 40	37 12	17 8
16:17:00	28 25	30 28	32 20	29 26	29 22	27 25	29 35	40 13	20 9
16:17:30	27 24	29 28	29 26	24 35	32 22	28 25	30 33	33 9	20 9
16:18:00	28 22	32 25	22 40	17 32	33 23	30 22	29 32	28 8	16 7
16:18:30	29 22	32 22	21 46	20 22	30 22	31 24	33 33	28 8	19 9
16:19:00	27 23	31 22	24 42	26 17	28 20	33 27	36 32	27 8	21 10
16:19:30	23 33	31 25	25 39	26 13	29 19	31 28	34 32	28 8	20 10
16:20:00	21 36	26 34	25 41	26 12	29 17	27 25	32 29	29 9	19 10
16:20:30	24 31	20 48	25 39	27 13	29 18	29 24	31 31	30 9	18 9
16:21:00	27 34	20 52	28 31	28 14	29 18	32 22	31 33	32 10	21 10
16:21:30	27 28	22 50	28 32	28 14	28 18	30 18	32 32	29 8	21 10
16:22:00	24 22	22 47	26 35	28 14	29 19	26 15	30 30	28 8	24 13
16:22:30	19 31	24 43	28 33	28 15	32 18	28 17	27 24	36 11	24 12
16:23:00	17 42	26 39	29 32	28 15	30 16	31 20	33 27	36 11	22 10
16:23:30	21 42	26 40	29 32	28 13	27 15	30 22	37 32	29 8	23 10
16:24:00	23 41	23 43	28 35	28 15	29 15	31 25	35 32	27 7	26 11
16:24:30	22 38	24 39	27 37	28 14	28 12	29 22	31 31	28 7	25 10
16:25:00	22 31	26 35	28 35	28 13	27 11	27 21	31 30	30 7	24 9
16:25:30	22 31	27 32	26 29	29 14	26 11	33 26	33 31	30 7	19 8
16:26:00	22 32	27 29	28 31	28 14	28 12	32 24	31 31	28 7	16 8
16:26:30	22 33	26 29	28 36	28 14	26 11	27 20	30 28	30 7	24 11
16:27:00	22 34	28 32	27 33	28 15	25 10	28 21	30 26	34 10	23 9
16.27:30	23 33	27 29	29 30	26 14	28 12	29 22	30 27	28 8	22 10

† Column 1 indicates volume (veh/min);
‡ Column 2 indicates occupancy (%).

INCIDENT DETECTION ALGORITHM

Reducing the number of false alarms that can result from short-term traffic inhomogeneities is a major objective of the incident detection algorithm developed in this study. To increase the transferability potential of the algorithm, it is kept as simple as possible. Following results from earlier work, the algorithm operates based on real-time traffic measurement at pairs of adjacent stations.

Earlier work has sought to achieve detection performance by exploiting the smoothed, normalized spatial occupancy difference between adjacent stations through time (*3*). To simplify the detection process, only raw data are employed in this work. Further, to take full advantage of all possible patterns presented by such data in real time, both occupancy and volume are presented to the neural network. These data are routinely available at traffic management centers across the United States and in other countries. Figures 3 and 4 show occupancy and volume data from a typical incident in our data set, occurring on December 6, 1989, at 16:18:00.

Several neural networks were investigated, all with 41 elements in the input layer and one in the output layer. The network that performed best had 30 nodes in the hidden layer and was selected for the remainder of this work. The number of training iterations was set at two levels; the first level, at which error was adequately reduced, was 1,500, and the second, at which the effects of over-learning are clear, was 4,000. The performance of the neural network for the data set was evaluated with respect to detection rate, false alarm rate, and time-to-detect. The best neural network algorithm was compared with incident detection algorithms previously found to perform best (i.e., DELOS, California, and Algorithm 7) (*10*).

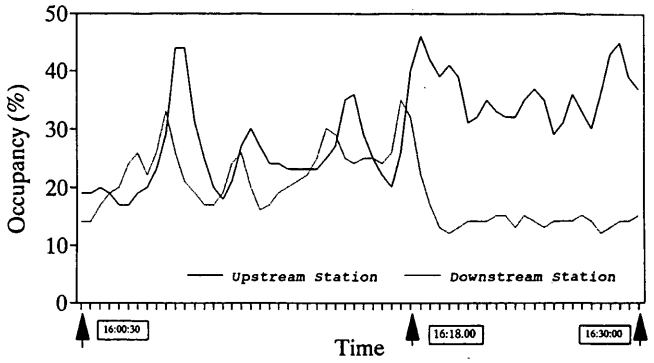


FIGURE 3 Detection occupancy data.

Training

The training of the neural network involves training with freeway data that include 31 reported incidents and with incident-free data randomly acquired from 14 of the 72 days in the data set. These are 30-sec station lane-average volume and occupancy data from 14 stations along the 5.5 mi freeway section described previously. For each incident, one or more 5-min patterns is introduced to the network depending on the duration of the incident, for a total of 89 training incident patterns.

Because the size of the dataset is limited, each consecutive incident pattern is placed at a 2.5-min overlap with its preceding pattern so the number of patterns with which the network is trained increases. The selection of 5-min pattern length and 2.5-min pattern overlap reflects findings from preliminary analysis of the data and is a function of station location, incident duration, and size of data set. Sensitivity analysis could be performed to determine the best pattern length and overlap in terms of algorithm performance. Training could be extended with additional pattern combinations from the data set. Because volume and occupancy data are collected every 30 sec, each sample includes 10 volume and 10 occupancy measurements from each of the two detector stations. As a result, each sample contains a total of 5 min $\times$ 2 measurements/min $\times$ 2 variables $\times$ 2 stations = 40 measurements.

The input vector of the neural network has 41 elements, including one element of constant 1 for adjusting the internal threshold, Θ_k, in Equation 2. The output of the neural network is one element and can vary between 0 and 1. During classification, an output value greater than a user-specified threshold (see later discussion on the

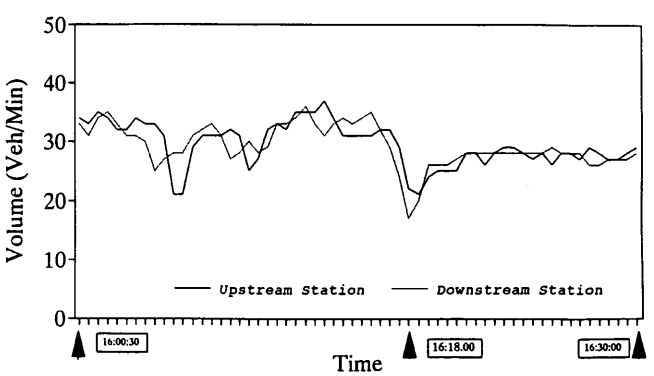


FIGURE 4 Detection volume data.

sensitivity of algorithm performance on the value of this threshold; default value is 0.5) indicates an incident, otherwise an incident-free traffic state is indicated. The desired output of an incident pattern is 1 and that of an incident-free pattern is 0. To illustrate, for an accident occurring between Stations 50S and 51S on December 6, 1989, at 16:18:00 (see original traffic data in Table 1), the first three incident patterns selected are listed in Table 2.

For training the network with incident-free patterns, 14 days in the data set were randomly selected, 336 5-min incident-free patterns were randomly acquired such that they average one per peak hour from each adjacent-station pair, for a total of 14 days × 2 hr/day × 12 station pairs. Because of the large number of available patterns, no pattern overlap was employed. In acquiring these patterns more weight was placed in areas in which previous work (*3*) had indicated a higher number of false alarms. For instance, the first three incident-free patterns between Stations 31S and 35S on December 6, 1989, beginning at 16:00:30, are listed in Table 3.

In each iteration of the training process, all 425 training patterns were presented to the network in random order. Because of the large number of patterns, a small gain was used for updating the network weights; as a result, a high number (above 1,000) of iterations were employed.

Testing

The neural network was tested through application to the complete data set over the 72-day period. Although the data set is the same as the one used for training, the testing procedure involved a substantially larger number of patterns, with both patterns containing incidents and incident-free patterns. For every two adjacent stations, a new pattern was defined in the data every 30 sec for a total of 211,536 test patterns. For example, the first three input patterns of station pair 50S and 51S on December 6, 1989, are shown in Table 4.

Every 30 sec the neural network classifies the state of traffic as either incident or incident-free based on the threshold defined by the user. After a persistence test, an incident alarm is declared. Four types of detection were tested based on persistence values of $P = 0$, 1, 2, and 3. For instance, if $P = 0$ following an incident classification at t, an incident alarm is declared. If $P = 3$ following an incident classification at time t, an incident alarm is declared only if an incident classification is also recorded at $t + 30$ sec, $t + 60$ sec, and $t + 90$ sec.

TABLE 2 Incident Patterns

	50S	51S			50S	51S			50S	51S		
	V	O	V	O		V	O	V	O		V	O	V	O
16 15:30	31	29	34	25	16:18:00	22	40	17	32	16:20:30	25	39	27	13
16 16 00	31	25	35	25	16:18:30	21	46	20	22	16:21:00	28	31	28	14
16:16:30	32	22	32	24	16 19:00	24	42	26	17	16:21:30	28	32	28	14
16:17 00	32	20	29	26	16:19.30	25	39	26	13	16:22:00	26	35	28	14
16:17.30	29	26	24	35	16 20:00	25	41	26	12	16:22:30	28	33	28	15
16 18 00	22	40	17	32	16 20 30	25	39	27	13	16:23:00	29	32	28	15
16.18:30	21	46	20	22	16 21:00	28	31	28	14	16:23:30	29	32	28	13
16:19:00	24	42	26	17	16.21:30	28	32	28	14	16:24:00	28	35	28	15
16.19:30	25	39	26	13	16:22.00	26	35	28	14	16:24:30	27	37	28	14
16.20:00	25	41	26	12	16:22:30	28	33	28	15	16·25:00	28	35	28	13

TABLE 3 Incident-Free Patterns

	31S	35S			31S	35S			31S	35S		
	V	O	V	O		V	O	V	O		V	O	V	O
16.00·30	23	10	24	09	16 05 30	27	12	24	09	16:10:30	28	12	25	09
16:01 00	25	12	21	08	16:06·00	29	12	25	09	16:11·00	28	12	24	09
16:01 30	24	11	24	09	16:06.30	27	11	28	10	16·11:30	28	12	26	10
16:02:00	24	11	25	08	16·07 00	28	12	26	10	16:12·00	28	12	26	10
16:02:30	28	12	21	08	16 07 30	33	15	24	09	16:12 30	29	13	24	09
16.03:00	29	12	25	09	16 08 00	29	12	30	11	16:13.00	29	13	27	10
16:03:30	30	13	27	10	16 08·30	26	11	28	10	16:13 30	32	14	27	11
16:04·00	31	13	26	10	16 09:00	26	11	24	08	16:14:00	32	14	28	11
16:04 30	30	12	28	10	16·09·30	24	10	24	09	16.14:30	30	13	28	11
16:05 00	27	11	28	10	16·10:00	27	11	23	08	16.15.00	31	13	27	10.

The output of the test module is illustrated in Table 5, which reflects part of the testing on the December 6, 1989, data. The output indicates a correctly detected incident and a number of false alarms. The incident began between Stations 51S and 55S at 16:18:00 and was continuously detected from 16:19:30 (time-to-detect = 1.5 min) until 16:54:30. Further, four false alarms at zero persistence were indicated at 16:07:30 (Stations 46S-50S), 16:10:30 (Stations 50S-51S), 16:16:30 (Stations 61S-62S), and 16:18:00 (Stations 55S-60S). For every continuous false alarm series, one false alarm is recorded.

RESULTS

Results from testing indicate the neural network's sensitivity to the number of iterations, the user-specified threshold, and the persistence value. Higher threshold and persistence values reduce the false alarm rate, but also reduce the detection rate and increase the average time-to-detection. For instance, Table 6 indicates that at 1,500 iterations with zero persistence, increasing the user-specified threshold from 0.5 to 0.8 reduces the false alarm rate from 1.4 to 0.40 percent, but also reduces detection rate from 94 to 81 percent and increases average time-to-detection from 2.5 to 4.1 min. Similarly, if user-specified threshold value is kept constant at 0.5, increasing persistence from 0 to 3 reduces false alarm rate from 1.4

TABLE 4 Testing Input Patterns

	50S	51S			50S	51S			50S	51S		
	V	O	V	O		V	O	V	O		V	O	V	O
16:00.30	34	19	33	14	16 01.00	33	19	31	14	16:01 30	35	20	34	17
16.01 00	33	19	31	14	16 01:30	35	20	34	17	16:02·00	34	19	35	19
16:01.30	35	20	34	17	16 02:00	34	19	35	19	16:02·30	32	17	33	20
16:02:00	34	19	35	19	16.02:30	32	17	33	20	16:03.00	32	17	31	24
16 02 30	32	17	33	20	16 03.00	32	17	31	24	16:03 30	34	19	31	26
16 03·00	32	17	31	24	16 03:30	34	19	31	26	16.04.00	33	20	,30	22
16:03 30	34	19	31	26	16.04:00	33	20	30	22	16:04 30	33	23	25	26
16 04 00	33	20	30	22	16·04:30	33	23	25	26	16·05:00	31	29	27	33
16·04 30	33	23	25	26	16 05:00	31	29	27	33	16:05:30	21	44	28	26
16.05 00	31	29	27	33	16 05:30	21	44	28	26	16:06 00	21	44	28	21

TABLE 5 Output of Neural Network

Date: 12/06/1989 southbound, afternoon

Station / Time	042S	046S	050S	051S	055S	060S	061S	062S	063S
16:05:30	-	-	-	-	-	-	-	-	-
16:06:00	-	-	-	-	-	-	-	-	-
16:06:30	-	-	-	-	-	-	-	-	-
16:07:00	-	-	-	-	-	-	-	-	-
16:07:30		Inc	-	-				-	-
16:08:00	-	Inc	-	-	-	-	-	-	-
16:08:30	-	-	-	-	-	-	-	-	-
16:09:00	-	-	-	-	-	-	-	-	-
16:09:30	-	-	-	-	-	-	-	-	-
16:10:00	-	-	-	-	-	-	-	-	-
16:10:30	-	-	Inc	-	-	-	-	-	-
16:11:00	-	-	Inc	-	-	-	-	-	-
16:11:30		-	-	-	-	-	-	-	-
16:12:00	-	-	-	-	-	-	-	-	-
16:12:30	-	-	-	-	-	-	-	-	-
16:13:00	-	-	-	-	-	-	-	-	-
16:13:30	-	-	-	-	-	-	-	-	-
16:14:00	-	-	-	-	-	-	-	-	-
16:14:30		-	-	-	-	-	-	-	-
16:15:00	-	-	-	-	-	-	-	-	-
16:15:30	-	-	-	-	-	-	-	-	-
16:16:00	-	-	-	-	-	-	-	-	-
16:16:30	-	-	-	-	-	-	Inc	-	-
16:17:00	-	-	-	-	-	-	Inc	-	-
16:17:30	-	-	-	-	-	-	Inc	-	-
16:18:00	-	-	-	-	Inc	-	Inc	-	-
16:18:30	-	-	-	-	-	-	Inc	-	-
16:19:00	-	-	-		-	-	Inc	-	-
16:19:30	-	-	-	Inc	-	-	-	-	-
16:20:00	-	-	-	Inc	-	-	-	-	-
16:20:30	-	-	-	-	-	-	-	-	-
16:21:00	-	-	-	-	-	-	-	-	-
16:21:30	-	-	-	-	-	-	-	-	-
16:22:00	-	-	-	Inc*	-	-	-	-	-
16:22:30	-	-	-	Inc	-	-	-	-	-
...	...	...	...	...	...	...	...	...	...
16.54:30	-	-	-	Inc	-	-	-	-	-

* Neural network detects incident.

to 0.40 percent, but also reduces detection rate from 94 to 84 percent and increases time-to-detect from 2.5 to 4.7 min. These effects are illustrated in Figure 5, in which the performance envelope of the neural network is also demonstrated.

The performance of the neural network algorithm was evaluated at different numbers of iterations. Two of these, 1,500 and 4,000 iterations, are reported in this study. A comparison of Tables 6 and 7 or Figures 5 and 6 indicates that the performance of the network at 4,000 iterations is worse than that at 1,500 iterations. Although additional sensitivity analysis may further specify the range of iterations for best performance of the neural network, the results indicate that at least 1,500 iterations are required for the network to be adequately trained and that 4,000 iterations will result in overtraining, which reduces the network's performance.

The performance of the neural network algorithm was compared with the best of the existing algorithms that have been calibrated and extensively tested and evaluated for this data set (*10*). These include Minnesota Algorithm DELOS 3.3 (0.05,6). Minnesota Algorithm DELOS 1.1 (10,6), Algorithm 7, and the California algorithm, in order of decreasing performance.

The California algorithm consists of three comparison tests to preset thresholds. An incident is detected (*a*) when upstream occupancy is significantly higher than downstream occupancy both in absolute value and relative to upstream occupancy and (*b*) when downstream occupancy has adequately decreased during the past 2 min. The last test distinguishes an incident from a bottleneck by indicating that a reduction in downstream occupancy has occurred over a short period of time as a result of the incident.

TABLE 6 Performance Results at 1,500 iterations

Persistence	Threshold	Detection Rate (%)	False Alarm Rate (%)	Average Detection Time (Min.)
0	0.5	94	1.4	2.5
	0.6	90	1.1	2.9
	0.7	81	0.70	3.4
	0.8	81	0.40	4.1
1	0.5	94	0.92	3.2
	0.6	87	0.68	3.9
	0.7	81	0.36	4.4
	0.8	74	0.19	4.8
2	0.4	87	0.75	3.3
	0.5	87	0.58	4.1
	0.6	84	0.40	4.7
	0.7	74	0.19	5.2
	0.8	58	0.10	5.6
3	0.4	87	0.53	3.9
	0.5	84	0.40	4.7
	0.6	81	0.26	5.4
	0.7	71	0.12	5.7
	0.8	48	0.069	6.2

Algorithm 7 is similar to the California algorithm, but it replaces the temporal downstream occupancy difference in the third test with the present downstream occupancy measurement. This replacement seeks to reduce the false alarms produced by compression waves.

The Minnesota algorithms use low-pass filtering of the occupancy measurements to distinguish short-term traffic inhomogeneities from incidents. Further, the algorithms attempt to distinguish recurrent congestion from incident congestion based on slow

TABLE 7 Performance Results at 4,000 Iterations

Persistence	Threshold	Detection Rate (%)	False Alarm Rate (%)	Average Detection Time (Min.)
0	0.5	94	2.9	1.9
	0.6	94	2.5	2.1
	0.7	94	2.2	2.3
	0.8	90	1.8	2.2
1	0.5	94	1.4	3.1
	0.6	94	1.2	3.2
	0.7	94	0.99	3.3
	0.8	84	0.78	3.2
2	0.4	90	0.86	3.7
	0.5	90	0.73	3.7
	0.6	90	0.62	4.2
	0.7	81	0.49	4.6
	0.8	65	0.38	4.5
	0.9	39	0.25	4.4
3	0.4	87	0.55	4.5
	0.5	81	0.47	4.0
	0.6	74	0.37	4.4
	0.7	61	0.29	4.9
	0.8	52	0.21	4.8
	0.9	26	0.11	4.0

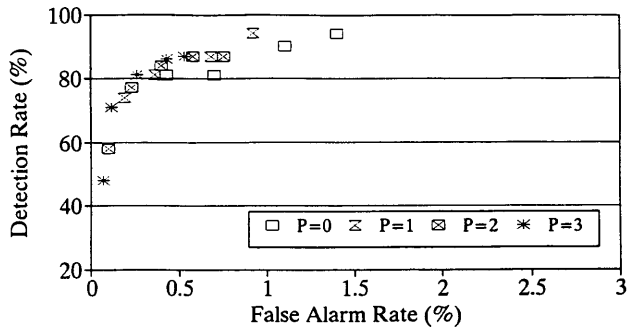


FIGURE 5 Neural network performance at 1,500 iterations.

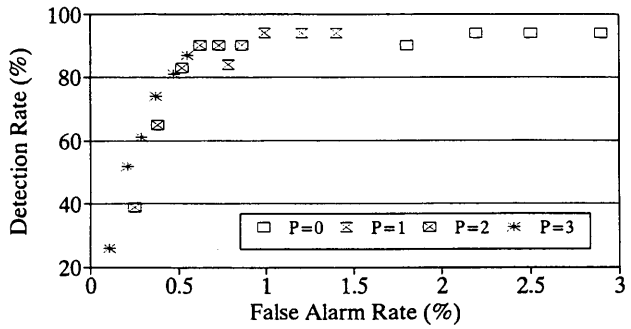


FIGURE 6 Neural network performance at 4,000 iterations.

or fast evolution of the congestion, respectively. The distinguishing logic of the two tests used is based on a temporal comparison of spatial occupancy difference between adjacent stations. Assuming an incident occurs at *t*, the congestion test considers the smoothed spatial occupancy difference from *k* time increments after *t*, normalized by the highest value of the smoothed upstream and downstream occupancies from *n* increments before *t*. The incident test compares the smoothed spatial occupancy difference for the period after *t* with the corresponding value from the past period. In particular, DELOS 1.1 (10,6) uses a moving average to smooth 10, 30-sec past occupancy values, and 6 present values. DELOS 3.3 (0.05,6) uses exponential smoothing with a smoothing factor of 0.05, and a time lag of 6 between the periods before and after the incident.

The evaluation results indicate that the neural network performs better than all algorithms in the set and, at a detection rate of 70 percent or higher, performs as well as the best algorithm, DELOS 3.3. This performance is noteworthy because the neural network represents the initial results in its class, developed in Minnesota with real data, whereas DELOS 3.3 was developed after considerable research. A more fair comparison would be between the neural network and DELOS 1.1, which also represents the initial findings in its class *(3)*. To illustrate, as Figure 7 suggests, at 70 percent detection rate, the false alarm rate of the neural network is 0.12 percent and that of DELOS 3.3 is 0.13 percent. The false alarm rate of DELOS 1.1 is 0.25 percent, or twice the number of false alarms of the neural network; that of Algorithm 7 is 0.34 percent, or approximately three times as many false alarms; and the false alarm rate of the California algorithm is 0.52 percent, or more than four times the number of false alarms produced by the neural network. Future versions of the neural network are expected to (*a*) improve time-to-detection by using a more appropriate pattern size, and (*b*) further decrease the number of false alarms by preprocessing and normalizing the field data before analysis.

CONCLUSION

A neural network algorithm that can be used to improve automated incident detection in freeways was discussed. Based on real-time occupancy and volume counts from pairs of adjacent loop detector stations, the three-layer feedforward network with approximately 1,000 nodes was trained with actual data, including 31 incidents from I-35W (a typical freeway in the Twin Cities Metropolitan

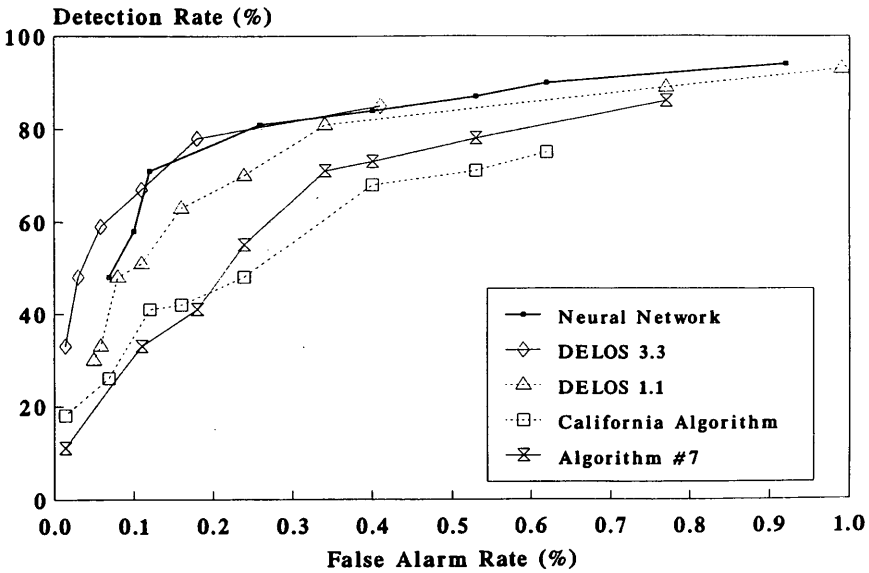


FIGURE 7 Algorithm performance comparison.

Area) over the afternoon peak period. Training with data from 72 days yielded promising test results and indicated that the algorithm was able to learn and classify incident and incident-free patterns effectively. Methods for improving the time-to-detect incidents are currently being developed by the authors.

Test results also indicated the sensitivity of algorithm performance to values of user-specified threshold, persistence, and the number of iterations used for training. Algorithm performance in terms of detection and false alarm rates was superior to most of the best algorithms that have been tested with this data-set. At a detection rate of 70 to 80 percent, the trained network has a false alarm rate of 0.12 percent to 0.26 percent. The computation time for one test is 4 msec on IBM-486/33MHz, indicating that it is practical to conduct incident detection in real time. Algorithm testing is continuing with the collection of additional incident data in the Metropolitan Area of Minneapolis-St. Paul.

ACKNOWLEDGMENT

The authors gratefully acknowledge Athanasios Chassiakos, Department of Civil-Mineral Engineering, University of Minnesota, for his helpful remarks. This work was supported in part by the National Science Foundation and the Minnesota Supercomputer Institute. The Center for Transportation Studies, Department of Civil-Mineral Engineering, University of Minnesota, is also acknowledged for its support. The Traffic Management Center, Minnesota Department of Transportation, cooperated in the study by providing the necessary data.

REFERENCES

1. Judicky, D., and J. Robinson. Managing Traffic During Nonrecurring Congestion. *Institute of Transportation Engineers Journal,* Vol. 62, No. 3, March 1992, pp. 21–26.
2. Kay, J. Intelligent Vehicle-Highway Systems and Incident Management. *Institute of Transportation Engineers Journal,* Vol. 62, No. 3, March 1992, pp. 55–57.
3. Stephanedes, Y., and A. Chassiakos. Application of Filtering Techniques for Incident Detection. *ASCE Journal of Transportation Engineering,* Vol. 119, No. 1, Jan./Feb. 1993, pp. 13–26.
4. Lippmann, R. An Introduction to Computing with Neural Nets. *ASSP Magazine,* Vol. 4, No. 2, IEEE, April 1987, pp. 4–22.
5. Hopfield, J., and D. Tank. Computing with Neural Circuits: A Model. *Science,* Vol. 233, Aug. 1986, pp. 625–633.
6. Carpenter, G., and S. Grossberg. The Art of Adaptive Pattern Recognition by a Self-Organizing Neural Network. *Computer,* Vol. 21, 1987, pp. 77–88.
7. Kohonen, T. *Self-Organization and Associative Memory.* Springer-Verlag, Berlin, 1989.
8. Rumelhart, E., G. Hinton, and R. Williams. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations* (D. Rumelhart and J. McClelland, eds), MIT Press, 1986, pp. 318–362.
9. Cheu, R., S. Ritchie, W. Recker, and B. Bavarian. Investigation of a Neural Network Model for Freeway Incident Detection. In *Artificial Intelligence and Civil and Structural Engineering* (B. Topping, ed), Civil-Comp Press pp. 267–274.
10. Chassiakos, A., and Y. Stephanedes. Smoothing Algorithms for Incident Detection. In *Transportation Research Record 1394,* TRB, National Research Council, Washington, D.C., 1993, pp 8–16.

Publication of this paper sponsored by Committee on Freeway Operations.

Development of Advanced Traffic Signal Control Strategies for Intelligent Transportation Systems: Multilevel Design

NATHAN H. GARTNER, CHRONIS STAMATIADIS, AND PHILIP J. TARNOFF

The development of advanced traffic signal control strategies suitable for advanced traffic management within Intelligent Transportation Systems is described. The strategies consist of a multilevel design for the real-time, traffic-adaptive control of an urban signal network. This design permits the system to be built up gradually to offer varying degrees of responsiveness, depending on particular network and traffic characteristics. The more advanced control levels in the hierarchy incorporate the capabilities of the lower control levels. A principal goal of the multilevel design is to invoke a selected control strategy when it can provide the greatest benefits and thus maximize the overall effectiveness of the system.

In November 1991 the FHWA issued a solicitation for the development and evaluation of a real-time, traffic-adaptive signal control system (RT-TRACS) suitable for use in an Intelligent Vehicle Highway Systems (IVHS) environment (*1*). The following information was provided as background:

> The FHWA's IVHS program consists of research and operational tests designed to combat traffic congestion. The thrust of the program is to develop and implement the technology necessary to mitigate the effects of congestion by maximizing the utility of existing transportation facilities. Some of the elements included in IVHS are transportation management, in-vehicle route guidance systems, integration of multi-modal transportation, integration of surface street and freeway control, traveler information systems, and incident management. In order for these elements to be integrated, a sophisticated traffic surveillance system must be deployed to provide the information necessary to enable the real-time traffic management.
>
> One of the key elements to these systems is a real-time, traffic adaptive signal control logic that enables the implementation of the traffic management and control strategies specified above. This control logic needs to not only assess the current status of the network, it must include forecasting capabilities such that proactive, not reactive, control is provided. To the extent possible, the signal control logic must be interfaced with freeway performance data and provide integrated, network-wide control.
>
> Current technology in real-time control is somewhat limited. In the United States, for example, there are no true real-time systems. Elsewhere, at least two systems have been developed and deployed at several locations; however, the applicability of these foreign systems to IVHS is only now being tested. This study is designed to answer this and many other similar questions from a technical perspective and be the focal point for the development of the signal control logic needed to support IVHS. (*1*) (Note: the term IVHS has since been changed to ITS.)

In this paper a multilevel design for RT-TRACS is presented. Each level incorporates a different methodology and has a different set of characteristics. The more advanced levels incorporate in a nested fashion the capabilities of the lower levels. This kind of design enables the system to be built up gradually to offer varying degrees of responsiveness, depending on particular network and traffic characteristics. This design provides the best combination of features for the overall optimization of traffic performance. Among the potential features that may be included in the RT-TRACS design are

- Both distributed and centralized traffic control;
- Traffic-responsive, on-line optimization, as well as background fixed-cycle control;
- Capability to interact with dynamic traffic assignment to implement proactive control;
- Dynamic priority control on selected routes;
- Congestion avoidance and congestion relief strategies; and
- Artificial intelligence technology to optimize strategy selection.

In addition, the following features are also included:

- Effective use of existing resources in a community,
- Coexistence of different control generations within one system,
- Improved fallback capabilities in case of surveillance system failure, and most important
- Effective use of the accumulated experience with real-time control.

The traffic engineering profession has more than 30 years of experience with computer control of traffic signals and the development and testing of various types of adaptive control strategies. It is important to make maximum use of this experience in the development of any new systems. In the next section, this experience is briefly reviewed.

REAL-TIME SIGNAL CONTROL: PAST EXPERIENCE

A thorough understanding of past experiences with advanced traffic signal control strategies is critical to the development of effective RT-TRACs strategies for ITS. Failure to do so may cause past mistakes to be repeated and the same pitfalls as encountered by past developers (*2*).

After the introduction of computer-based traffic signal control systems in the 1960s, numerous experiments were conducted to develop more advanced (i.e., more responsive) control strategies. One of the most comprehensive studies was the (Urban Traffic Con-

N. H. Gartner and C. Stamatiadis, University of Massachusetts-Lowell, Department of Civil Engineering, Lowell, Mass. 01854. P. J. Tarnoff, Faradyne Systems, Inc., 3204 Tower Oaks Blvd., Rockville, Md. 20852.

trol System (UTCS) experiment in the 1970s by the FHWA. The UTCS project was directed toward developing and testing a variety of advanced network control concepts and strategies and lasted for almost a decade. Its results defined the state of the art in the United States to the present. The UTCS experiment was described in detail in a work by MacGowan and Fullerton (*3*).

Research and testing of control strategies in the UTCS project was divided into three generations, as shown in Table 1. The same nomenclature is used in this paper. The different generations are characterized as follows.

First-Generation Control (1-GC)—This mode of control uses prestored signal timing plans that are calculated off-line based on historical traffic data. The plan controlling the traffic system can be selected on the basis of time of day, by direct operator selection, or by matching from the existing library a plan best suited to recently measured traffic conditions (volumes and occupancies). This is named the traffic-responsive (TRSP) mode of plan selection. The mode of plan selection is determined by the operator. Frequency of update in the traffic-responsive mode is 15 min. 1-GC software also includes logic to enable a smooth transition between different signal timing plans, and a critical intersection control (CIC) feature that enables vehicle-actuated adjustment of green splits at selected signals. NCHRP research (*4*) has cast doubt on the efficacy of the CIC algorithm and proposed to consider Optimization Policies for Adaptive Control (OPAC) driven controllers as substitutes. Plans in 1-GC can be calculated by any off-line signal optimization method, such as TRANSYT-generated plans, or progression optimization methods such as MAXBAND. The 1½-GC is a strategy in which new timing plans are generated automatically when traffic conditions warrant it.

Second-Generation Control (2-GC)—This is an on-line strategy that computes in real-time and implements signal timing plans based on surveillance data and predicted values. The optimization process can be repeated at 5-min intervals; however, to avoid transition disturbances, new timing plans cannot be implemented more often than once every 10 min. 2-GC software contains an optimization algorithm (SIGOP), a traffic prediction model, subnetwork configuration models, CIC, and a transition model to minimize transition time between two plans.

Third-Generation Control (3-GC)—This strategy was designed to implement and evaluate a fully responsive, on-line traffic control system. Similar to 2-GC, it computed control plans to minimize a networkwide objective using predicted traffic conditions for input. The differences compared to 2-GC were that the period after which timing plans were revised was shortened to 3 to 5 min, and that cycle length was allowed to vary among the signals as well as the same signal during the control period (CP). This was accomplished by dividing the CP into an integral number of intersection-specific control intervals (CI), which were calculated on the basis of predicted volume and capacity ratios on each approach to the intersection using a Webster-like method. Thus, control intervals were approximately equal to the expected value of cycle length at each intersection. However, the switching points within each CI were also determined by the on-line optimization procedure. In this method, a simplified model of traffic flow scenarios was used to trace performance, in an iterative manner, through strings of mini-star networks composed of the signalized node and all its incoming and outgoing links (the CYRANO model, *5*).

The dynamics of the control plan generation and implementation for the three UTCS strategies are illustrated in Figure 1. It is noteworthy that 3-GC is similar in concept to 2-GC. In both strategies the traffic data used in the timing plan being implemented are displaced by at least two control periods from the actual flow measurements. Therefore, the effectiveness of the control system response depends entirely on the quality of the prediction model. The different UTCS control strategies were designed to provide an increasing degree of traffic responsiveness, with an expectation to capitalize on the variability in traffic flows and to provide an improvement in urban street network performance. However, results of extensive field testing demonstrated that these expectations were not entirely fulfilled *(3)*.

1-GC, in its various modes of operation, performed best overall and demonstrated that it can provide measurable reductions in total travel time over that which could be attained with a well-timed three-dial system (see Table 2). The traffic-responsive mode of 1-GC plan selection was generally more effective than the time-of-day mode. The 2-GC strategy was mixed but was overall inferior compared with 1-GC; it demonstrated some small improvements on

TABLE 1 Characteristics of UTCS Control Strategies

FEATURE	1-GC	2-GC	3-GC
Update interval	15 min	5-10 min	3-5 min
(Control Period)			(Variable)
Control plan generation	Off-line optimization selection from library by time-of-day, traffic responsive, or manual mode (7 plans used)	On-line optimization	On-line optimization
Traffic prediction	None	Historically based	Smoothed values
CIC	Fine tuning of splits	Fine tuning of splits and offsets	NA
Cycle length	Fixed within each section	Fixed within variable groups of intersections	Variable in time and space Predetermined for control period

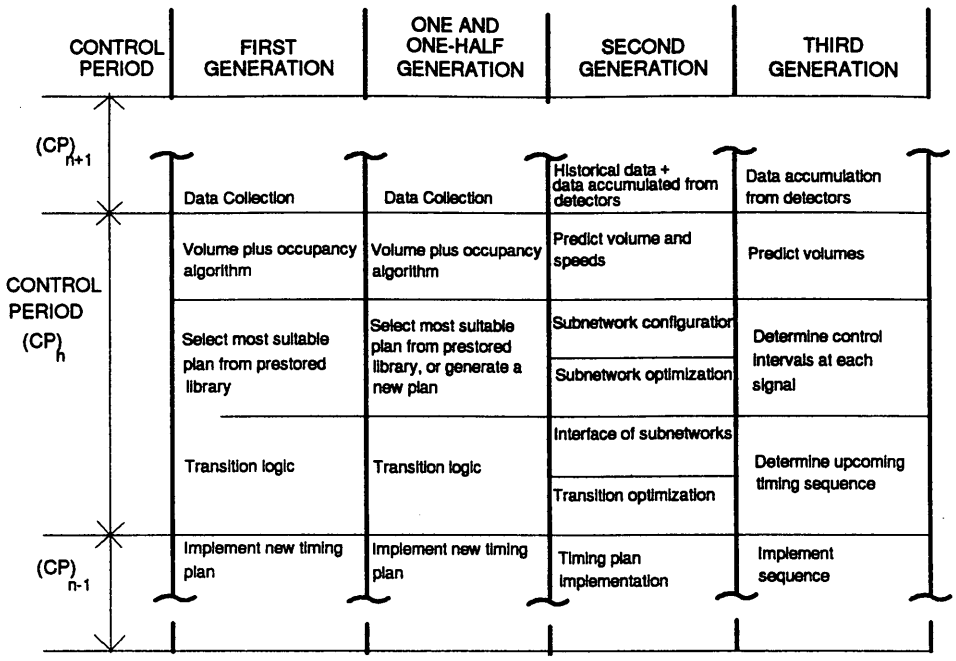


FIGURE 1 Dynamics of control plan generation and implementation in UTCS strategies.

the arterial but degraded traffic flow in the network. The 3-GC strategy, in the form tested in the UTCS system, was unsuccessful in responding to traffic flows and degraded performance under almost all conditions for which it was evaluated.

Thus, the more responsive strategies resulted in poorer performance than the fixed-cycle, nonresponsive strategies. This appeared to be counterintuitive. On the basis of these results, one might erroneously conclude that a library of timing plans generated off-line, based on historical data (from another day, another month, perhaps another year, but for the same time period of the day), is more effective than timing plans generated on-line, based on very recent data (the past 15, 5, or 3 min). However, a closer examination of the experiments reveals that the expectations were not fulfilled not because their rationale was wrong (that traffic-responsive control should provide benefits over fixed-time control), but because of a failure of the models and procedures that were used in the UTCS study to deliver the desired results.

Possible causes of the poor showing of the responsive UTCS strategies (*6,7*) follow.

- Because of the inherent inaccuracies in the measurement-prediction cycle, neither the 2-GC nor the 3-GC strategies could respond adequately to rapid changes in traffic flows.
- The frequent transition in signal timing may be more harmful than the adaptiveness sought by on-line optimization in 2-GC and in 3-GC. Considerable delays are incurred during the transition process.
- Although 3-GC had a variable-cycle feature, the entire signal switching sequence was predetermined for every control period and therefore not dynamically responsive to the actual traffic conditions on the street. In essence, it was a 2-GC strategy with an imposition of a different cycle length at each intersection. The benefits obtained from a locally optimal cycle length were insufficient to outweigh the loss of the benefits of synchronization among the intersections of the network.
- The control strategy used by 3-GC was centralized and the optimization procedure required a comparatively long time for convergence (much longer than 2-GC for the same size network). The

TABLE 2 Comparison of Results of UTCS Strategies

Traffic Responsive Strategy		% change in aggregate veh.-min. of travel with respect to base			
		AM Peak	Off Peak	PM Peak	All Day Average
1-GC	Arterial	-2.6	-4.0	-12.2	NA
	Network	-3.2	+1.9	-1.6	NA
2-GC	Arterial	-1.3	-3.8	+0.5	-2.1
	Network	+4.4	+1.9	+10.7	+5.2
3-GC	Arterial	+9.2	+24.0	+21	+16.9
	Network	+14.1	-0.5	+7.0	+8.2

time allotted for this procedure was insufficient for reaching a good optimum.

The failure of the more responsive strategies to accomplish their stated goals was not limited to the UTCS experiment. Similar results were experienced by the British Transportation Research Laboratory (*8*) and by Metropolitan Toronto Traffic Department (*9*). The basic premise has always been that on-line traffic control strategies should be capable of providing results that are better than those produced by the off-line methods. Because this premise was not accomplished in the experiments that were conducted during the 1970s, it was clear that new strategies had to be developed to be able to implement it successfully. To achieve this goal, the following prescription for the development of an effective demand-responsive traffic control system was offered in 1982 (*10*):

1. The system must be designed to provide better performance than off-line methods. Although this may appear self-evident, it was not always recognized explicitly in the development of responsive strategies in the past.
2. Development of new concepts is needed and not merely the extension of existing concepts. Effective responsiveness is not achieved by implementing off-line methods at an increased frequency. New methods that are better suited to the variability of traffic must be developed.
3. The system must be truly demand-responsive, that is, to adapt to actual traffic conditions and not to predicted values that may be far off from the actual conditions. As a corollary to this principle, if the traffic conditions cannot be adequately sensed or predicted, then lower-generation strategies may be more advantageous than higher-generation strategies.
4. It should not be arbitrarily restricted to control periods of a specified length, but be capable of frequent updating of plans as necessary.

A significant advance toward these goals was achieved during the 1980s with the introduction of SCOOT in the United Kingdom (*11*), which may be considered a 2½-GC strategy, and by SCATS in Australia (*12*), which is considered a 1-GC/TRSP variant. These strategies are noteworthy, especially in their ability to generate timings on-line. However, a close examination of field test results reveals that the methods score both successes and failures compared with traditional fixed-time control. For example, floating car surveys that were conducted in the course of the evaluation of the SCOOT method produced results of the type shown in Figure 2 (*11*). It is evident that the advanced method does not always perform better than the base method. Sometimes its performance is inferior or indistinguishable.

These principles serve as a basis for the development of a new family of highly advanced adaptive control strategies: OPAC (*13*), PRODYN (*14*), and UTOPIA (*15*). After the OPAC lead, all strategies adopted a dynamic programming optimization methodology within a rolling horizon framework. This approach shows great promise as an element of the RT-TRACS design but has not yet

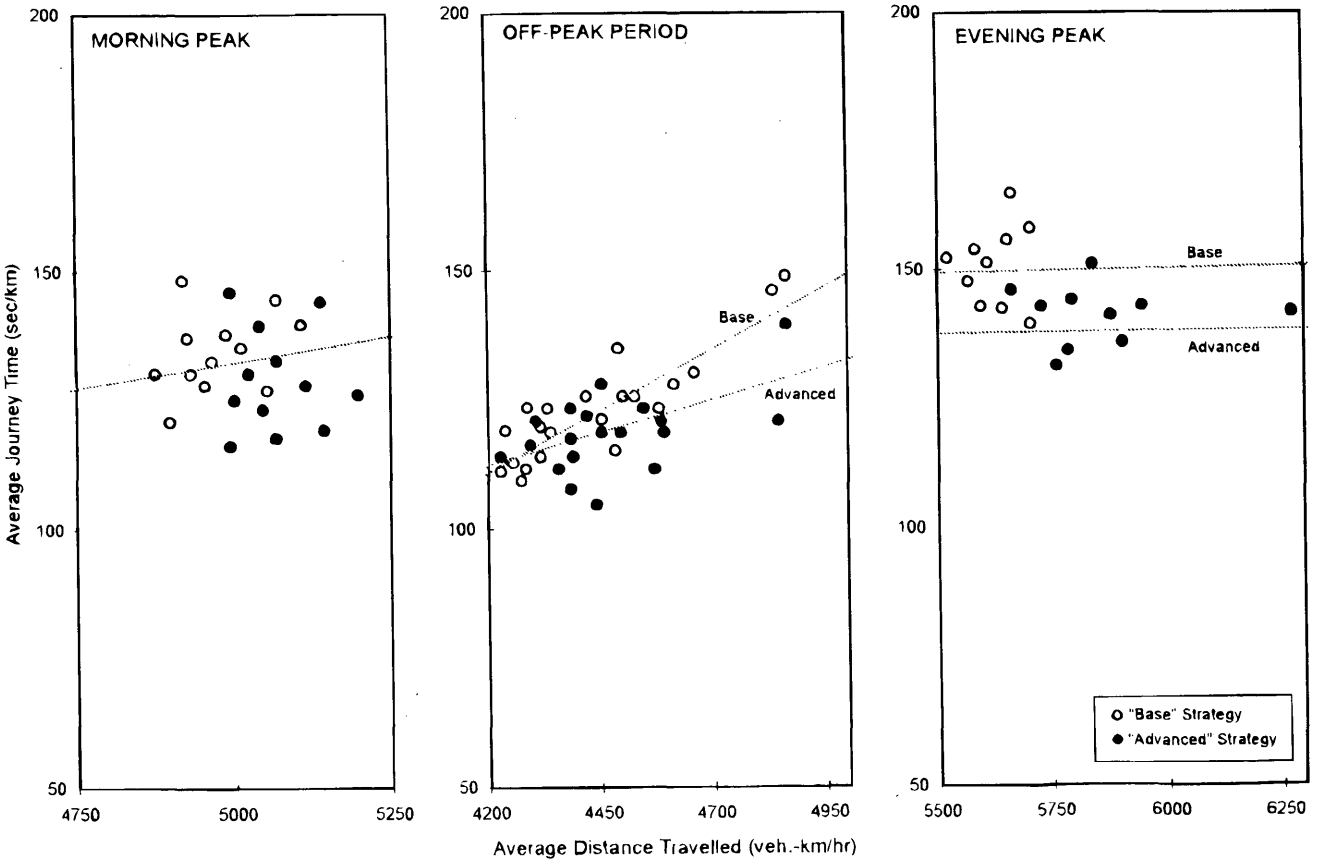


FIGURE 2 Examples of floating car surveys in evaluation of advanced strategies.

been fully developed and tested. Further discussion of this approach is given in the next section, which provides new framework, or architecture for the progressive development of the RT-TRACS strategies to optimize system performance.

ADVANCED CONTROL STRATEGIES: MULTILEVEL DESIGN

In the new framework for advanced traffic control, a multilevel design is offered in which each level in the hierarchy encompasses, in a nested fashion, the capabilities of the lower levels. The control levels correspond, in some respects, to the different UTCS control generations; however, they also contain significant differences and enhancements. Descriptions of the envisioned levels of control follow.

0-LC: Base Control—0-LC: Base Control is the most basic type of signal control and is a mix of fixed-time and traffic-actuated controllers and some arterial coordinated systems. The traffic engineer determines timing patterns manually, or by some PC-based programs, using available historical volume data. Various arterial progression schemes may be used. Manual adjustment of timings in the field is used because of inaccurate or inadequate data. Timings evolve by trial and error. There is infrequent updating of plans, typical of small-town operation or isolated sections of urban areas with limited technical support.

1-LC: First-Level Control—This level is similar to the UTCS/1-GC system. There is centralized control with limited availability of traffic surveillance and communications. Timing plans are calculated off-line using historical data and stored in the computer's data base. A variety of arterial and network optimization packages may be used, such as TRANSYT, PASSER II, MAXBAND, MULTIBAND, and so forth. The plans can be implemented by time of day, manual activation, special events, or traffic-responsive modes. SCATS is an advanced version of the latter mode. 1½-LC is available as an option to enable automatic updating of timing plans for changing traffic and network conditions.

2-LC: Second Level Control—A basic 2-GC system follows the UTCS designation: a centralized control with on-line optimization was a fixed, common cycle time for dynamic subnetwork configurations. Typical timeframes for the optimization are between 5 and 10 min. Traffic volumes are predicted for the upcoming interval during which the new timing is to be implemented. Optimization is performed by an off-line strategy that was adapted for on-line control, such as SIGOP, RTOP, or PRINET (PRIority NETwork optimization). It requires more advanced computational, surveillance, and communication capabilities than 1-LC. Advanced versions of 2-LC systems (perhaps designated as 2½-LC) include SCOOT.

3-LC: Third Level Control. This is a fully adaptive traffic signal control system that incorporates the capabilities of all the previous levels, yet may relinquish their restrictive characteristics in favor of improved traffic performance; for example, a common cycle time is not required for coordination, although it is not necessarily excluded from the control parameter set. Optimization of phase sequence capabilities may be available at selected locations. It consists of on-line, dynamic optimization that can be performed centrally or and distributively, or both, through a network of smart controllers. Subnetworks can be configured dynamically to carry out optimal policies as required by the overall optimization objective. An advanced surveillance system is required for this control level. It meets specifications of the original UTCS/3-GC.

An example of a strategy that meets the requirements of 3-LC is the OPAC strategy originally developed by Gartner *(13)*. OPAC is based on dynamic programming (DP) optimization, which generates optimal signal switching sequences in real time. The OPAC strategy was initially implemented for individual, distributed intersection control and has exhibited good performance in field testing *(16)*. It is now being extended for network operation under the RT-TRACS research program. A bilevel hierarchy is being used, where the upper level ensures the optimal coordinated operation of the individual smart controllers at the lower level. A schematic of the operation is shown in Figure 3. The primary signal is a smart controller that runs the OPAC algorithm and interacts with the neighboring satellite controllers to optimize local performance. Each signal is, in turn, a primary signal in its own mininetwork. In this way, the entire network becomes interconnected and coordinated.

4-LC: Fourth Level Control—This system level includes all the capabilities of 3-LC with additional intelligence as follows:

- Can interact with a dynamic traffic assignment module in the ATMS to implement proactive control (*i.e.*, in anticipation of the projected traffic volumes and routes),
- Can provide dynamic priority control on selected routes,
- Can implement congestion avoidance strategies, and
- Can implement congestion relief strategies.

5-LC: Fifth Level Control—This is a super level that incorporates the capabilities of all the previous levels. Most important, it makes the most efficient use of the control strategies in those systems based on accumulated expertise and experience under local conditions. Selection of the appropriate control strategy for the particular conditions is done by artificial intelligence technology. An analysis and explanation of the underlying basis for this design is given in the next section.

INTELLIGENT STRATEGY SELECTION

The hierarchical framework described will offer the possibility to tailor the system's capabilities to particular needs and means. Advanced strategies can be deployed gradually and can coexist with lower-level strategies. The full spectrum of capabilities can be built up gradually in terms of deployment of sensors, surveillance equipment, controllers, and communications, as well as central control hardware and software.

It has been common wisdom that increasing responsiveness contributes to improved traffic performance. As indicated previously, this is not always true. Consider the following facts based on evidence obtained from carefully conducted field experiments that were reported in the literature *(10)*.

- FACT 1. TRANSYT settings are frequently used as the base for comparison of advanced strategies (such as SCOOT). But it has been demonstrated that, in some cases, other off-line methods can provide significant improvements in performance over TRANSYT. Therefore, when an advantage of a responsive method is claimed, it should be analyzed whether a more suitable off-line method would have given the same, or better results.
- FACT 2. Experiments with 2-GC methods have shown swings in both directions: improvement in some cases, degradation in others *(3,9)*. The situations in which 2-GC methods are advantageous

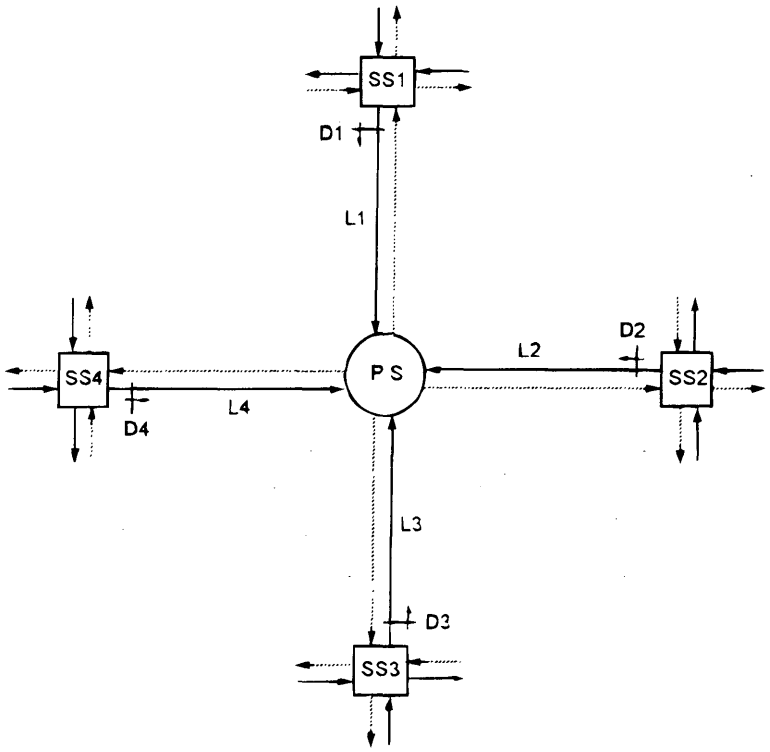


FIGURE 3 Architecture of OPAC network operation. (PS = primary signal, SS = satellite signal, D = detector, L = link)

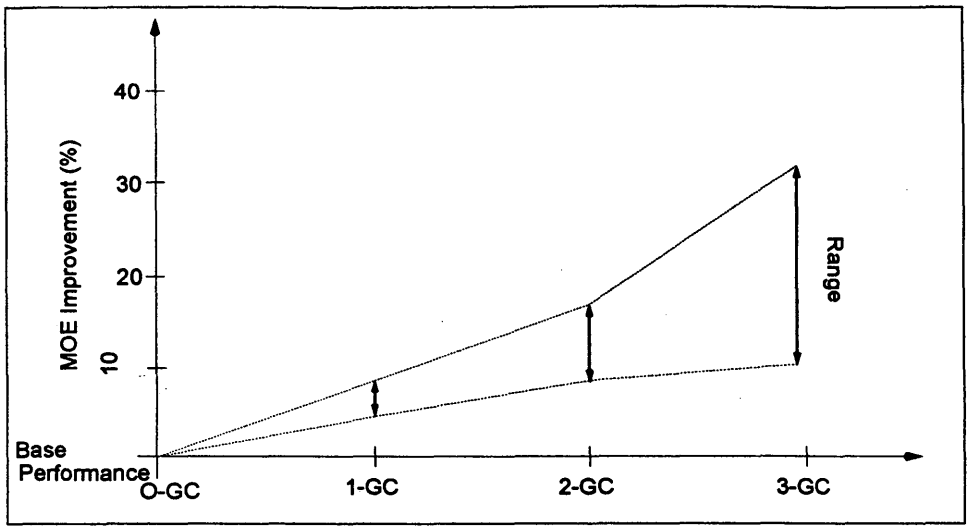
should be characterized and not used when they are disadvantageous.

• FACT 3. Traffic-responsive methods such as SCOOT or OPAC also have shown considerable swings in performance under different traffic conditions. An example of this behavior is shown in Fig. 2, which illustrates the field evaluation of a responsive method relative to a base fixed-time plan. It is evident that (a) in some cases, no overall advantages are realized compared to nonresponsive methods and (b) in most cases, the performance data have a wide spread around the average. Thus, although responsive methods can provide substantial benefits compared with nonresponsive methods, they are also likely to degenerate into poor performance if not properly applied.

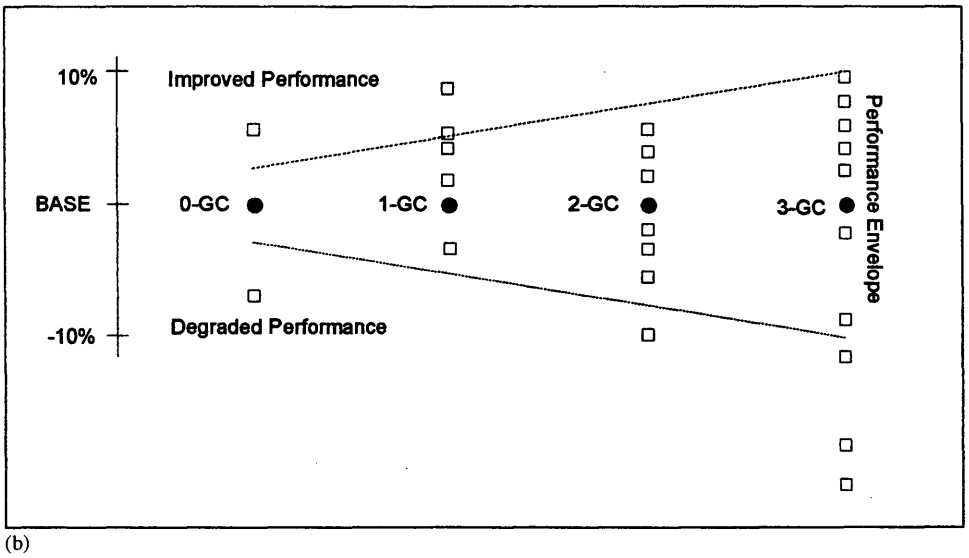
The argument is best illustrated by Figure 4. It shows the spreads in performance that are perceived to exist among the different control generations (Figure 4a), as well as the spreads in performance that were actually measured in the field (Figure 4b). The points shown in the figure were collected from published reports in the literature (*3,8,9,11*). These figures illustrate an interesting phenomenon: the more advanced (ie., responsive) strategies do not lead to improved performance all the time, notwithstanding the common perception that they do. More likely, they lead to a wider spread in performance results compared with a common basis. In some cases, because of reasons that are not completely explained, traditional off-line methods perform better than responsive methods. Therefore, one of the principal objectives of any intelligent control system would be to invoke a particular control strategy that will be most suitable for existing conditions so that overall performance of the system is optimized.

It is envisioned that an expert system can be developed to ensure that a particular generation of control strategies will be implemented when it will provide the maximum benefits. Development of the system will require a careful characterization of signal networks and the identification of the particular traffic flow patterns that would be most amenable to benefit from a particular control strategy. The underlying proposition is that lower-level strategies may often be as good or better than higher level more advanced strategies. By recognizing the conditions under which the particular strategies perform best, overall system performance can be optimized. This would be the role of the 5-LC control level.

An overview of the operational flow diagram of the 5-LC strategy selection process is shown in Figure 5. This is a dynamic system in which the first step is the evaluation of the initial traffic conditions, as well as evaluation of the impact of external factors like weather conditions, type of day, and so forth. Based on this information, an initial control strategy is selected, best suited for the identified conditions. The strategy is then implemented and its performance is continually evaluated. The performance of the strategy, coupled with a dynamic traffic assignment model and the expected response of the users of the system to the control strategy, is used to predict traffic characteristics in the next control period. On the basis of these predictions, the system determines whether a new control policy is warranted. In case a new control policy is required, the most beneficial one is selected. If no change is needed, the performance continues to be monitored. In this way the performance



(a)



(b)

FIGURE 4 *(a)* **Expected relative performance of control generations;** *(b)* **reported relative performance of control generations.**

of the system is continuously evaluated and control strategies are updated only when conditions warrant. On-line simulation can be used as an aid in evaluating the performance of the system.

CONCLUSIONS

Although advanced technologies enable development of ever more sophisticated strategies, experience has taught that such strategies do not always result in improved performance. A major contribution of the new technologies will be to enable identification and recognition of the particular characteristics of the traffic system and selection of the most appropriate strategy for any existing situation. Frequently, such strategy might be a traditional fixed-cycle network optimized signal pattern or an arterial progression scheme. It does not necessarily have to be a real-time optimized control strategy. In other words, by providing a menu of strategies, one is likely to do better than by tying oneself to one particular strategy. The conclusion to be drawn from examining previous studies is the following: let us not throw out the experience gained in the past 30 years, but let us build on it gradually to develop improved traffic signal control operations. This is the basis for the framework proposed in this paper. Much of the development will depend on experimentation with alternative strategies and learning from experience in the field.

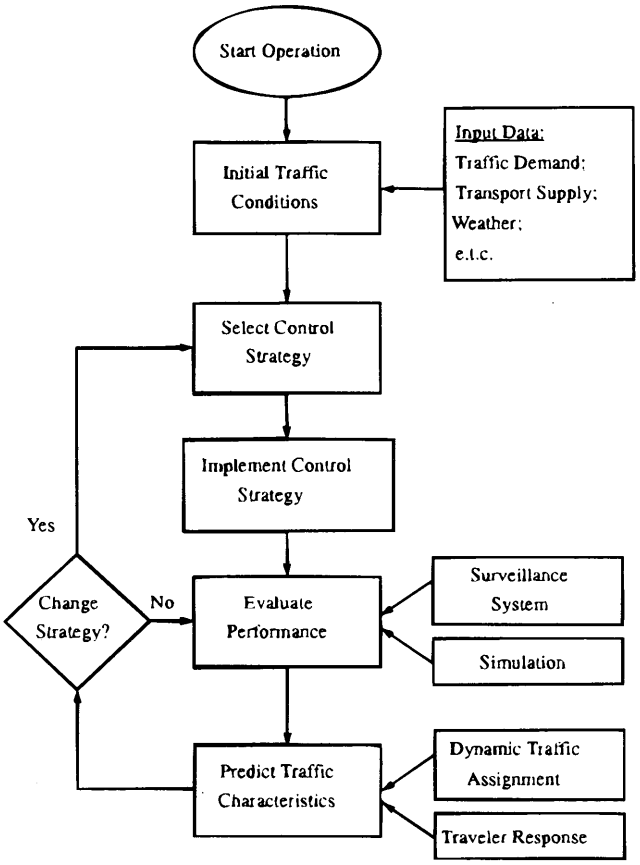


FIGURE 5 Operational flow diagram for strategy selection process.

ACKNOWLEDGMENTS

The paper is based, in part, on a research project sponsored by FHWA of the U.S. Department of Transportation with Farradyne Systems, Inc., as prime contractor and the University of Massachusetts, Lowell, as subcontractor.

REFERENCES

1. *Real-Time, Traffic Adaptive Control for IVHS.* Solicitation No. DTFH61-92-R-00001, FHWA, U.S. Dept. of Transportation, November 1991.
2. Tarnoff, P. J., and N. H., Gartner. Real-Time, Traffic Adaptive Signal Control. *Proc., Advanced Traffic Management Conference,* St. Petersburg, Fla., Oct. 1993.
3. MacGowan, J., and I. J. Fullerton. Development and Testing of Advanced Control Strategies in the Urban Traffic Control System (three articles). *Public Roads,* Vol. 43 Nos. 2–4, 1979–1980.
4. *Traffic Adaptive Control, Phase I: Critical Intersection Control Strategies.* Farradyne Systems, Inc., NCHRP Project 3–38, Vol. 3, Sept. 1989.
5. *Variable Cycle Signal Timing Program.* National Technical Information Service, KLD Associates, Inc., May 1974.
6. Gartner, N. H. Urban Traffic Control Strategies: The Generation Gap. *Proc., 2nd International ATEC Congress on Traffic and Transportation Control Systems,* Paris, France, 1980.
7. Gartner, N. H. Demand-Responsive Traffic Signal Control Research. *Transportation Res.19A,* 1985.
8. Robertson, D. I. Traffic Models and Optimum Strategies: A Review. *Proc., International Symposium on Traffic Control Systems,* Berkeley, Calif., Aug. 1979.
9. *Improved Operation of Urban Transportation Systems,* Vol. 1–3, Metropolitan Toronto Corp., Toronto, Canada, 1974–1976.
10. Gartner, N. H. Prescription for Demand-Responsive Urban Traffic Control. In *Transport Research Record 881,* TRB, National Research Council, Washington, D.C., 1982, pp. 73–75.
11. Hunt, P. B., D. I. Robertson, R. D. Bretherton, and R. I. Winton. *SCOOT: A Traffic Responsive Method of Coordinating Signals.* TRRL Report LR 1014, United Kingdom, 1981.
12. Lowrie, P. R. SCATS: The Sydney Co-Ordinated Adaptive Traffic System. *International Conference on Road Traffic Signalling.* IEE, London, England, 1982.
13. Gartner, N. H. OPAC: A Demand-Responsive Strategy for Traffic Signal Control. In *Transportation Research Record 906,* TRB, National Research Council, Washington, D.C., 1983.
14. Henry, J. J., and J. L. Farges. PRODYN. *Proc., 6th IFAC/IFIP/IFORS Symposium on Transportation,* Paris, France, 1989.
15. Di Taranto, C., and V. Mauro. UTOPIA. *Proc., 6th IFAC/IFIP/IFORS Symposium on Transportation,* Paris, France, 1989.
16. Gartner, N. H., P. J. Tarnoff, and C. M. Andrews. Evaluation of Optimized Policies for Adaptive Control Strategy. In *Transportation Research Record 1324,* TRB, National Research Council, Washington, D.C., 1991, pp. 105–114.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

REALBAND: An Approach for Real-Time Coordination of Traffic Flows on Networks

PAOLO DELL'OLMO AND PITU B. MIRCHANDANI

An approach is proposed for real-time coordination of signal phase timings for a network. Currently, network coordination is done using off-line methods, such as MAXBAND, PASSER II, and TRANSYT, which are based on average traffic volumes for various movements. On-line approaches such as SCOOT adapt off-line methods by constantly inputting updated average volumes computed from detector data over the "last" decision horizon. REALBAND first identifies platoons and predicts their movement in the network (i.e., their arrival times at intersections, their sizes, and their speeds) by fusing and filtering the traffic data obtained, from various sources, in the last few minutes. An approximate traffic model, APRES-NET, is used to propagate the predicted platoons through the network for a given time horizon. The signals are set so that the predicted platoons are provided appropriate green times to optimize a given performance criterion. If two platoons demanding conflicting movements arrive at an intersection at the same time, then either one or the other will be given priority for green time, or one of them is split to maximize the given measure of performance. This study discusses how such conflicts are resolved and the corresponding algorithmic procedure of REALBAND.

Since the early 1970s, several cities in the United States, Australia, Europe, and elsewhere have implemented traffic control systems in which a network of intersections is centrally controlled by a mainframe or a minicomputer. Most of these systems have (*a*) magnetic loop detectors near the intersection to detect arriving vehicles and (*b*) a microprocessor-based local controller at each intersection where traffic control parameters are input manually or downloaded through communication links, such as telephone lines, twisted pair cable, or cable television lines.

Current implementations, even those that are state of the art, have some drawbacks due to inherent technological constraints imposed on the system design. However, these drawbacks are gradually being eliminated with the rapid advances in detector, communication, and computer technologies provided by the Intelligent Transportation Systems (ITS) program in the United States, and similar high-technology-based programs in Europe and Japan.

These modern technologies, combined with methodological advances in control theory and operations research, can be used to develop a control system for real-time traffic management to improve overall traffic system performance. A hierarchical control architecture recently proposed by Head, Mirchandani, and Sheppard (*1*) uses the capabilities of modern technologies and exploits availability of real-time data. The system that is being developed based on this architecture, referred to as RHODES, calls for a modular implementation of the subsystems responding to the various hierarchical control functions within the control structure, namely network load control, network flow control, and intersection control. The hierarchical control system is schematically represented in Figure 1.

This study deals with the second level of the hierarchy: network flow control. At this level, decisions and actions for real-time coordination of traffic flows on the network are implemented by coordinated intersection phasing. This has proven to be a challenging real-time control problem.

Prototypical off-line approaches to network coordination are TRANSYT (*2*), MAXBAND (*3*), and PASSER II (*4*), the latter two being predominately for arterial coordination. Although the original MAXBAND model allowed the optimization of signal timings in a network, the model has been used primarily for coordinating arterials. Recent enhancements of MAXBAND, embodied in MAXBAND-86 (*5*) and PASSER IV (*6*), have made its implementation to grid networks possible; however, applications to actual networks are still lacking.

The basic ingredients of these methods include (*a*) a traffic flow model and (*b*) an algorithm for optimizing a specified performance criterion (this criterion could be a weighted sum of several performance indices). For example, in TRANSYT, vehicles are "loaded" onto the network at given origins and are propagated through the network in accordance with a traffic flow model. Traffic controls affect the movement of these vehicles, and numerical optimization (gradient search) is performed to find controls that optimize the specified performance criterion. In MAXBAND and PASSER II, vehicles are loaded on an arterial, and traffic signals on that arterial are coordinated to optimize a performance criterion, which often relates to the number of stops. Because these are off-line methods, assumptions on the traffic loads are based on historical average volumes, which are uniformly loaded onto the arterials. This results in an assumption of platoons of uniform size and identical speeds.

A notable extension is the MULTIBAND model (*7*), which allows the bandwidth in each section of an arterial to be different. This accounts for turn-in and turn-out traffic that causes volume differences in the various arterial sections. Although simulations have shown that MULTIBAND performs better than MAXBAND, it is still an off-line method that uses historical average volumes and assumes platoons of uniform size and identical speeds on the arterial sections.

TRANSYT may be used in an on-line fashion to compute signal settings every few minutes and download those settings to the field. In a way, this is exactly what SCOOT (*8*) does. However, the current versions of SCOOT that the authors know about have the disadvantage that platoons in the network may not experience sufficient platoon progression, or any other desired platoon-based performance. Ad-hoc approaches have been suggested to enhance SCOOT to consider platoon progression; however, to the authors' knowledge, these have not been implemented. Although

P. Dell'Olmo, Electronic Engineering Department, University of Rome "Tor Vergatta," Viale della Ricerca Scientifica, 00133 Rome, Italy. P. B. Mirchandani, Systems and Industrial Engineering Department, University of Arizona, Tucson, Ariz. 85721.

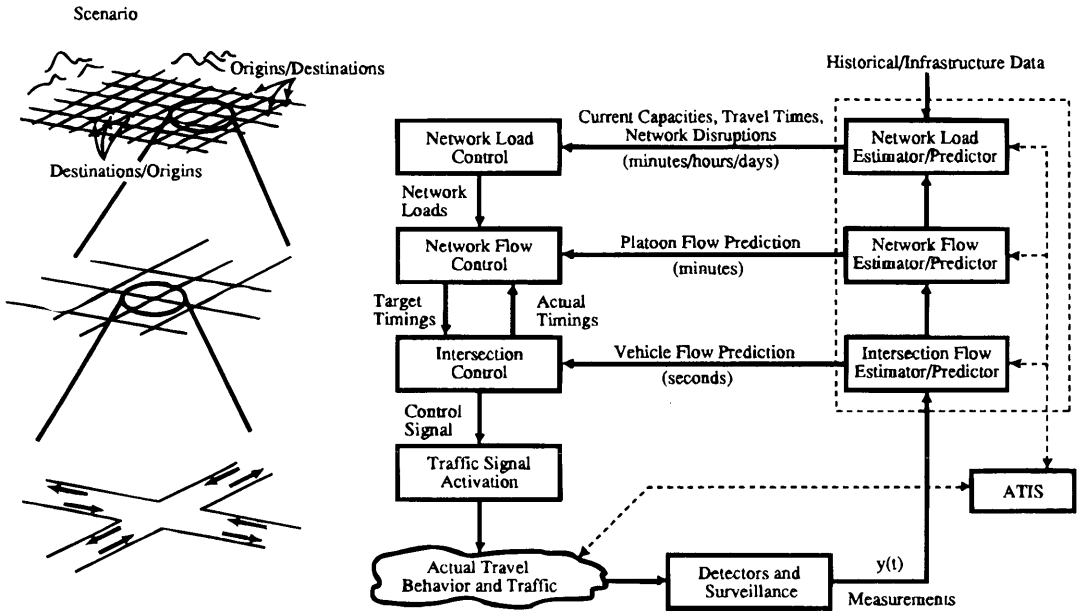


FIGURE 1 Hierarchical control architecture for traffic management.

TRANSYT has been modified to include progression opportunities (*9*), it has not been implemented for real-time applications. It is not clear whether this approach is amenable for real-time applications due to its excessive computational requirements. Furthermore, TRANSYT, and for that matter SCOOT, do not explicitly consider the currently measured, in real-time, traffic flows (i.e., platoons and their speeds), but instead take the current data and assume a uniform flow of the current volumes.

THE "REALBAND" APPROACH

The approach presented in this study explicitly considers available real-time information for computing signal timings. It first identifies platoons and predicts their movement in the network (i.e., their arrival times at intersections, their sizes, and their speeds) by fusing and filtering the traffic data obtained, from various sources, in the last few minutes. An approximate traffic model is used to propagate the predicted platoons through the network for a given time horizon. The signals are set so that the predicted platoons are provided appropriate green times to optimize a given performance criterion.

Two platoons demanding conflicting movements may arrive at an intersection at the same time. In that case one will be given priority on the green time, or one of the platoons will be split to maximize the given measure of performance. Optimally resolving such conflicts in real time is the main objective of the algorithm presented, which, for brevity, is referred to as *REALBAND*.

The time-distance diagram on a single arterial is shown in Figure 2. The goal of off-line arterial progression algorithms, such as MAXBAND and PASSER II, is to set the signal timings so that the

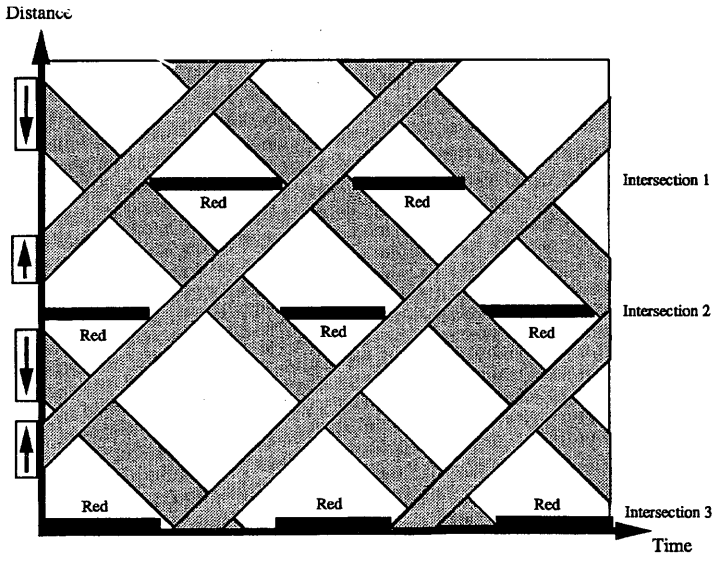


FIGURE 2 The MAXBAND concept.

number of vehicles that can traverse the arterial in either direction without stopping (other similar criteria may be incorporated) is maximized. The figure shows these bands of green times. Note the following drawbacks: it is assumed that sets of platoons of equal size are distributed in a cyclic manner, and that platoons travel at the same constant speed.

The time-distance diagram in Figure 3 (*a*) shows platoons of different sizes and different speeds. Because the green times required for these platoons are different from those required for the uniform case shown in Figure 2, the smooth anticipated progression is disrupted. By slightly adjusting the red times, it may be possible to reinstate the green bands for the given platoons with their own sizes and speeds [see Figure 3(*b*)]. Of course, the identified platoons on the cross streets must also be considered when the green times are adjusted so that cross street traffic does not get delayed unnecessarily; *REALBAND* does consider this. Platoon dispersion and compression, although not shown in the figure, may also be included. Also, the illustrations do not show turning vehicles, which could increase or decrease platoon sizes and the speed differences given in the figures also have been purposefully exaggerated. This is so that the proposed concept for network flow control is more easily visualized. The approximate flow prediction model (discussed later) addresses each of these characteristics.

If the intersecting platoons fit exactly within the red times shown here, then it is not necessary to resolve green-time demand of conflicting movements. On the other hand, if flows at an intersection produced a concurrent green-time demand for conflicting movements, then the conflict must be resolved by determining to which movement the green time must be allocated. Figure 4, which shows platoons on two other perpendicular arterials at Intersections 2 and 3, illustrates this scenario.

REALBAND makes a forward pass in time. When a conflict arises a decision node in a tree is formed; the types of decisions at this node include: (*a*) give green time to Platoon A, (*b*) give green time to platoon B, or (*c*) split Platoon A (or Platoon B, because only one or the other platoon needs to be split). Each branch of the tree is propagated over time to keep track of the total performance up to the decision node plus the performance on the link associated with the potential decision. An implicit approximation is used on the additive nature of the performance measure to propagate from node to node in the decision tree.

Figure 5 gives the current prediction of the movement of the platoons shown in Figure 4. The first demand conflict arises between Platoons N and W3 at Intersection 3. To resolve the conflict, Platoon N (Figure 6) is split or platoon W3 is stopped (Figure 7). Considering the resulting predictions shown in Figure 6, the next conflict arises between Platoons S and E3. Here the decision is either to stop Platoon S (Figure 8) or Stop E3 (Figure 9). In this way, a decision tree is formed that keeps track of various candidate decisions as demand conflicts arise. For this illustration, the decision tree for the predictions that arise for various decisions is given in Figure 10.

When the time horizon is reached, associated with each end node will be the total cost of the all the decisions leading up to the node on the path from the root of the decision tree to the end node (leaf) of the decision tree. Selecting the one with minimum cost provides the least cost trajectory of conflict resolution decisions. A final backward pass provides a phase plan within the time horizon considered for the identified platoons. This is passed to the third level of the hierarchical traffic control system (intersection control logic) as constraints (and, hence, an initial cut at a sequence of phases) that specify the "winning phase" from the outcome of each conflict resolution on the optimal root-to-leaf path in the decision tree. Further optimization is performed at the intersection level, at which more detailed data on individual vehicle movement are gathered. For the platoons shown in Figure 4, choosing the path with optimum performance (in this case minimum total delay), the resulting optimal decisions from the decision tree are shown in Figure 11, which includes the red and green times for the N-S arterial. It indicates that at Intersection 3 Platoon N should not be stopped but Platoon W3 should be stopped and, later, Platoon E3 should not be stopped but Platoon S should be stopped, when the corresponding demand conflicts arise.

The advantages of the *REALBAND* approach include:

1. Using real-time data, REALBAND explicitly identifies the platoons and predicts their movement in the network; the method also sets traffic signals to respond to the identified platoons.
2. REALBAND does not necessarily require a predetermined sequence of phases. The output provides an initial cut at a sequence of phases for further optimization at the lower intersection level.

A final issue that needs to be resolved in the REALBAND method is the computation of performance measures, (e.g. the total number of stops, total delay, etc.). To do this, concepts from TRAF-NETSIM (*10*) and TRANSYT are used to create a quick-and-dirty simulation to evaluate the performance of a set of signal settings. For real-time applications, these performance measures are needed quickly so that the performance criterion may be optimized in real time. A detailed simulation becomes computationally unwieldy when the simulation model is used as a function evaluator (i.e., for evaluating the performance function for each candidate signal setting) in an optimization routine. To be in the proper range for the optimizations being performed at the intersection level, only approximate values for optimal signal timings are necessary at this second level of hierarchy. The simulator, to evaluate performance measures in *REALBAND*, is referred to as the *Approximate Prediction in Response to a Signal Network* (*APRES-NET*) model (*11*).

The flow chart for the algorithmic process for network flow control optimization is given in Figure 12. To begin the recursion it is suggested that an initial signal plan be given so that the network flow control optimization begins in the proper range. The initial plan may be obtained from an off-line method such as TRANSYT using volumes obtained from the upper level network load control module (if available) or using historical data.

To use the fast simulation model for function evaluation, the spatial region for the simulation model must be bigger than the area of control for network coordination so that the movements of all real-time platoons within the region of control can be predicted for several minutes in the future. Figure 13 (*a*) shows the region of control for the network flow control logic, and Figure 13 (*b*) shows the area simulated using the *APRES-NET* simulator.

IMPLEMENTATION ISSUES AND SOME RESULTS

Although Figure 12 shows a rather simple flow-chart, the following issues must be considered for the code to be effective in providing good if not optimal solutions and efficient for real-time applications.

- Filtering detector data for identifying platoons,
- Initialization of *REALBAND,* and
- Propagating *REALBAND* through time using *APRES-NET.*

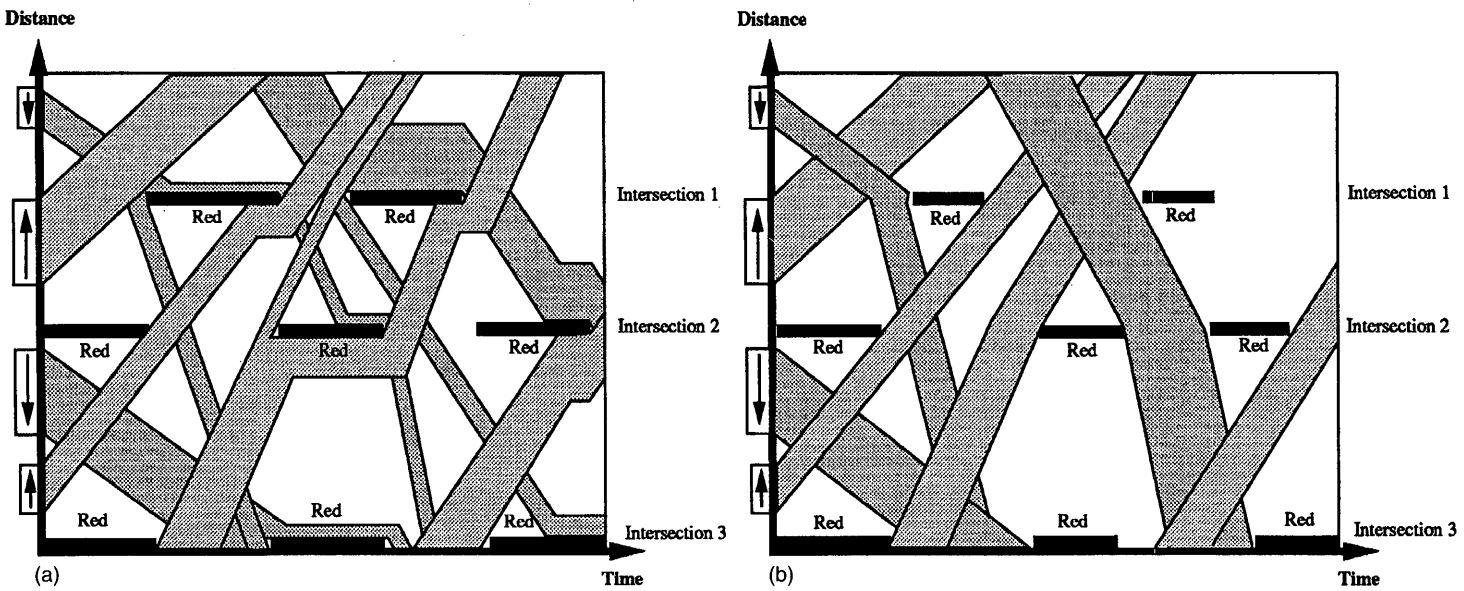


FIGURE 3 (*a*) Actual MAXBAND performance; (*b*) the REALBAND concept for single arteries.

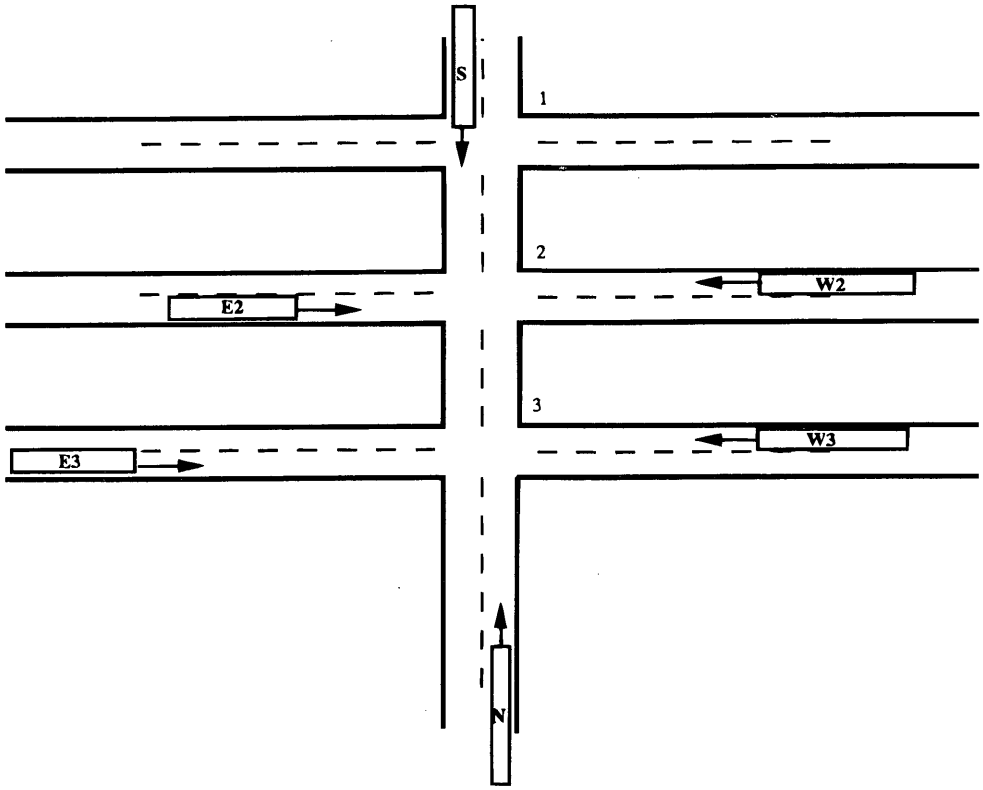


FIGURE 4 REALBAND example network.

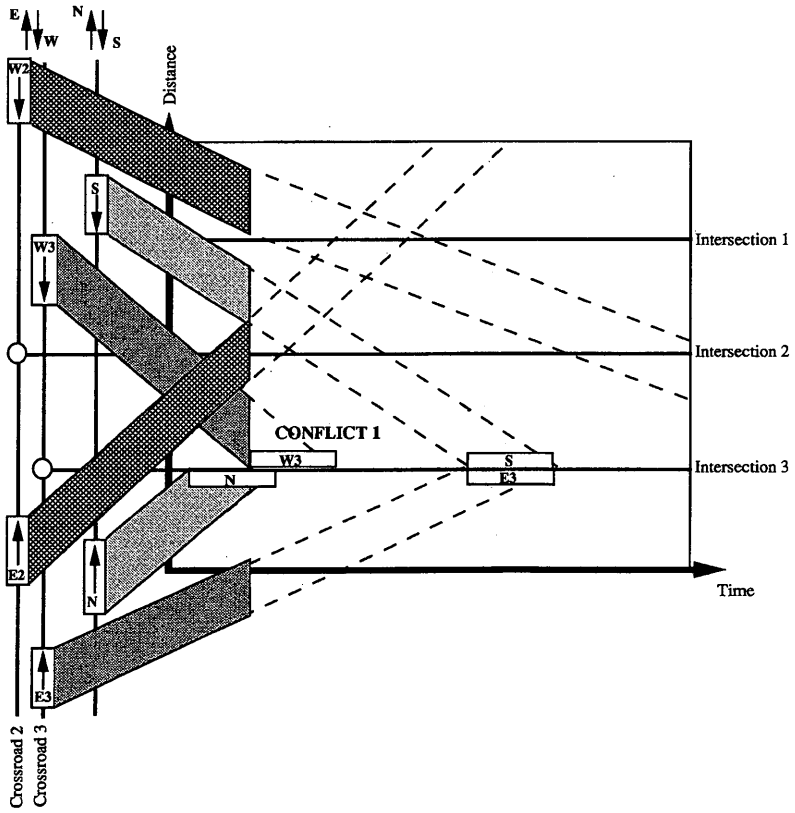


FIGURE 5 Current prediction of platoon movement.

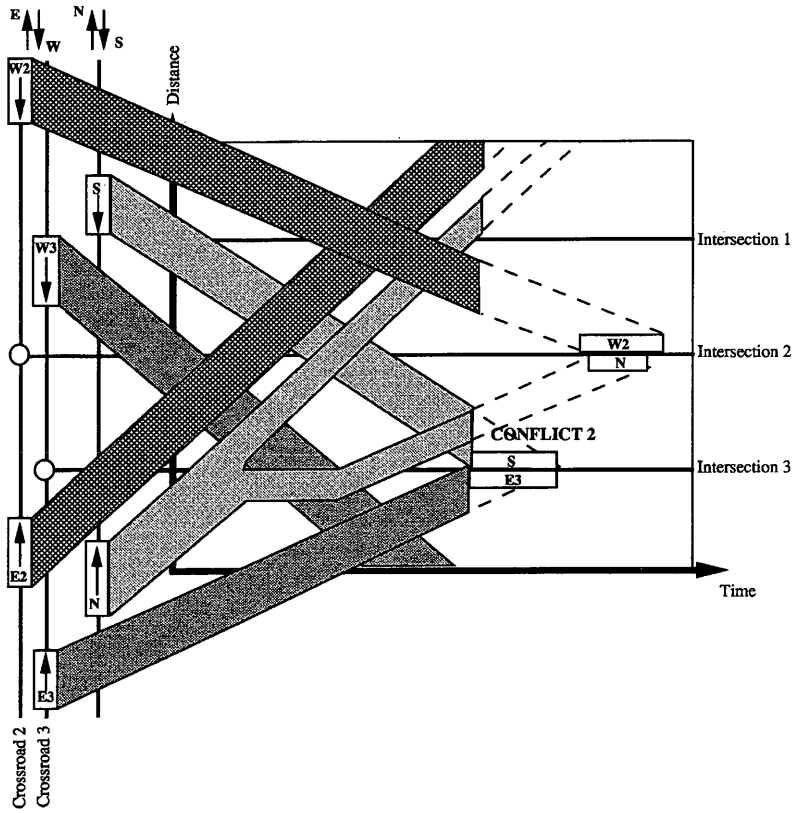


FIGURE 6 Decision to split Platoon N at Intersection 3.

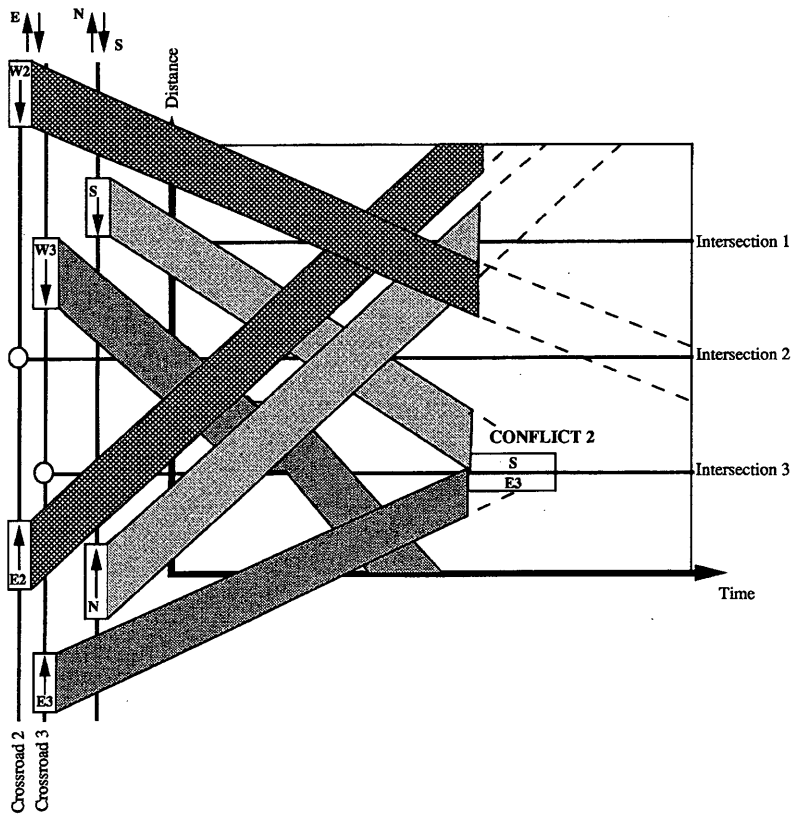


FIGURE 7 Decision to stop Platoon W3 at Intersection 3.

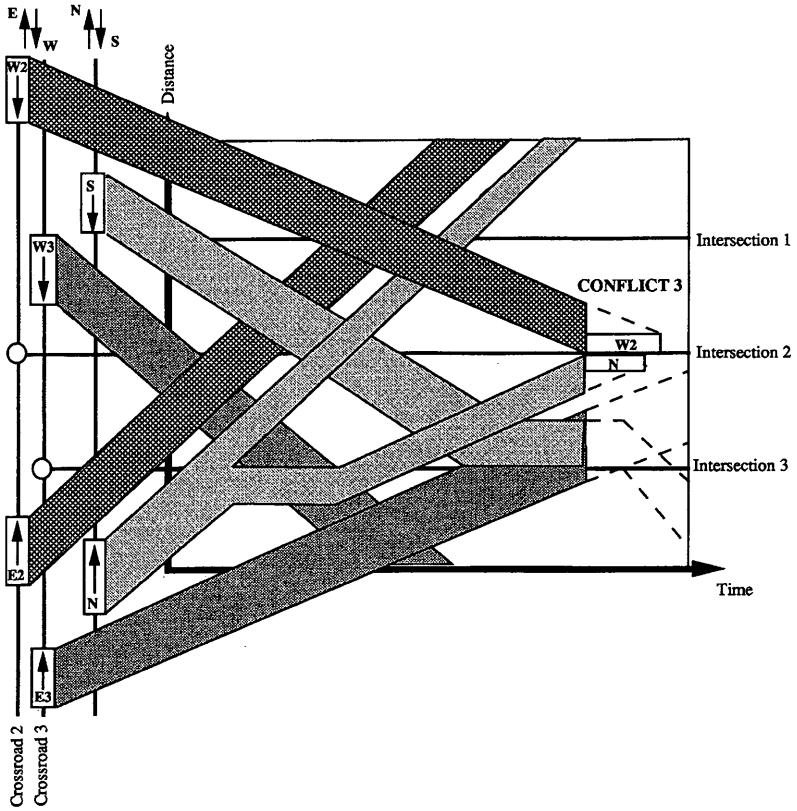


FIGURE 8 Decision to stop Platoon S at Intersection 3.

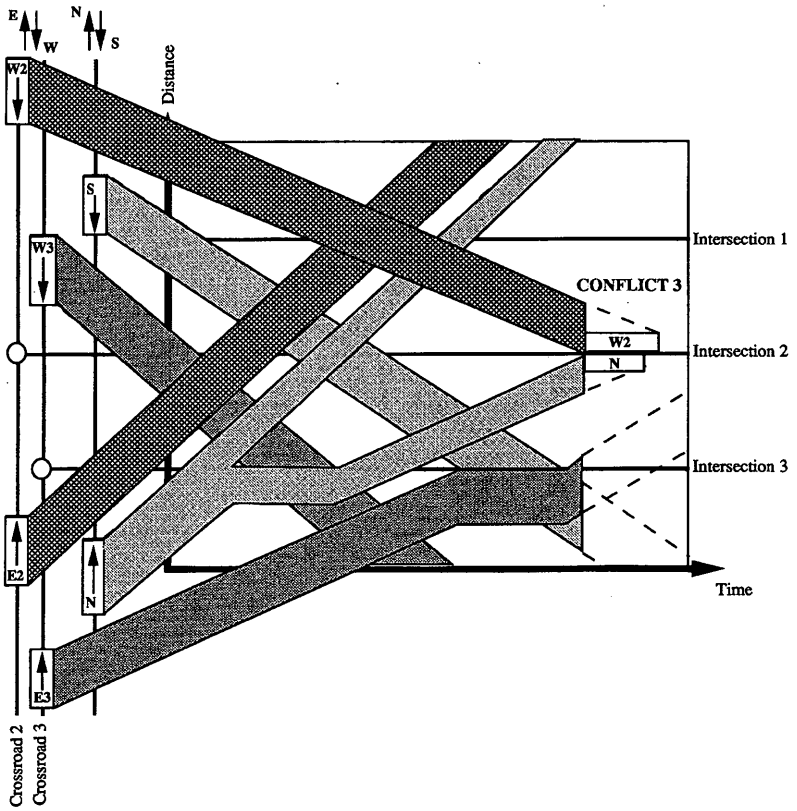


FIGURE 9 Decision to stop Platoon E3 at Intersection 3.

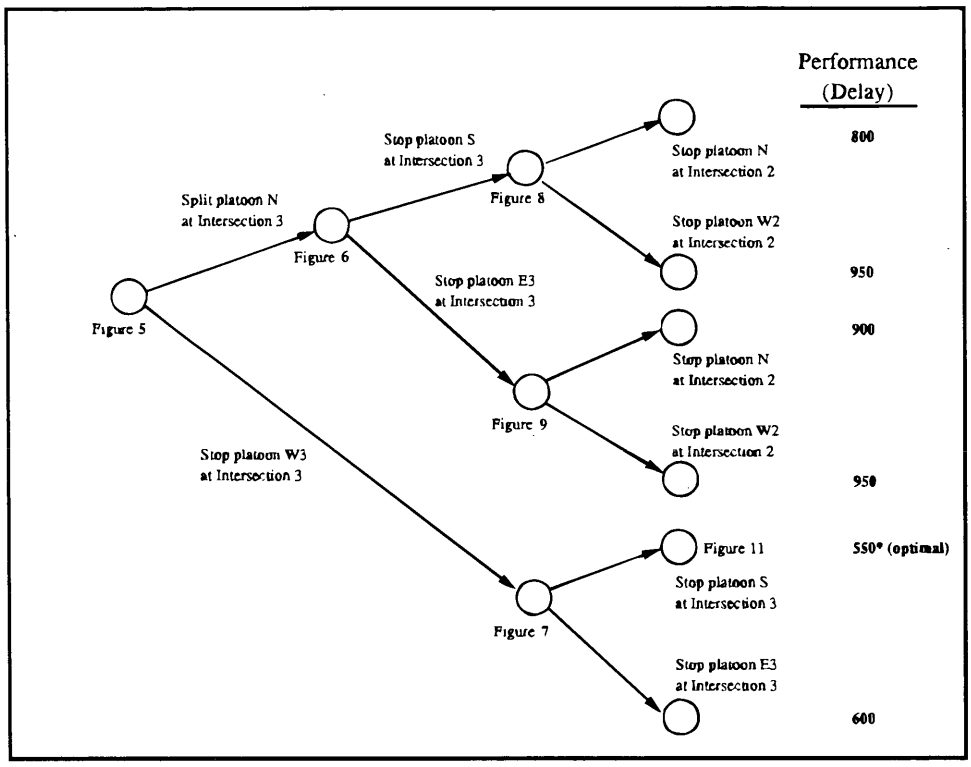


FIGURE 10 Decision tree for an illustrative problem.

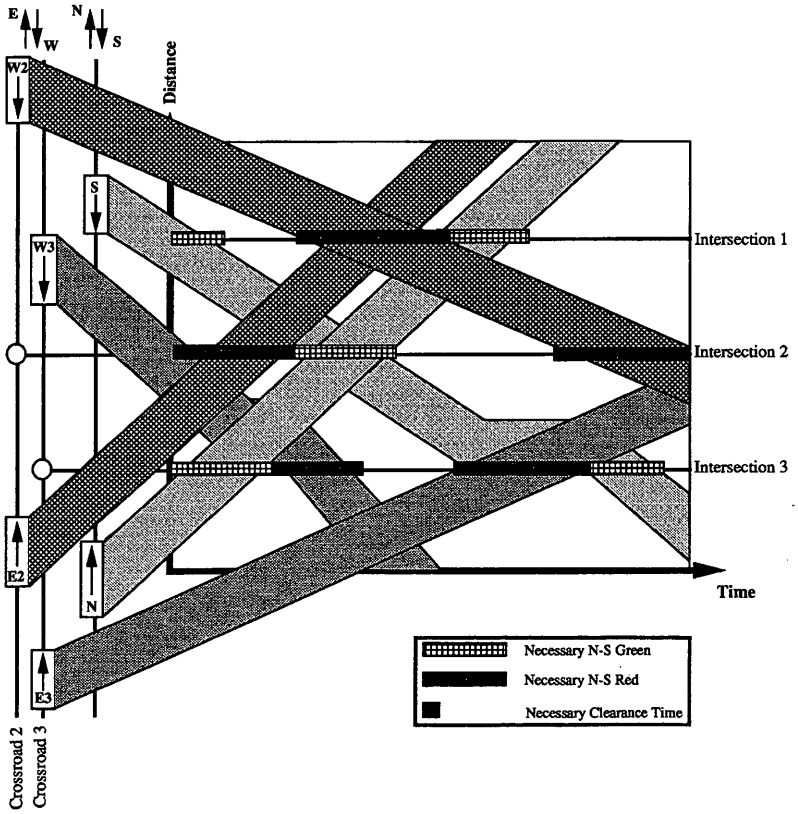


FIGURE 11 The north-south "red" and "green" phases for optimum decisions (see Figure 10).

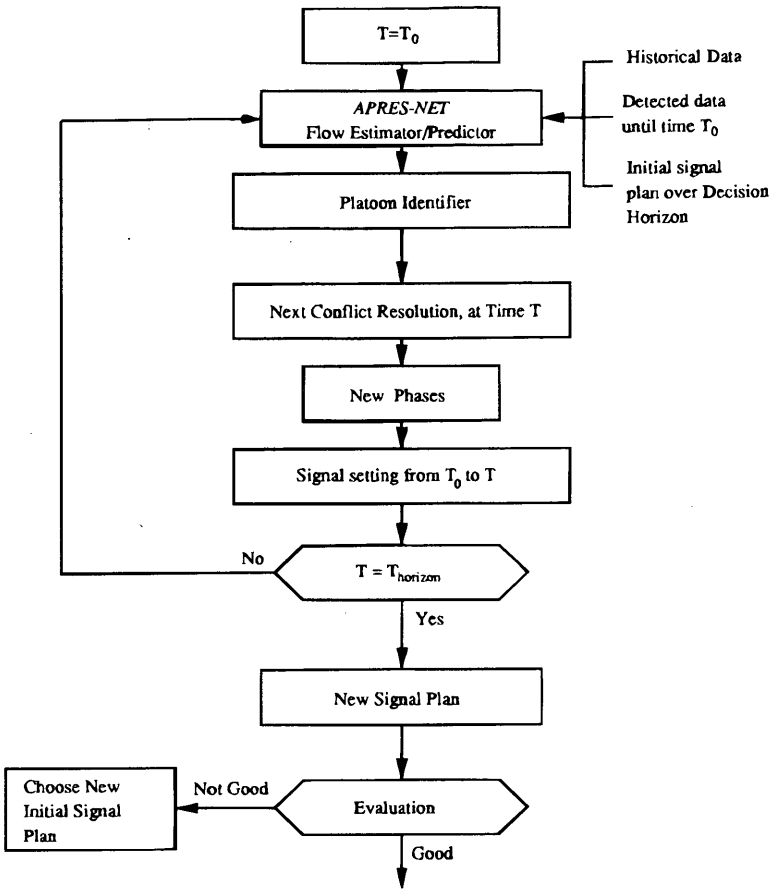


FIGURE 12 Flow chart for REALBAND.

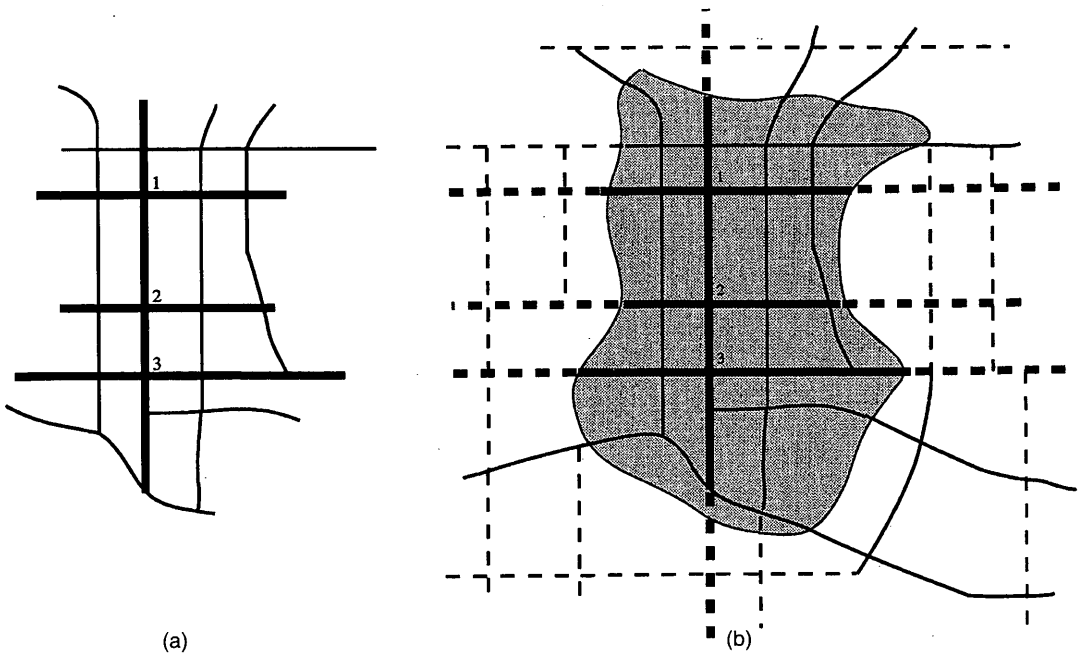


FIGURE 13 Region for (*a*) network flow control, and (*b*) simulation model.

The definition of platoons for traffic coordination depends on the level of traffic in the network. In low congestion with vehicles traveling at high speeds, a platoon may be composed of as few as three cars having an average headway of 2 secs. On a more heavily congested road, a platoon may consist of many more cars with an average headway of 1 sec. In any case, if a traffic engineer looks at vehicle detector data, that engineer can easily recognize platoons. Any platoon identification algorithm in which the goal of the algorithm is to filter out individual cars and identify platoons should be able to emulate the decisions of a traffic engineer. Several algorithms are being explored to identify platoons based on concepts of low-pass filters, threshold rules, etc. To develop the *REALBAND* algorithm, a platoon identifier was used based on two user-specified threshold parameters: maximum headway between two vehicles in the same platoon and the minimum number of vehicles that constitute a platoon. Further research and evaluation is recommended for the development of an appropriate platoon identifier.

REALBAND starts with an initial solution of phase timings and associated measure of performance obtained through *APRES-NET*. The initial phase timings could have been developed off-line using a program such as *TRANSYT* with traffic volumes obtained from the network load control level, or the initial phase timings could be given for the next few minutes. The initial phase timings define the first node on the decision tree, as shown in Figure 10. The measure of performance associated with the initial phasings becomes an upper bound (UB) on the performance because it is known that the signal network gives a feasible set of timings with at least this level of performance.

The platoon identifier then filters the detector data to identify platoons. The initial set of phasings resolves conflicts by default because traffic controllers have built-in signal phase control logic that does not permit conflicting movements at an intersection. However, examination of the platoon data from the initial run of *APRES-NET* will indicate that, at times, a platoon will be stopped or split so that another platoon can pass through a conflicting movement. *REALBAND* will identify the first time this occurs, say at time $T_0+\Delta T$. This is the first conflict to be resolved; hence, the time has been propagated by ΔT, and a new leaf node is formed in the decision tree. *APRES-NET* will then be run using the signals corresponding to having the stopped platoon pass through and the other (conflicting) platoon stopped. Then another UB is obtained, along with a new set of identified platoons (the platoons may have changed due to splitting, combination of platoons, or both). For the two sets of scenarios formed, the next conflict is identified and the process is repeated. In this manner, as *REALBAND* propagates through time, it develops a decision tree, keeping a feasible UB for the performance at each node. The algorithm terminates when *REALBAND* propagates to the planning horizon; the phase timings corresponding to the leaf node with minimum UB become the timings sent to the lower (intersection) level controllers. The algorithm can also be set to terminate when some performance threshold is satisfied with some UB (this performance threshold being given as a function of a lower bound or given a priori by the traffic engineer). Although an effective lower bound has not yet been developed for each node, the authors intend to to develop a procedure to create such lower bounds and prune the decision tree.

The authors are still conducting simulation tests for evaluating *REALBAND* performance. A 41-node, 42-link actual network (representing a section of Tucson) has been coded on *APRES-NET*. The initial phase timings have been provided by the city's traffic engineer based on TRANSYT and subsequent manual fine-tuning. In the tests, two intersections (e.g., Intersections 1 and 2) within this network were subjected to real-time control using *REALBAND*. The resulting changes in phase timings are shown in Figure 14. There is considerable difference between the initial timings and the timings downloaded by *REALBAND*. For a 200-sec planning horizon for this two-intersection problem, the total delay for the entire 41-node network decreased from 12,559 vehicle-sec to 11,275 vehicle-sec, yielding about a 10 percent improvement.

The authors are still developing an efficient mechanization for generating and pruning decision trees. Further laboratory evaluation of *REALBAND* also is planned. The network will be simulated using a micro-simulation model (TRAF-NETSIM will be used for this purpose) and will provide simulated detector data to *REALBAND* (through *APRES-NET*). *REALBAND* in turn will return to the microsimulator phase durations in real time. In this manner, it will be possible to perform considerable off-line evaluation before testing *REALBAND* in the field.

The *REALBAND* procedure for network flow control exploits the availability of real-time traffic data to control vehicular traffic through a network to optimize a given performance measure. It

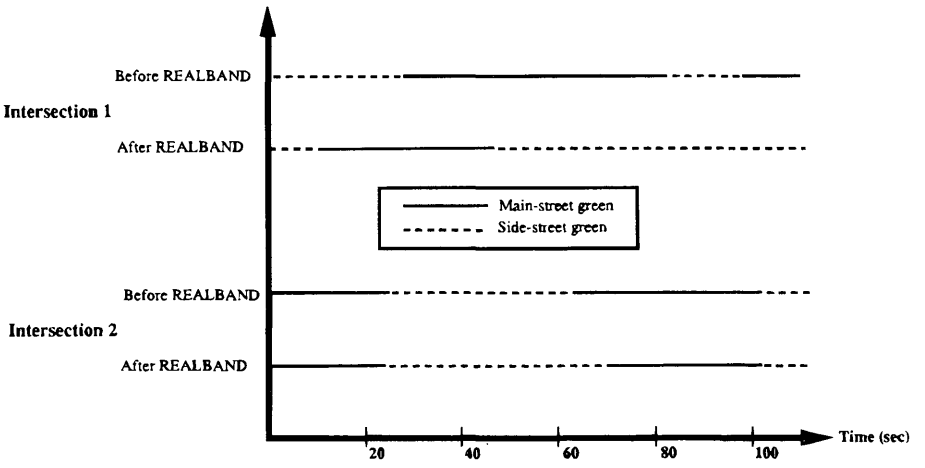


FIGURE 14 "Before" and "after" REALBAND application.

is envisioned that this procedure will be suitable for light-to-moderate traffic conditions, but not oversaturated conditions. *REALBAND* should perform as well as or better than off-line methods such as PASSER II, MAXBAND, and TRANSYT.

ACKNOWLEDGMENTS

The first author acknowledges Consiglio Nazionale delle Ricerche Progetto Finalizzato Transporti 2 for partially supporting this research; the second author acknowledges the support of Arizona Department of Transportation and the Federal Highway Administration. The authors also appreciate the insightful suggestions and helpful comments of Larry Head, Michael Whalen, Douglas Gettman, Vijitha Kaduwela, and Michael O'Brien.

REFERENCES

1. Head, K. L., P. B. Mirchandani, and D. Sheppard. A Hierarchical Framework for Real-Time Traffic Control. In *Transportation Research Record 1360*, TRB, National Research Council, Washington, D.C., 1992, pp. 82–88.
2. Wallace, C. E., K. Courage, D. R. Reaves, G. W. Schoene, and G. W. Euler. *TRANSYT-7F User's Manual.* Federal Highway Administration, U.S. Department of Transportation, 1981.
3. Little, J. D. C., M. D. Kelson, and N. H. Gartner. MAXBAND; A Program for Setting Signals on Arteries and Triangular Networks. In *Transportation Research Record 795,* TRB, National Research Council, Washington, D.C., 1981, pp. 40–46.
4. Chang, E. C., B. G. Marsden, and R. Derr. PASSER II-84 Microcomputer Environment System-Practical Signal-Timing Tool. *Journal of Transportation Engineering 113,* 1987, pp. 625–641.
5. Chang, E. C., S. L. Cohen, C. Liu, C. J. Messer, and N. A. Chaudhary. MAXBAND-86: Program for Optimizing Left-Turn Phase Sequences in Multiarterial Closed Networks. In *Transportation Research Record 1181*, TRB, National Research Council, Washington, D.C., 1988, pp. 61–67.
6. Chaudhary, N. A., and C. J. Messer. *PASSER IV, A Program for Optimizing Signal Timing in Grid Networks.* Paper No. 930825. Presented at the 72nd Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1993.
7. Gartner, N. H., S. F. Assmann, F. Lasaga, and D. L. Hou. MULTIBAND-A Variable-Bandwidth Arterial Progression Scheme. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 212–222.
8. Hunt, P. B., D. I. Robertson, R. D. Bretherton, and R. I. Winton. *SCOOT—A Traffic Responsive Method of Coordinating Signals.* Report No. LR 1014. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1981.
9. Hadi, M. A., and C. E. Wallace. *Improved Optimization Efficiency in TRANSYT-7F.* Technical Report. Transportation Research Center, University of Florida, Gainesville, for presentation at the ORSA/TIMS Meeting, Orlando, Fla., April 1992.
10. Rathi, A. K. and A. J. Santiago. The New NETSIM: TRAF-NETSIM Version 2.00 Simulation Program. Presented at the 68th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1989.
11. Dell'Olmo, P., and P. B. Mirchandani. A Model for Real-time Traffic Coordination Using Simulation Based Optimization. *Annals of Operations Research* (in press).

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Model to Evaluate the Impacts of Bus Priority on Signalized Intersections

SRINIVASA R. SUNKARI, PHILLIP S. BEASLEY, THOMAS URBANIK II, AND DANIEL B. FAMBRO

Transit service is being viewed increasingly as a reliable travel demand measure. Various means of attracting the motorist to move from the car to transit are being attempted all over the country. It has been found that delay at intersections is the primary cause of bus delay. Reducing delay at intersections can reduce overall trip time, improve schedule reliability, and reduce overall congestion. Providing priority for buses at signalized intersections is one way to reduce delay at intersections. Numerous organizations have developed priority strategies to do the same. But most of them cannot be operational for a long time for various reasons. Improved technology has prompted traffic engineers to renew efforts to develop newer bus priority strategies. This paper discusses the development of a model to evaluate the impacts of implementing a priority strategy at signalized intersections. The model uses the delay equation for signalized intersections in the *1985 Highway Capacity Manual*. A priority strategy was developed and implemented in the field. Data were collected, and delay in the field was measured. The model seems to be predicting delay reasonably accurately. In some cases, however, the model was overestimating delay. The model can be a useful tool to traffic engineers to evaluate the impacts and the feasibility of implementing a priority strategy.

The concept of providing priority to buses at traffic signals is by no means a newly conceived notion. In fact, as early as 1962 an experiment was conducted in Washington, D.C. in which the offsets of a signalized network were adjusted to match better the lower average speed of buses (*1*). The first bus-actuated, or active, signal priority experiment for buses occurred in Los Angeles in 1970 and soon was followed by other similar demonstrations across the United States (*2*). These early experiments concentrated on moving buses through an intersection as quickly as possible with little or no concern for other traffic. In general, experimentation with bus priority has yielded positive results for buses and traffic on the bus street (*3–12*). However, priority may increase delay to traffic on the cross-street. Since the concept emerged, however, experimentation and research have produced few operational systems.

RESEARCH OBJECTIVE

The primary objective of this research was to develop a model to evaluate a conditional, active bus priority strategy for a traffic signal in a coordinated signal system. An effective priority strategy would provide significant benefits to buses in terms of reduced travel time, delay incurred at an intersection, and increase in schedule reliability. The strategy should not disrupt the progression along the arterial and should not affect the cross-street operation seriously. The model was evaluated by implementing the strategy at a local intersection and observing the impacts of the strategy.

Texas Transportation Institute, College Station, Tex. 77843.

Priority at Traffic Signals

Signal priority is a method of providing preferential treatment to buses at traffic signals by altering the signal timing plan in a way that benefits buses. Buses may begin a movement within a vehicle platoon, but loading and unloading requirements may cause them to fall behind the platoon and become delayed at downstream traffic signals (*13*).

Delay at traffic signals is one of the largest components of bus delay on arterial streets. Bus delay at traffic signals comprises between 10 and 20 percent of overall bus trip times and nearly 50 percent of the delay experienced by a bus (*2*). Thus, by giving priority to buses at traffic signals, bus delay can be reduced. Potential short-term advantages of bus priority also include the decrease in bus travel times and increased speeds, decrease in schedule variability, and the improvement of non-bus traffic on the bus phase. At reasonable demand levels, bus priority can make transit a more attractive mode of transportation and may increase the passenger-carrying capacity of arterial streets (*14*). Signal priority treatments can be categorized as follows.

Passive Priority

In passive priority, predetermined timing plans are used to provide some benefit to the transit movements but do not require the presence of the transit vehicle to be active. The following passive priority treatments are low-cost methods aimed at improving transit operations (*15*).

Adjustment of Cycle Length

Reducing cycle lengths can provide benefits to transit vehicles by reducing the delay.

Splitting Phases

Splitting a priority phase movement into multiple phases and repeating it within a cycle can reduce transit delays without necessarily reducing the cycle length.

Areawide Timing Plans

Areawide timing plans provide priority treatment to buses through preferential progression, which can be accomplished simply by

designing the signal offsets in a coordinated signal system using bus travel times.

Metering Vehicles

Buses benefit from metering by allowing buses to bypass metered signals with special reserved bus lanes, special signal phases, or by rerouting buses to nonmetered signals.

Active Priority

In active priority treatments, priority is given only when the bus is actually present. There are mainly four types of active priority treatments.

Phase Extension

A phase extension is useful when the bus will arrive at the intersection just after the end of the normal green period and is usually limited to a maximum value (*13*).

Early Start

An early start priority is used when the bus arrives at the intersection during a red indication by truncating all non-bus phases (*13*).

Special Phase

A special phase occurs when a short green phase is injected into the normal phase sequence while all other phases are stopped (*13,16*).

Phase Suppression

To facilitate the provision of the priority bus phase, one or more nonpriority phases with low demand may be omitted from the normal phase sequence (*16*).

The four previous strategies are the most widely used forms of active priority. These strategies can be used alone or can be combined to provide priority to buses. Priority schemes sometimes also include the concept of compensation (*16*). In compensation, the nonpriority movements can be allocated additional green time in the form of a nonpriority phase extension after a priority to minimize deterioration of nonpriority phases.

Unconditional Priority

In unconditional signal priority (or preemption), priority is given whenever the bus detector places a call to the signal controller. After the bus is detected, the bus movement is given a green indication after all other vehicular and pedestrian clearance intervals are satisfied for safety reasons. Because unconditional priority is so disruptive to cross-street traffic, it is used mainly for emergency vehicle preemption of traffic signals only.

Conditional Priority

Conditional signal priority strategies attempt to limit the undesirable effects caused by unconditional priority through selective consideration of various factors. These factors include schedule adherence, bus occupancy, cross-street (or non-bus street) queue length, current traffic conditions, time since last priority, effect on coordination, and point in cycle at which the bus is detected.

DEVELOPMENT OF A PRIORITY MODEL

A model to simulate, evaluate, and estimate the effects of the priority scheme on intersection operation was developed (*17*). Priority is provided by phase extensions and early start of the priority phase at regular intervals. The development of the model is described in the following paragraphs.

It can be assumed safely that when a priority is granted to a bus on the coordinated approach, the result will be a decrease in delay to the bus and the vehicles on the coordinated approach. Similarly, because green time is taken from the cross-street, an increase in delay to the vehicles on the cross-street approaches is expected. These effects can be examined quantitatively using the input-output models shown in Figure 1.

Five cases have been defined and illustrated in Figure 1. Case 1 does not provide any priority. In Case 2, the priority phase gets a minimum extension to allow the bus to go through the intersection. Case 2 is most beneficial because the nonpriority phases are disrupted least and the bus would have had maximum waiting time to go through the intersection if no priority were provided.

Case 3 provides maximum extension (predefined) to the priority phase. A bus detected just before the arterial phase (priority phase) terminates can go through the intersection if the phase is extended by the travel time from the detection zone to the intersection. A 10-sec maximum extension was used in the model.

In Case 4, a minimum early start is provided. When a bus arrives on red very late in the cycle, a minimum phase time is provided for the nonpriority phase(s) on at that time, and the priority phase comes on early. The nonpriority phases are not affected seriously. Case 5 illustrates a maximum early start for the priority phase. When a bus arrives just after the termination of the arterial phase (priority phase), all of the nonpriority phases are provided minimum times and the priority phase comes on early.

Figure 1 illustrates the arrivals and departures for both the main street and the cross-street and the effects of priority phase extensions and early starts on delay. Extending the main street phase to accommodate the bus should cause a reduction in delay (reduction in size of triangle) for the vehicles on this approach. The length of the extension affects the amount by which delay is reduced. The effects on the cross-street are similar but opposite. A short extension likely will cause a small increase in delay (increase in size of triangle), whereas a large extension should cause a larger increase in delay. An early start priority affects delay similarly to an extension, as illustrated in Figure 1.

Analytical Tool to Evaluate Priority Scheme

The simple model developed uses the delay equation found in the *1985 Highway Capacity Manual* (HCM). Geometric, traffic, and signal timing values as required for the HCM model are obtained

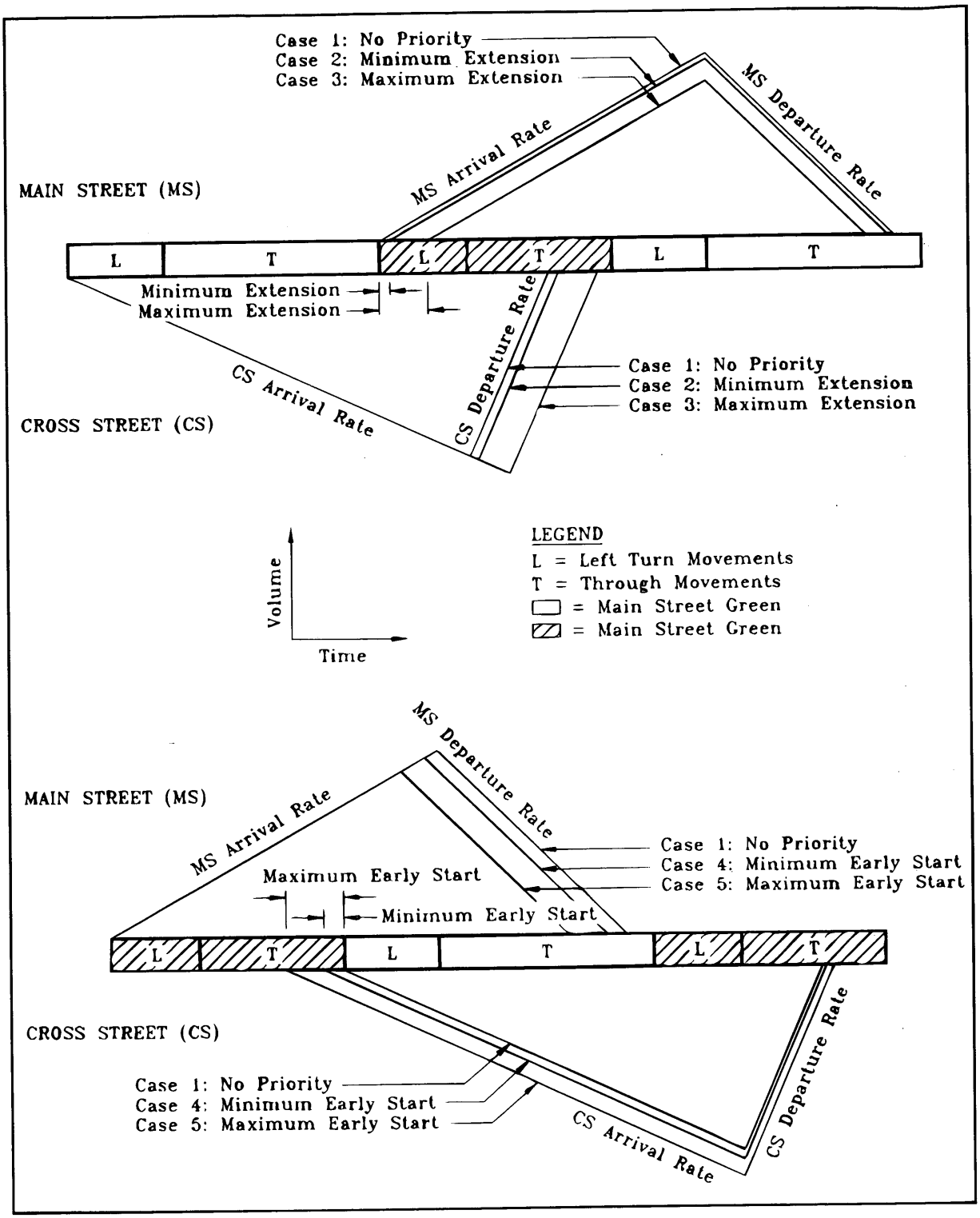


FIGURE 1 Illustration of the expected effects of the priority scheme on main and cross-street traffic.

from either the plans or the field observations. Different types of priority are modeled by adjusting the green times to represent the desired condition (no priority, phase extension, or early start). For example, to model the intersection without priority, the green times used in the spreadsheet would match average green times in the field. Similarly, for an extension or early start the cross-street green times would be decreased (and the coordinated phase green time increased) according to the type of priority and the length of the priority phase.

The HCM delay equation calculates the average seconds of delay per vehicle. These units are adequate for many applications. The increase in delay is experienced by a large number of vehicles with

small occupancies (passenger cars). But the benefits are presumably experienced by a large number of passengers in the bus. Thus, person delay is a more appropriate measure of effectiveness. Also, to compare the benefits gained by buses to the increase in delay to the cross-street, the effects are compared on a cycle-by-cycle basis. The HCM delay value is converted to person-sec of delay per cycle knowing the number of vehicles per cycle and the average automobile occupancy.

The magnitude of delay savings to the bus depends on the time at which it arrives at the intersection or is detected by a priority detector. If it arrives during the green portion and can pass safely through the intersection without an extension, then there is no delay savings to the bus. However, if the bus arrives at the intersection such that it can be accommodated by an extension, then the bus is saved an amount of time equal to the length of the cross-street period. If the bus arrives during the cross-street period, the delay savings to the bus increases the earlier it is detected in the phase.

Based on the green splits, the model estimates the period the bus has to wait when no priority is provided as well as when priority is provided. This is done for the buses arriving at different points in the cycle. For simplicity, an assumption is made that buses are arriving only on Phase 2 approach and priority is being provided only to the coordinate phases (Phases 2 and 6).

The model calculates the savings obtained by providing priority through a number of steps. Various terms used in the spreadsheet that is used in the model are defined and described below.

Person Delay/Cycle with No Bus

Person delay/cycle with no bus has been defined for two cases in the spreadsheet.

Person Delay/Cycle with No Bus for Original Splits ($D_{(NB\text{-}OS)}$)

$D_{(NB\text{-}OS)}$ is obtained by simply converting vehicle-stopped delay without modifying the green splits in sec/vehicle to sec/cycle and multiplying by the average auto occupancy. The average auto occupancy is assumed to be 1.25 and does not consider any bus arriving in that particular cycle.

Person Delay/Cycle with No Bus for Modified Splits ($D_{(NB\text{-}MS)}$)

$D_{(NB\text{-}MS)}$ is the same as $D_{(NB\text{-}OS)}$ except that the splits used to calculate stopped delay are modified as required for providing priority to bus.

Waiting Period

Waiting period is the period the bus has to wait at the intersection. It depends on the point in cycle at which the bus arrives and also whether priority is provided.

Person Delay/Cycle with One Bus and No Priority ($D_{(1B\text{-}NP)}$)

$D_{(1B\text{-}NP)}$ is obtained for each case of priority. $D_{(1B\text{-}NP)}$ is the same as $D_{(NB\text{-}OS)}$ for all phases except the Bus Phase (Phase 2), where a bus is assumed to arrive in the cycle. The delay for the bus phase is obtained by summing the bus phase delay in $D_{(NB\text{-}OS)}$ with the product of the bus occupancy (*40*) and the period for which the bus has to wait.

Person Delay/Cycle with One Bus and With Priority ($D_{(1B\text{-}P)}$)

$D_{(1B\text{-}P)}$ is obtained for each case of priority. Various scenarios are obtained by modifying the green splits (adding to priority phases and reducing from nonpriority phases). First, the delay with modified splits is calculated assuming there is no bus arrival in the cycle ($D_{(NB\text{-}MS)}$). The waiting period for the bus when priority is provided as well as when no priority is provided was calculated earlier. The delay values in $D_{(1B\text{-}P)}$ are the same as in $D_{(NB\text{-}MS)}$ except the bus phase. The delay for the bus phase is obtained by summing the delay in $D_{(NB\text{-}MS)}$ with the product of the bus occupancy (*40*) and the waiting period for the bus when priority is provided.

The delay values in $D_{(1B\text{-}NP)}$ and $D_{(1B\text{-}P)}$ are obtained assuming that a bus is arriving every cycle. However, a bus is arriving only once every 4 or 5 cycles or priority for the bus is being provided every 4 to 5 cycles. Providing priority in quick succession is harmful for two reasons. First, the controller may lose coordination and arterial progression is disrupted; second, the delays for some cross-street phases may get very high. A few cycles without priority will allow any phases disrupted caused by providing priority to recover. The cycles in which the bus is arriving (priority is being provided) can be called *bus arrival cycles.* Thus, there are only a limited number of *bus arrival cycles* in an hour, and their number depends on the cycle length of the intersection and the extent to which the cross-street can be disrupted.

Weighted Normal Delay ($W.D_{(NP)}$)

$W.D_{(NP)}$ is the delay experienced in person sec/cycle for an hour, in which buses are arriving every 4 or 5 cycles and no priority is provided. $W.D_{(NP)}$ is obtained by summing the product of the delays in $D_{(1B\text{-}NP)}$ with the number of bus arrival cycles and the product of the delays in $D_{(NB\text{-}OS)}$ with the number of normal cycles (non-bus arrival cycles) and dividing the sum by the total number of cycles in an hour:

$$W.D_{NP} = \frac{(D_{(1B\text{-}NP)} * \textit{No. Bus Arr. Cyc.}) + (D_{(NB\text{-}OS)} * \textit{No. Non-bus Arr. Cyc.})}{\textit{No. Cyc./hr}}$$

Weighted Delay with Priority ($W.D_{(P)}$)

$W.D_{(P)}$ is the delay experienced in person sec/cycle for an hour, in which buses are arriving every 4 or 5 cycles and the appropriate priority is provided. $W.D_{(P)}$ is obtained by summing the product of the delays in $D_{(1B\text{-}P)}$ with the number of bus arrival cycles and the product of the delays in $D_{(NB\text{-}OS)}$ with the number of normal cycles (non-bus arrival cycles) and dividing the sum by the total number of cycles in an hour:

$$W.D_{P} = \frac{(D_{(1B\text{-}P)} * \textit{No. Bus Arr. Cyc.}) + (D_{(NB\text{-}OS)} * \textit{No. Non-bus Arr. Cyc.})}{\textit{No. Cyc./hr}}$$

FIELD EVALUATION OF THE MODEL

The objective of the field evaluation was to investigate the reliability of the results predicted by the model and document any benefits to the bus and possible detriments to other traffic caused by the priority strategy. Stopped delay was the chosen measure of performance because intersection delay studies are very common and it is relatively easy to collect; it is also precise.

The HCM field delay measurement technique was used to collect stopped delay data for each of the approaches to the intersection for each of the cases. The number of seconds of delay per vehicle can then be calculated according to the following equation (*18*):

$$Delay = \frac{\sum V_s \times I}{V},$$

where

$Delay$ = stopped delay, in sec/vehicle,
ΣV_s = sum of stopped vehicle counts,
I = length of interval (sec), and
V = total volume observed during study period.

Site Selection

The following criteria were considered in the decision to choose the site:

- the intersection must be signalized;
- the intersection must be part of a coordinated system;
- intersection geometrics and signal phasing should be relatively simple such that priority is feasible;
- the site allows for the collection of the necessary data; and
- the intersection is not critical such that priority would disrupt traffic operations to a great degree.

Careful consideration of the above criteria resulted in the selection of the intersection of Texas Avenue and Southwest Parkway in College Station, Texas. The site is located at the intersection of a major north-south arterial (Texas Avenue) and a major east-west collector (Southwest Parkway). The intersection is controlled by an EPAC 300-actuated controller unit manufactured by Automatic Signal/Eagle Signal.

Data Collection

It was decided to simulate bus operation in the field for different types of bus arrivals on the southbound approach of Texas Avenue. Stopped vehicle counts were recorded for Case 1, Case 3, and Case 5 of the five conditions described earlier. Case 2 and Case 4 did not warrant a separate study, because the effect of providing priority on the nonpriority phases was not significant.

The data collected for each of the cases were reduced and input into the model. The green splits were obtained from the data downloaded from the controller, and their average values were used for each case. Field studies were used to calibrate the model to the local conditions. The progression factors specified in HCM were incorporated. Various factors defined in HCM were used to calculate the saturation flow rate. However, it should be noted that calibration may not result in very similar values of delays from the model and field studies. The HCM delay equation is suitable for fixed time operation. It is based on a number of empirical factors that may not apply accurately to the existing local conditions. Also, although every effort was made to maintain consistency and accuracy in the field data collection, there could be some minor errors. Hence, it is necessary to recognize that the model results may not match completely the field results.

Field Data Collection

NEMA phase designation (Figure 2) with phases 2 and 6 as coordinated phases was used to denote phases. Data collectors were positioned on each approach to record the number of vehicles stopped at 15-sec intervals. In Case 1 (No Priority) the stopped vehicle counts were recorded for 30 min.

In Cases 3 and 5 the stopped vehicle counts were required only during cycles in which the priority scheme was activated. Buses do not operate along Texas Avenue. Therefore, a push button was activated manually to simulate the arrival of the bus for various cases. To minimize the disruption to non-bus traffic, the intersection was allowed to recover between successive activations of the scheme.

The intersection was videotaped using two video cameras to determine the traffic volumes. Data were also collected from the traffic signal controller via a laptop computer. This information included the status of each detector and the current signal phase every 1/10 of a second. The phase status data were used to obtain green splits during the study.

The data collection was performed independently for Cases 3 and 5 on separate days. For each day a similar type of bus arrival was simulated. For Case 3 (maximum extension) the point in the cycle at which the coordinated phase would terminate (i.e., if no priority was to be provided) was determined. The push button was energized a few seconds before that point in the cycle and held for about 10 seconds after the point. The coordinated phase would be extended as long as the push button was held. The duration by which the coordinated phase was extended was reduced proportionately from the subsequent noncoordinated phases.

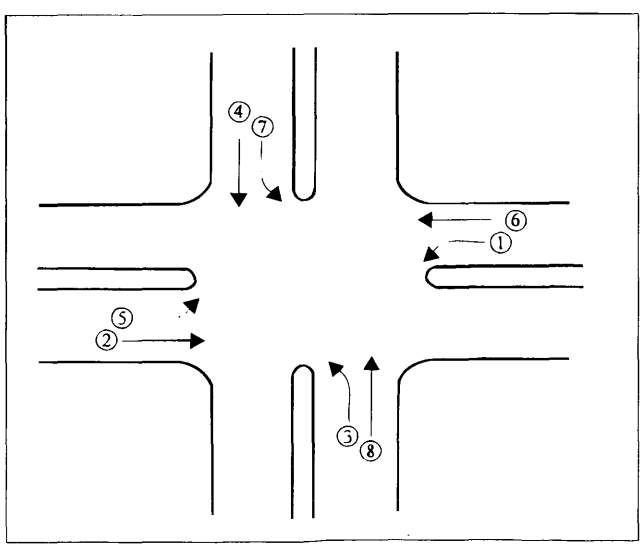


FIGURE 2 NEMA configuration for numbering phase movements.

For Case 5 (maximum early start) it was decided to provide a maximum of 7 sec for Phases 3 and 7, a maximum of 15 sec for Phases 4 and 8, and maximum of 5 seconds for Phases 1 and 5. The push button was energized 7 sec after the coordinated phase terminated and Phases 3 and 7 came on. This actuation forces Phases 3 and 7 to the subsequent phases. The push button was energized in a similar fashion to force out of other nonpriority phases after providing the earlier specified green times.

About 10 sample priority cycles were obtained for each of the two cases. Stopped delay data were collected for 4 intervals of 20 min each. Because the cycle length of the intersection was 115 sec, each 20-min period facilitated in getting two to three cycles in which priority was provided. Hence, the target of 10 priority cycles was achieved.

Data Reduction

The data collected (green splits, volumes, and stopped delay data) were reduced for each case separately to obtain stopped delay on a cycle-by-cycle basis. This was done to maintain uniformity in data reduction with the other cases. As mentioned earlier, in Cases 3 and 5 only the cycles in which priority was provided were considered.

Averages of the green splits (for each cycle) were input into the model. Volumes were obtained from the video tapes. These volumes were input into the model and used to estimate delays in the field. The same procedure was used for all of the cases (Cases 1, 3, and 5).

Stopped delay values obtained in the field were then compared with the delay values obtained from the model.

COMPARING FIELD AND MODEL RESULTS

Field data were reduced as described earlier. The average volumes and splits were input into the spreadsheet model. Intersection stopped delay observed in the field and predicted by the model were computed and compared by approach as well as for the entire intersection. Table 1 illustrates the comparison of these delays.

Data in Table 1 indicate that the delay predicted by the model is slightly higher than the delay observed in the field. The difference in delays is more apparent at higher vehicle-to-cycle (v/c) ratios ($v/c >$ 0.85). This indicates that the model is good at predicting delays with low v/c ratios and as v/c ratios increase, the model overestimates the delay values. This finding is consistent with the belief that the delay equation in the HCM overestimates the delay at high v/c ratios.

TABLE 1 Comparison of Field Delay with Model Delay

Cases	Approach	v/c Ratio	Field Delay (sec/veh)	Model Delay (sec/veh)
Case 1	N. Bound	0.39	11.7	12.0
	S. Bound	0.43	6.8	12.0
	E. Bound	0.73	21.7	41.8
	W. Bound	0.71	35.5	46.9
	Total Intersection Delay		19.1	21.1
Case 3	N. Bound	0.37	13.1	11.0
	S. Bound	0.49	4.4	11.0
	E. Bound	0.97	39.9	77.7
	W. Bound	0.79	39.1	55.4
	Total Intersection Delay		18.1	26.0
Case 5	N. Bound	0.39	9.9	10.4
	S. Bound	0.42	4.9	10.9
	E. Bound	0.86	55.2	57.8
	W. Bound	0.85	40.6	69.7
	Total Intersection Delay		19.2	20.2

The approaches with v/c ratios > 0.85 were removed (refer to Table 1), and a regression analysis was performed with the field delay values as independent values and the model values as dependent values. The result of the regression analysis follows:

$$R^2 = 0.904$$
$$\text{Intercept} = 0$$
$$X\text{-coefficient} = 1.413$$

The desirable values for R^2 and X-coefficient are 1. Although an R^2 of 1 indicates that there is a strong linear relationship between the field delay and model delay, an X-coefficient of 1 indicates that model delay is equal to field delay.

The regression analysis indicates that there is a strong linear relationship between the delay predicted by the model and the delay observed in the field. However, the model is overestimating delay by about 41 percent.

To investigate the overestimation of the delay by the model, the data were reduced further to obtain stop delay for each phase for all the three cases. Table 2 illustrates the delay experienced by each phase along with their v/c ratios. It is seen that although the delay predicted by the model is slightly higher than the delay observed in the field for most of the phases, the difference is more apparent for phases with high v/c ratios and for left-turn phases. The difference can be attributed to two reasons. First, the HCM delay equation used in the model overestimates delay at high v/c ratios. Second, delay observed in the field for left-turn phases (mainly Phases 1 and 5) is higher than delay predicted by the model. This is because left-

TABLE 2 Comparison of Field Delay with Model Delay for Each Phase

Case	Phase	Volume to Capacity Ratio	Field Delay (sec/veh)	Model Delay (sec/veh)
Case 1	1	0.26	20.7	41.5
	6	0.42	10.8	8.8
	5	0.26	16.5	41.0
	2	0.47	5.70	8.9
	3	0.78	61.0	48.8
	8	0.71	37.3	38.0
	7	0.59	37.8	42.5
	4	0.78	35.2	48.7
Case 3	1	0.36	24.7	41.1
	6	0.43	11.8	7.2
	5	0.24	15.3	42.2
	2	0.53	3.40	8.6
	3	1.07	44.9	115.5
	8	0.92	37.0	55.9
	7	0.71	43.5	49.2
	4	0.84	37.2	57.9
Case 5	1	0.29	16.1	41.2
	6	0.41	9.20	6.7
	5	0.31	18.4	41.7
	2	0.44	3.20	7.0
	3	0.92	80.4	73.5
	8	0.84	41.8	49.4
	7	0.70	56.1	46.4
	4	0.93	34.7	77.9

turning vehicles have protected-permitted operation and, thus, have the long duration of arterial through movements to make left turns. The model does not estimate delay very well for left turns having protected-permitted operation.

To examine the model under less complicated conditions, delays for left-turn phases and phases with high *v/c* ratios were removed from the data in Table 2. A regression analysis performed on the remaining data gave the following results:

$$R^2 = 0.904$$
$$\text{Intercept} = 0$$
$$X\text{-coefficient} = 1.413$$

Results of the regression analysis indicate that although the model is overestimating delay by about 25 percent, there is a strong linear relationship between the field delay and model delay. Although the delay estimation is very good at lower values, the model is overestimating the delay at higher values. The delay for the arterial phases are the low values and are being predicted very well. However, delay for the cross-street phases may be estimated by using the *X*-coefficient as a reduction factor. Although it is recognized that using only 10 observations to perform a regression analysis may not be ideal, lack of more data did not allow a more thorough analysis.

CONCLUSIONS

Based on data collected and reduced and analysis performed with the model, it can be said that a model has been developed to evaluate the effect of a bus priority strategy on the intersection operations. The model is very simple to use and estimates the effects of bus priority at an intersection reasonably accurately. The model seems to overestimate delay for some phases. Overestimation of delay, however, will only present a picture that is worse than what actually is in the field, that is, the delay experienced by the critical phases is less than the delay predicted by the model. Hence, even if the model predicts that the implementation of a priority strategy may worsen significantly the intersection operation, it may not be the case. The results of the model should be looked at closely, and engineering judgment should be used to evaluate the feasibility of any priority strategy.

ACKNOWLEDGMENTS

The authors thank the Southwest Region University Transportation Center for sponsoring this project. Acknowledgments are also due to the engineers and technicians from Eagle Signal and the City of College Station for assisting in conducting this research.

REFERENCES

1. Sperry Rand Corporation. *Urban Traffic Control and Bus Priority System Design and Installation.* Sperry Systems Management Division, November 1972.
2. Evans, H. and G. Skiles. Improving Public Transit Through Bus Preemption of Traffic Signals. *Traffic Quarterly.* Vol. 24, No. 4, October 1970, pp. 531–543.
3. Ludwick, J. S., Jr. *Simulation of an Unconditional Preemption Bus Priority System.* Report MTP-400, The Mitre Corporation, December 1974.
4. Labell, L. N., C. P. Schweiger, and M. Kihl. *Advanced Public Transportation Systems: The State of the Art. Update '92.* U.S. Department of Transportation, Federal Transit Administration, Washington, D.C., April 1992.
5. Elias, W. J. *The Greenback Experiment. Signal Preemption for Express Buses: A Demonstration Project.* Caltrans, Sacramento County, Calif., April 1976.
6. Wattleworth, J. A., K. G. Courage, and C. E. Wallace. *Evaluation of Some Bus Priority Strategies on NW 7th Avenue in Miami.* In *Transportation Research Record 626,* National Research Council, Washington, D.C., 1977, pp 32–35.
7. *Evaluation of the Bus Priority Traffic Signal Preemption Demonstration Project.* Ventura Bus Priority Traffic Signal Pre-emption Demonstration Project, City of Los Angeles Department of Transportation, Los Angeles, Calif., January, 1990.
8. Finger, W. B. *Express Bus Preemption That Can Work With Signal System Progression.* Charlotte, North Carolina Express Bus Preemption System, July 1992.
9. New Type of Bus Pre-Emption System Tested. *Urban Transportation Monitor,* Vol. 7, No. 11, June 1993.
10. Press Release. Kitsap Transit, June 25, 1992.
11. *Bus Priority: Traffic Signal Preemption.* A Report by the Transit/Highways Task Force, Chicago Area Transportation Study, Chicago, Ill., December 1989.
12. Press Release. *The Maryland Department of Transportation Announces New Bus Pre-Emption System for MD 2 in Anne Arundel County.* May 11, 1993.
13. Taube, R. N. *Bus Actuated Signal Preemption Systems: A Planning Methodology.* Report UMTA-WI-11-0003-77-1, University of Wisconsin-Milwaukee Center for Urban Transportation Studies, Milwaukee, Wis., May 1976.
14. Paine, K. L. *Bus Priority Signal Systems.* 3M Company, January 1977.
15. Urbanik II, T. *Evaluation of Priority Techniques for High Occupancy Vehicles on Arterial Streets.* Research Report 205-5. Texas Transportation Institute, Texas State Department of Highways and Public Transportation, July 1977.
16. Bernhard, H. *Bus Priority: A Focus on the City of Melbourne.* Civil Engineering Working Paper, Monash University, Clayton, Victoria, Australia, August 1990.
17. Sunkari, S. R., P. S. Beasley, T. Urbanik II, and D. B. Fambro. *A Model to Evaluate Bus Priority at Signalized Intersections.* Southwest Region University Transportation Center, August 1994.
18. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

REALTRAN: An Off-Line Emulator for Estimating the Effects of SCOOT

H. RAKHA AND M. VAN AERDE

An off-line emulation tool, entitled REALTRAN (REAL-time TRANsyt), which can emulate the SCOOT version 2.2 signal optimization logic, has been developed. REALTRAN was derived from the TRANSYT-7F model (TRANSYT version 7F) by introduction of various constraints into the optimization logic of TRANSYT-7F. These constraints allow the user to select optimization parameters that enable REALTRAN to operate in a fashion similar to that of the original SCOOT signal optimizer logic. The REALTRAN model is currently intended to serve as an educational tool but can, in the future, serve as a tool for fine tuning the operation of the real-time controls of the SCOOT system in a laboratory environment, where scientific and statistically valid testing and sensitivity analyses of the signal optimization algorithms can be performed. Alternatively, REALTRAN can be utilized to estimate off-line the expected benefits of SCOOT by use of location-specific network and flow data.

Field studies have indicated that various real-time urban traffic control (UTC) systems, including the SCOOT system, are capable of attaining reductions in the range of 10 percent in the network travel time compared with conventional fixed-time signal control (*1*). However, it has been impossible to achieve these reductions consistently. It appears that the main reasons for the lack of more extensive success of these real-time UTC systems are related to the complexity of the problem, the variability of the link flows, and the inaccuracy of the vehicle detector measurements. To address each of the factors there is a need for tools to fine-tune the operation of the real-time controls reliably off-line in a laboratory environment. This environment would allow sensitivity analyses on the settings and signal optimization algorithms of such real-time UTC systems to be performed.

Initially, the structure of the TRANSYT model is reviewed, as this model forms the basis for both the SCOOT system and the REALTRAN model. The SCOOT system is also described in the following section because the REALTRAN model can, depending on user-specified inputs, attempt to replicate the SCOOT signal optimization logic. It must be noted at this point that any reference to TRANSYT in this paper refers to the general optimization logic of TRANSYT including that utilized in the TRANSYT-7F version.

First, the general concept of the REALTRAN model is presented. Subsequently, the details of the REALTRAN hill-climbing procedure and the associated use of high-sensitivity parameters are presented, together with the cycle-length optimization process and a description of how the cycle-length, offset, and phase-split optimization procedures operate together.

Because of the limited space available in this paper, in the following section only a simple example illustration is presented to illustrate briefly how the REALTRAN model makes frequent but minor alterations to the signal settings such that the signal plan evolves incrementally toward the near-optimum signal settings. A more detailed application of the REALTRAN model can be found in the literature (*2*). Finally, a summary and the conclusions of the paper are provided.

Department of Civil Engineering, Queen's University, Kingston, Ontario K7L 3N6, Canada.

BACKGROUND

In this section we describe the more macroscopic concepts of the TRANSYT model, followed by a microscopic description of the TRANSYT hill-climbing procedure. The latter detailed description will provide the reader with an appreciation of the difference between the TRANSYT and SCOOT logic. In addition, this section provides a description of the SCOOT signal optimization logic before describing how REALTRAN can model the SCOOT logic.

Conceptual Description of TRANSYT

The TRANSYT model is perhaps the most widely used off-line signal optimization tool (*3*). The TRANSYT model was developed at the Transport Research Laboratory, and many versions have evolved. One of these TRANSYT versions is TRANSYT-7F, which was developed at the University of Florida (*4*).

TRANSYT is a macroscopic, deterministic simulation and optimization model. The model requires the link flows and link turning proportions as inputs and assumes them to be constant for the entire simulation period. The TRANSYT program simulates the traffic conditions for the duration of one complete cycle length, and these conditions are assumed to be representative of all other cycles.

The TRANSYT model is macroscopic because the traffic module models the flow of vehicles as cyclic flow profiles (CFPs) rather than modeling individual vehicles. Specifically, the cycle length is divided into a number of short time steps, which are typically 1–5 sec long. The CFP records platoons of vehicles as successive steps within the representative cycle, and the shape of the CFP is calculated by the model for each one-way flow in the study area.

TRANSYT Hill-Climbing Procedure

The TRANSYT program carries out a sequence of iterations between the traffic simulation module and the signal setting optimization module. For the initial signal settings the traffic module estimates the performance index (PI) by simulating the traffic as it reaches each intersection in the network. Subsequently, alterations are made to the signal settings by the optimization module. These

signal setting changes are sent to the traffic module, which alters the CFP that leaves each signal, which in turn affects the arrival profile at any downstream signals.

The TRANSYT program searches for the optimum signal settings by using a two-stage procedure. In the first stage the optimum cycle length is found through a search, at user-specified intervals, within a user-specified range of minimum and maximum cycle lengths. Subsequently, in the second stage, the cycle length that produced the lowest PI in the former search is investigated in further detail by a hill-climbing procedure to determine the optimum offsets and phase splits for this cycle length.

As the shape of the PI objective function versus offset, phase split, and cycle length is not always convex, local minima may exist. Thus most conventional derivative methods would fail to find the global minimum and could frequently be caught in local valleys. The offset and phase-split searches verify that the global minimum PI is found by use of a combination of small, medium, and large step sizes that usually move the optimizer away from a local minimum. Although there are no absolute guarantees that the search will find the global minimum, the TRANSYT heuristic has during the past 25 years been found to yield a very practical trade-off between accuracy and efficiency. Considerable work has been conducted to test and evaluate other search methods; however, no major improvements have been made to the TRANSYT search heuristic (*5,6*).

Overview of the SCOOT System

The SCOOT real-time UTC software (*7,8*) uses a traffic simulation model similar to that used by TRANSYT (*9*). This simulation model is used on-line, however, during every cycle by the optimizer to evaluate alternative signal timings and thus find the best signal settings based on the prevailing dynamic traffic conditions. The objective of SCOOT, as in the TRANSYT model, is to minimize the PI. Traffic is also modeled as a CFP in the SCOOT traffic model; however, the time interval is fixed at 4 sec, and each link's inflow CFP is measured directly from the street by detectors, as opposed to being inferred from the turning movements of the upstream intersection.

The SCOOT optimizer updates the traffic signal plan on a cycle-by-cycle basis. In doing this the optimizer uses the previous cycle's signal settings as a seed in the search for new timings and makes minor, but very frequent, alterations to these seed signal settings. The changes to the signal settings are made based on a restricted search for a minimum PI in the immediate vicinity of the seed signal settings, rather than by an exhaustive search for a global minimum PI, as in TRANSYT. The SCOOT signal optimizer effectively uses an elastic coordination plan that stretches and shrinks the coordination scheme to match the latest situation recorded by the real-time cyclic flow profiles. The changes made to the current plan, while minor, are frequent, so that over time the plan evolves considerably without causing major disruptions to traffic.

The three key principles of the SCOOT real-time UTC system that make it different from the TRANSYT model are as follows:

- to measure the cyclic flow profile in real time as opposed to deriving it from upstream turning movements,
- to update an on-line model of queues continuously as opposed to only updating once, and
- to make incremental as opposed to global optimizations to the signal settings.

OVERVIEW OF THE REALTRAN CONCEPT

We developed the REALTRAN model by adding to and altering the TRANSYT-7F optimization logic to perform the following functions: (a) to estimate iteratively the optimum signal timings of a network of traffic signals for a time series of link flows, (b) to evaluate these optimum signal timings, using a second set of link flows, and (c) to allow the user to specify constraints to the standard TRANSYT signal optimization logic.

The REALTRAN model, as does the SCOOT logic, involves the application of the TRANSYT optimization module every minute to an externally specified data stream. Each TRANSYT application is seeded with the signal timings that were found during the previous minute. The search can be constrained, depending on the user-specified parameters, to look only for those new signal timing solutions that are very similar to the previous minute's signal timings. The use of a good seed, plus the constraints on the optimization, can therefore significantly reduce every minute's computational requirements, because the optimizer starts from signal settings that are already very close to the optimum signal settings. Furthermore, as only minor changes are made to the signal settings, this approach can also avoid disruptions to the traffic during signal plan changes in a fashion similar to the SCOOT logic.

Details of the Optimization Procedure

The REALTRAN model can perform standard TRANSYT signal optimization, a restrained SCOOT-like signal optimization, or any user-specified restrained signal optimization, depending on the user-specified input parameters. The REALTRAN model, in simulating the SCOOT logic, makes three main restrictions to the hill-climbing process of the TRANSYT model, as follows: constraining the offset optimizer by allowing changes of only a few seconds for each optimization, constraining the phase-split optimizer to only a few-second changes, and making cycle-length optimizations at intervals of not less than 3 min, using a limited range of potential cycle-length choices.

The first key to the potential success of this effort derives from forcing the TRANSYT optimization to start the search for signal timings for the subsequent minute at the signal timings that were found at the conclusion of the previous minute.

The second key to the practical success of this effort derives from the implementation of a user-specified constraint on the number and size of the optimization steps that can be taken each minute to find improvements on this previous minute's timings.

Although the above two steps toward making a real-time version of TRANSYT satisfy the offset and phase-duration considerations of REALTRAN, a further addition to the logic was required to enable TRANSYT to mimic SCOOT's changes in cycle length. Specifically, at user-specified time intervals (typically every 2–5 min) the model determines whether the overall network PI can be decreased by moving to either a longer or a shorter cycle length. The maximum amount of permitted change in the cycle length is again user specified and cycle-length dependent to replicate SCOOT's different cycle-length increments.

Details of Input Requirements

The REALTRAN simulation program requires four input files, namely, a master file, the standard TRANSYT input file, a link flow

file to be used for signal optimization, and a link flow file to be used for the evaluation of the signal settings.

In the master file the names of the various input and output files are specified. In addition, the cycle-length increment thresholds, cycle-length increments, and the maximum number of steps to be used by the hill-climbing procedure are specified. This provides the user with the flexibility of testing different optimization constraints on the potential traffic signal settings. Furthermore, the "optimization link flow file" identifies the flows to be used by the optimizer to select the new signal settings each minute, and the "evaluation link flow file" identifies the flows to be used in evaluating these new signal settings. In this fashion the model can simulate either a time lag in the optimization procedure or the fact that the link flows input to REALTRAN may be filtered flows and thus differ from the actual flows.

The above input data requirements imply that the present version of the REALTRAN program is intended to be a simulation model that can replicate SCOOT's real-time controls and not a real-time control system that is to be a competitor for SCOOT. What is important, however, is that the user either can specify the cycle-length thresholds, increments, and frequency of full optimization as those used by the actual SCOOT system, making the REALTRAN model simulate control algorithms in a fashion similar to the SCOOT signal optimizer, or vary these parameters to study the impact on the PI.

SPECIFICS OF THE REALTRAN OPTIMIZATION MODULE

The main reason for the success of the modified hill-climbing routine that is described in this section is the elimination of the arbitrariness of the seed solution that is utilized to initiate the search for the next minute's signal timings. In this section we discuss the modified hill-climbing module in further detail and also illustrate how the use of high-sensitivity parameters in the standard TRANSYT input file can help to speed up the optimization process and assist in modeling mini-areas to mimic the SCOOT logic. We also briefly describe the cycle-length search process and the combined cycle-length, phase-split, and offset optimization process.

Modified Hill-Climbing Module

We mimicked the incremental nature of SCOOT in REALTRAN by setting up the modified hill-climbing module so it uses the previous interval's signal timing settings as the seed to initiate the search for the optimum signal settings for the following time period. Figure 1 demonstrates some of the details of how the modified hill-climbing module operates by using a typical two-step constraint, for all possible combinations of the shape of the PI curve, as indicated below.

Case (*1*) illustrates an optimization scenario in which an initial step reduces the PI and therefore leads the algorithm to proceed with a second step. This second step leads to a further reduction of the PI. However, as the limit of a maximum of two steps in a given search direction prevents the algorithm from proceeding any further, the signal settings that are found following the first two steps are retained and are considered to be the new approximation of the global optimum.

In Case (*2*) of Figure 1 the first optimization step is again shown to reduce the PI. However, when the second step is taken in the same direction, the PI is shown to start to increase again. Consequently it appears that the algorithm has found a local minimum, and the algorithm returns to the signal timing settings that were found following the first step as the new-found approximation to the global optimum signal settings.

Case (*3*) illustrates how an attempted shift in the signal timings to the right results in an increase in the PI. Consequently the algorithm reverses its search direction and doubles its step size to make a shift in the signal timings to the left past the initial signal settings and returns the step size to its original value. This move is shown to lead to a reduction in the PI compared with the initial settings. The limit of a maximum of two steps in any direction prevents the algorithm from proceeding any further in this direction. Case (*4*) is similar to Case (*3*) but makes only one step to the left.

Finally, the example in case (*5*) illustrates that a shift in either direction leads to an increase in the PI. Consequently the optimum signal settings are considered to be retained by simply keeping the initial signal settings that existed when the search was initiated.

The above hill-climbing decision logic is used in both the offset and the phase-split optimization processes within REALTRAN, using the minimum resolution if not specified otherwise in card type 4 of the standard TRANSYT input file. Utilizing card 4, one can investigate the effect on SCOOT of different step sizes. In addition, the maximum number of steps utilized by the REALTRAN model in each direction can be set by the user. This value is typically set to two to mimic SCOOT's minimum signal changes while maintaining the ability to escape local minima.

Use of High-Sensitivity Parameters

The sensitivity parameter sets a limit below which any downstream changes to the CFP are neglected. The base TRANSYT model permits the use of a sensitivity parameter card (card type 6) to limit the extent to which the downstream effects, at other intersections, of a change in signal timings at a given intersection will be examined. The setting of these parameters affects the time required for execution of the model. Because the SCOOT optimization logic considers only the flows arriving at the traffic signal in generating the cycle-length duration and phase splits, and considers only the surrounding signals in estimating the optimum offsets, high-sensitivity parameters of 20 percent were utilized as the default in the REALTRAN model.

Search for the Optimum Cycle Length

To minimize any disruptions to traffic within SCOOT, the changes in cycle length are restricted to be very small. This objective is achieved within REALTRAN in three ways: by controlling the number of cycle lengths to be evaluated, by controlling the cycle-length increment, and by controlling the frequency of cycle-length optimization runs.

During each cycle optimization the REALTRAN program uses the previous interval's cycle length as a seed. The minimum cycle length to be considered is then selected as the initial seed cycle length minus the user-specified cycle-length increment. Similarly, the maximum cycle length to be considered is chosen as the seed cycle length plus the cycle-length increment. These optimizations are performed at a user-specified interval and thus can be performed, for example, every 3 min, as is the case with the SCOOT system or at more/less frequent intervals.

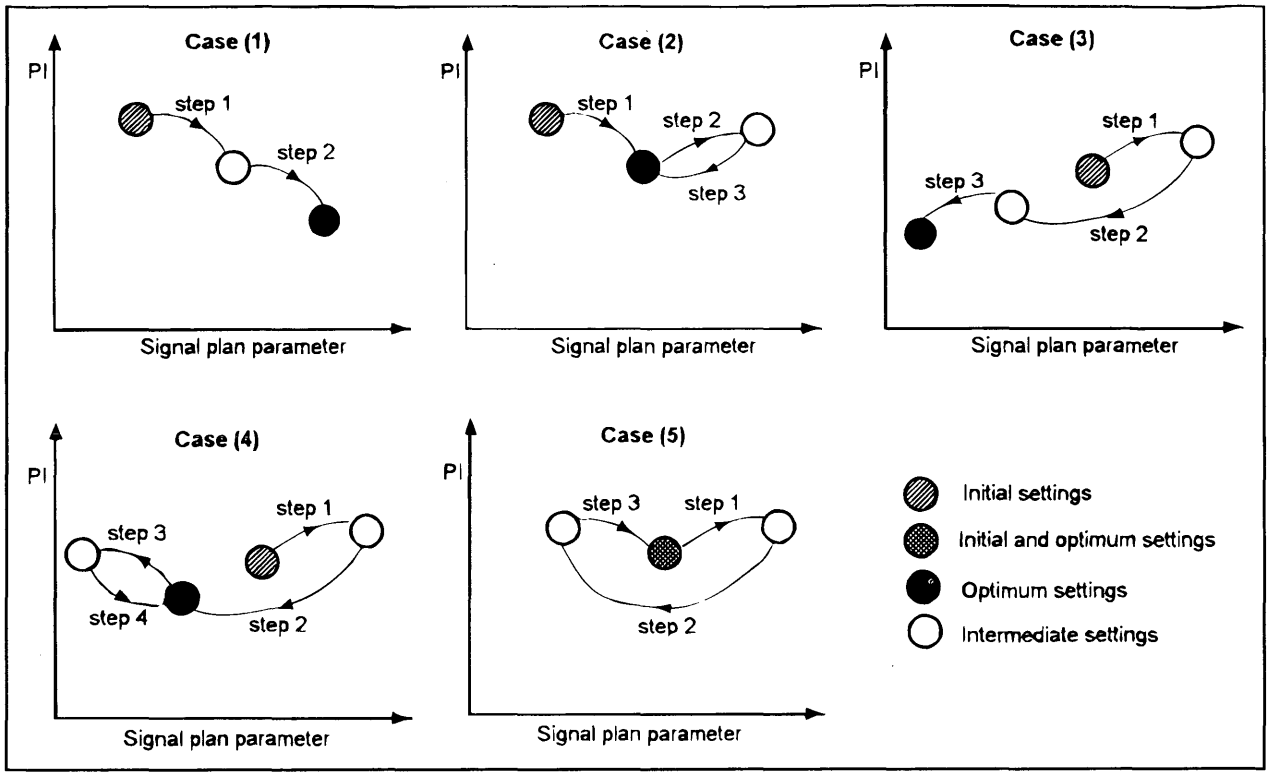


FIGURE 1 Modified hill-climbing mechanism for a two-step constraint.

The REALTRAN model can use as many as three different cycle-length increments, depending on the cycle length. The model can therefore again replicate the SCOOT system's cycle optimization logic to a large extent.

EXAMPLE ILLUSTRATION

To illustrate briefly how the REALTRAN model can examine the impacts of a constrained SCOOT-like optimization for different traffic flow patterns, an extract of the results from a nine-traffic-signal grid network is presented in this section. Because of limited space, only a very brief summary of the results is presented; however, the details of the network and results can be found in the literature (*1,10*).

The network was simulated for a hypothetical sequence of 5 hr within which the flows experienced a peak in the eastbound direction followed by a peak in the westbound direction. Each minute, the REALTRAN optimizer optimized the signal settings, using two optimization scenarios. In the first, the REALTRAN model utilized the standard unconstrained TRANSYT full optimization every minute. Subsequently, in the second scenario the REALTRAN optimizer was constrained to emulate the SCOOT signal optimization logic. To simplify the illustration of the traffic flow pattern, only a sample of the link flows for 10 typical minutes during the simulation period, for the four approaches to the traffic signal located at intersection 5, is provided in Table 1. For these sample arrival link flows Table 2 illustrates the signal settings that were selected by the REALTRAN optimizer for the two scenarios studied, namely, the TRANSYT and SCOOT emulations.

It can be noted from Table 1 that during this time period the flow on the westbound, southbound, and northbound approaches to signal 5 remained constant while the flow on the eastbound approach increased from 525 to 750 vehicles/hr. This change represents

TABLE 1 Link flows arriving at signal 5

Time into Simulation (min)	Approach flows to signal 5 (vehicles/hr)			
	Eastbound	Westbound	Southbound	Northbound
35	525	350	500	250
36	550	350	500	250
37	575	350	500	250
38	600	350	500	250
39	625	350	500	250
40	650	350	500	250
41	675	350	500	250
42	700	350	500	250
43	725	350	500	250
44	750	350	500	250

TABLE 2 Signal settings chosen by the TRANSYT and SCOOT emulations (Signal 5)

Time	TRANSYT Emulation					SCOOT Emulation				
(min)	Cycle	Offset	Phase 1	Phase 2	PI	Cycle	Offset	Phase 1	Phase 2	PI
35	48	0	24	24	80.2	40	12	20	20	78.3
36	44	20	23	21	83.7	40	12	21	19	82.9
37	52	0	27	25	87.7	44	13	23	21	82.0
38	52	3	28	24	91.8	44	11	24	20	92.5
39	52	3	28	24	97.8	44	12	24	20	99.3
40	72	9	40	32	100.1	48	13	26	22	103.8
41	64	6	36	28	106.8	48	9	27	21	115.3
42	68	5	39	29	112.3	48	8	27	21	133.3
43	76	9	44	32	119.9	52	11	29	23	156.9
44	80	4	47	33	129.8	52	9	30	22	173.4

approximately a 50-percent increase in flow in 10 min. Although such an increase within 10 min may not be very likely to occur in practice, the intent was to investigate, by means of such a rapid hypothetical change in flows, the robustness and responsiveness of the SCOOT emulator.

It can be noted from Table 2 that the TRANSYT emulator responded to these drastic changes in flows with major changes in the signal settings. Specifically, the cycle length changed from 52 to 72 sec, which is equivalent to a 35-percent change, at the onset of the 40th min. In contrast, the SCOOT emulator altered the cycle length by only 4 sec from 44 to 48 sec, following which the SCOOT emulator was restricted from performing another cycle-length optimization for 3 min. Also, the TRANSYT emulator made a drastic change in the offset, from an offset of 0 sec to an offset of 20 sec, at the start of the 36th min, for a change in cycle length from 48 to 44 sec, while the SCOOT emulator was restricted to minor alterations. The difference between the maximum and minimum offsets selected by the TRANSYT emulator was 20 sec (0 to 20 sec), as opposed to a 4-sec difference for the SCOOT emulator (13 to 9 sec). The TRANSYT emulator also varied the phase 1 duration from 23 to 47 sec at the onset of the 40th min, which is equivalent to a 14-sec variation, while the SCOOT emulator varied the phase 1 duration only from 24 to 26 sec.

In comparing the PI at each minute (columns 6 and 11 of Table 2), it is evident that initially the PI for the SCOOT emulator was lower. However, as the TRANSYT emulator was not restricted, this trend changed as the link flows varied. However, because the unconstrained optimizer made larger variations to the signal settings, it caused major disruptions to the traffic and thus added inefficiencies that are not accounted for in Table 2. It must be noted also, based on these limited results, that the SCOOT emulator succeeded, albeit with a lag, in following the trend in the variation of cycle length, offset, and phase split, thus allowing the signal timings to evolve over time in a fashion similar to those of the SCOOT signal optimizer.

SUMMARY AND CONCLUSIONS

We believe that in this paper we have made a significant step toward addressing the need for a simulation tool that is capable of emulating the SCOOT optimization logic. The structure of such a model, entitled REALTRAN, has been presented. The REALTRAN model is based on the well-known TRANSYT program, specifically TRANSYT-7F, by constraining the hill-climbing procedure within the TRANSYT model. The user can specify the maximum number of steps allowed in each direction and thus allow REALTRAN to model various constraining conditions that are different from the SCOOT default. The REALTRAN model permits the specification of a separate link flow file that is used to select the optimum signal settings and of an optional link flow file that is used to evaluate these signal settings. Furthermore, the user can specify, through the master file, both the cycle-length increments to be evaluated and the cycle-length thresholds separating various cycle-length increment zones.

It must be noted that the REALTRAN program, like the SCOOT, SCAT, and PRODYN traffic models, is built on the vertical queue model and thus cannot consider in detail the effect of downstream link congestion on the signal output. These models operate well as long as the network is not overly congested. However, they fail to model the effect of downstream congestion on the capacity of upstream intersections during queue spillback. In the case of SCOOT the queuing model is updated by queue measurements from the field. In addition, the REALTRAN model cannot model the rerouting of traffic in response to changes in signal timings.

REFERENCES

1. Boillot, F., J. M. Blosseville, J. B. Lesort, V. Motyka, M. Papageorgiou, and S. Sellam, Optimal Signal Control of Urban Traffic Networks. *Proc., IEE International Conference on Traffic Road Monitoring and Control*, 1992, pp. 75–79.
2. Rakha, H., M. Van Aerde, and E. R. Case, Experiments in Incremental Real-Time Optimization of Phase, Cycle, and Offset Times Using an On-Line Adaptation of TRANSYT-7F. Presented at the Engineering Foundation Conference on Traffic Management: Issues and Techniques, Palm Coast, Florida, April 1991.
3. Robertson, D. I. *TRANSYT: A Traffic Network Study Tool*, RRL Report 253, Crowthorne, UK, 1969.
4. Courage K., and C. Wallace. *TRANSYT-7F Users Guide*. University of Florida, Gainesville, 1991.
5. Timmermans, W. F., P. P. Van Den Bosch, and J. J. Klijnhout. Improved Network Control by Using TRANSYT. *Traffic Engineering and Control*, Vol. 21, 1979, pp. 353–356.
6. Foulds, L. R. TRANSYT Traffic Engineering Program Efficiency Improvement via Fibonacci Search. *Transportation Research*, Vol. 20A, No. 4, 1986, pp. 331–335.
7. Hunt, P. B., D. I. Robertson, R. D. Bretherton, and R. I. Winton. *SCOOT: A Traffic Responsive Method for Coordinating Signals*. TRRL Report LR, 1981, p.1014.
8. Hunt, P. B., D. I. Robertson, R. D. Bretherton, and M. C. Royle. The SCOOT On-Line Traffic Signal Optimization Technique. *Proc., IEE International Conference on Road Traffic Signalling*, Publication No. 207, 1982, pp. 59–62.
9. Robertson, D. I., and R. D. Bretherton. Optimizing Networks of Traffic Signals in Real-Time—The SCOOT Method. *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 1, February, 1991, pp. 11–15.
10. Rakha, H. *A Simulation Approach for Modeling Real-Time Traffic Signal Controls*. Ph.D. dissertation. Queen's University, Kingston, Ontario, Canada, 1993.

Publication of this paper sponsored by the Committee on Traffic Signal Systems.

Pioneer Application of Passer IV in the Houston Metro-RCTSS Project

CHANG LIU, NADEEM A. CHAUDHARY, HARRY C. SIMEONIDIS, AND SIREESHA SIRIGIRI

The Metropolitan Transit Authority of Harris County (METRO) is currently conducting the Regional Computerized Traffic Signal System project in the Houston metropolitan area. This project involves optimization of signal timings at 3000 intersections and is one of the largest undertakings of this type in the country. This study concerns the application of PASSER IV, a new progression-bandwidth–based program that optimizes network signal timing, to two subsystems located at the Texas Medical Center, south of downtown Houston. These systems were selected by METRO for early implementation.

The Metropolitan Transit Authority of Harris County, Texas (METRO) is currently conducting the Regional Computerized Traffic Signal System (RCTSS) project in the Houston metropolitan area. Consisting of 3000 signalized intersections, this undertaking is emerging as one of the largest signal systems in the country. Coordination of traffic signals in the system will provide for improved traffic flow. As an advanced traffic management system, RCTSS will be integrated with other intelligent vehicle highway systems projects in the Houston area to improve the efficiency of the transportation system, reduce urban congestion and air pollution, and enhance motorist safety.

As a consultant to METRO, Rust Lichliter/Jameson was responsible for developing system operations plans for two signal networks in the Texas Medical Center/Astrodome area, which was identified as one of two areas for early implementation. The consultant team had about 4 months to provide its report to METRO. Due to the limited time, the work had to be performed expediently. A description of this work begins with the study area. Study objectives are outlined, including the optimization and simulation tools utilized to perform the work. Finally, the results of the study are presented.

STUDY AREA

Network Description

The Texas Medical Center/Astrodome area is bounded by I-610 on the south, US-59 on the north, Buffalo Speedway on the west, and State Highway 288 on the east. As shown in Figure 1, it consists of two subsystems; Main/Fannin (MF) and Holcombe/Old Spanish Trail (HOST). The MF subsystem is a grid network consisting of 10 two-way arterials, 2 one-way arterials, 1 mixed (partially one-way) arterial, and 26 signalized intersections. Adjacent to the MF subsystem, the HOST subsystem is composed of 3 two-way arterials, 2 one-way frontage roads, and 13 signalized intersections (including 2 diamond interchanges). Among many important institutions and establishments in this area, Hermann Park and Texas Medical Center are located between these two subsystems. Texas Medical Center is one of the largest medical research centers in the world. Rice University is located to the west of the MF subsystem.

C. Liu, H. C. Simeonidis, and S. Sirigiri, Rust Lichliter/Jameson, 11111 Brooklet Dr., Suite 100, Houston, Texas 77099-3596. N. A. Chaudhary, Texas Transportation Institute, Texas A&M University System, College Station, Texas 77843-3135.

Existing Network Traffic Flow Conditions

Traffic congestion in this area is common, with significant peaking characteristics due to the unique land use pattern of the area. Main Street, the primary north-south arterial with three lanes in each direction reaches flow rates in excess of 2300 vehicles per hour northbound during p.m. peak periods, while Holcombe, also with three lanes in each direction, reaches flow rates in excess of 2000 vehicles per hour eastbound during a.m. peak periods.

Traffic flow rates for each arterial during various peak periods are presented in Tables 1 and 2. For the MF network, traffic flow rates are very light on Montrose northbound and Sunset westbound during the a.m. peak period. In the same network, the heaviest traffic flow rates are on Main northbound during the p.m. peak period and Holcombe eastbound in the a.m. peak period. For the HOST network, traffic flow rates are very light on Almeda southbound during the a.m. peak period and Holcombe westbound during the noon peak period. In the same network, the heaviest traffic flow rates are on the State Highway 288 west service road and Old Spanish Trail westbound during the p.m. peak period.

The volume to capacity (v-c) ratios are summarized in Tables 3 and 4. For the MF network, the p.m. peak period has the worst operating conditions, in which approximately 20 percent of exclusive left turn phases and 14 percent of the through phases are operating at or over capacity. For the HOST network, the worst operating condition also occurs in the p.m. peak period, during which 55 percent of the left turn phases and 20 percent of the through phases are operating at or over capacity. The most congested intersections in the MF network are the following:

Morning peak:

- Main at Sunset, University, Dryden, and Holcombe;
- Fannin at N. McGregor and Holcombe

Noon peak:

- Fannin at Binz, N. McGregor, M. D. Anderson and Holcombe;

Afternoon peak:

- Main at Blodgett, Sunset, N. McGregor, University, and Holcombe;

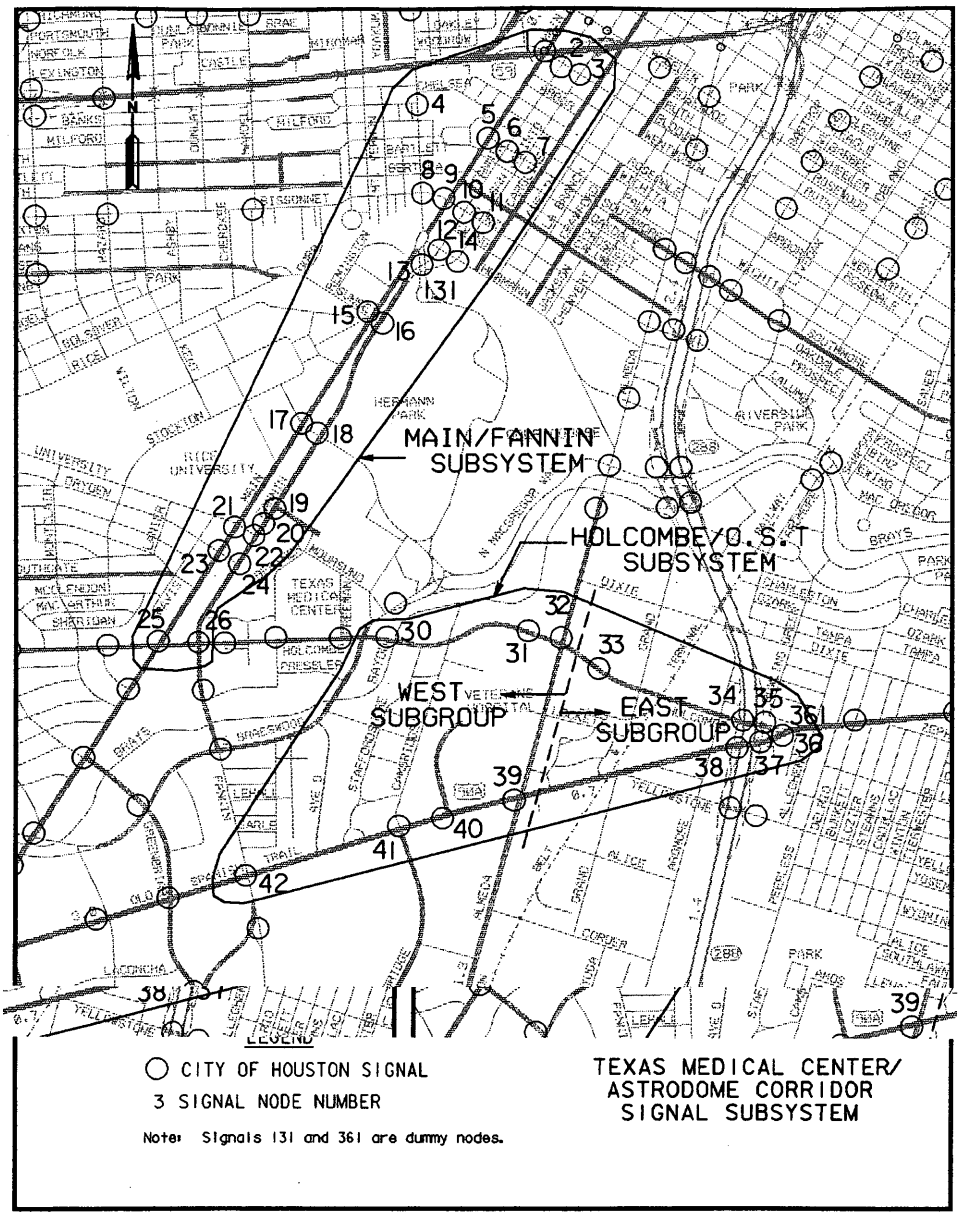


FIGURE 1 Map of study area.

- Fannin at N. McGregor, M.D. Anderson, Dryden, and Holcombe; and
- Binz at Montrose and San Jacinto.

The most congested intersections for the HOST network are the following:

Morning peak:

- Holcombe at Braeswood and State Highway 288 frontage road;
- Old Spanish Trail at Fannin, Almeda, and State Highway 288 east frontage road;

Noon peak:

- Holcombe at Braeswood and Veterans Administration Hospital;
- Old Spanish Trail at Fannin, Mixon, Almeda, and State Highway 288 frontage road;

Afternoon peak:

- Holcombe at Braeswood, Almeda, and State Highway 288 frontage road; and
- Old Spanish Trail at Fannin, Cambridge, Almeda, and State Highway 288 frontage road.

STUDY OBJECTIVES

Bandwidth-based signal timings are very popular among drivers and engineers in Texas. Thus, the main objective of the study was to generate signal timings that provide maximum progression bands

TABLE 1 MF Subsystem Flow Rate

Artery	Street	Flow Rate (vph)					
		A M Peak		Noon Peak		P M Peak	
No	Name	NB/EB	SB/WB	NB/EB	SB/WB	NB/EB	SB/WB
1	Montrose	596-832	1107-1293	820-844	603-865	1552-1608	808-1064
2	Main	441-1980	483-1379	275-1488	222-1643	407-2340	412-2092
3	Fannin N	N/A	919-1490	N/A	956-1968	N/A	948-1810
4	Fannin S	387-1380	443-1500	632-1428	524-2232	292-1153	417 1793
5	San Jacinto	634 1388	N/A	952 2208	N/A	488-2332	N/A
6	Blodgett	N/A	304-552	N/A	304-668	N/A	364-828
7	Southmore	44-96	128-248	100-236	152 204	116-352	156-452
8	Binz	652-776	452-516	384-452	268-661	857-1004	528-904
9	Hermann	192-280	76-136	144-364	116-200	340-436	108 156
10	Sunset	544-776	36	184 408	68	156-376	136
11	N MacGregor	24-332	548-621	100-384	536-573	172 369	694-1056
12	University	536-860	42-231	532-676	92-559	383-672	66-668
13	Dryden	309 367	61-167	147-236	124-209	176-198	267-329
14	Holcombe	1935-2047	907-1348	813-1752	892-1188	814-979	1739-1914

on all the arterials of the two networks. However, it should also be pointed out that realistic progression bands can only by produced under certain conditions, including little or no platoon dispersion, low volumes, and few vehicles turning from the cross streets (*1*). Furthermore, even under ideal traffic conditions, good progression may not be achievable for some combination of cycle length, travel speeds, and link distances. Since traffic conditions in the study networks were less than ideal, delay minimization was selected as the secondary objective.

The consulting team had about 4 months to complete all tasks, including data collection, generation of optimal signal timing plans, assessment of results (estimating measures of effectiveness), and production of the final report. Due to time limitations, efficient personal computer-based optimization and simulation tools had to be used. Because of the grid layout and the presence of one-way arterials in the networks, the consultant team was reluctant to use PASSER II (*2*) (a program to optimize progression bandwidth on two-way arterials), which would have required breaking the networks into separate arterials. This approach would also have required manual adjustments to combine the optimal signal timings for individual arterial. A prerelease beta version of PASSER IV (*3,4*), a signal timing optimization package for multiarterial networks, became available from Texas Department of Transportation (TxDOT) and was acquired for use in this study. It was also decided to use TRANSYT-7F (*5*) to simulate the optimal signal timings for both networks. PASSER IV, described in the next section, provides a number of features that were needed for completing this study efficiently.

DESCRIPTION OF PASSER IV-94

PASSER IV-94 is the latest addition to the PASSER (II and III) (*2,6*) family of bandwidth-based signal timing optimization programs developed for (TxDOT) by Texas Transportation Institute, Texas A&M University System. PASSER IV is applicable to single

TABLE 2 HOST Subsystem Flow Rate

Artery	Street	F ow Rate (vph)					
		A M Peak		Noon Peak		P M Peak	
No	Name	NB/EB	SB/WB	NB/EB	SB/WB	NB/EB	SB/WB
1	Holcombe	195-971	760-1346	272-800	312-685	390-1350	244-1456
2	O S T	768-904	456-1288	641-1472	456-968	660-1700	460-1448
3	Almeda	688 1416	288-471	472-972	444-448	556-1836	620-900
4	SH-288 WSR	N/A	948-1519	N/A	824-1136	N/A	1028-1188
5	SH-288 ESR	948-1072	N/A	537-716	N/A	533 916	N/A

TABLE 3 MF Subsystem V-C Summary

Time	Phase	% of Phase						Total
Period		v/c > 1	v/c > .9	v/c > .8	v/c > .7	v/c > .6	v/c <= .6	
A M Peak	Left Tur	3 5	0	3 5	7	10	76	100
	Through	7 5	3	7	4	5 5	73	100
Noon Peak	Left Tur	10	3 5	10	14	3.5	59	100
	Through	2	4	2	8	11	73	100
P M Peak	Left Tur	13	7	0	10	0	70	100
	Through	7 5	7 5	11	9	12	53	100

arterials and multiarterial (closed loop or open) networks. PASSER IV has the following outstanding features:

- Determines signal cycle length, green splits, offsets, and phase sequences that simultaneously maximize progression bands on all arterials in the network.
- Handles either one-way or two-way arterials.
- Allows user to specify arterial priorities.
- Allows link speeds to vary between user-specified limits.
- Prints a specified number of best signal timing solutions.
- Provides efficient optimization techniques.
- Produces an extensive output report that includes time-space diagrams, bandwidth efficiencies, signal timing tables, and measures of effectiveness (approach delays, v-c ratios, and level of service).
- Optionally generates input data files for TRANSYT 7F program. These files can be used either to simulate bandwidth solutions or to perform bandwidth-constrained delay optimization of bandwidth based signal timings.
- Provides a menu-driven graphic user interface (GUI) with pull down menus and full mouse support. The PASSER IV GUI is capable of running the TRANSYT-7F program from its main menu.

PASSER IV, like other programs available to traffic engineers, is limited to applications in which undersaturated flows exist. In addition, heavy turning volumes reduce its ability to produce meaningful progression bands. These deficiencies are due to the inherent weakness of the bandwidth optimization approach. PASSER IV can still be a useful tool; however, when the traffic conditions are less than ideal, engineering judgement should be used to select proper cycle length, green splits, and queue clearance times.

In summary, PASSER IV provides a comprehensive signal timing optimization environment for undersaturated multi-arterial networks. It is capable of activities for networks similar to those available for arterials through the combined use of Arterial Analysis Package (*7*) and PASSER II.

DEVELOPMENT OF SIGNAL TIMING PLANS

This task was carried out in three main steps: collection and preparation of traffic data, generation of optimal signal timing plans, and assessment of the benefits of proposed signal timing plans through comparison with existing signal timing plans. The following subsections describe these steps and present the results of the study.

TABLE 4 HOST Subsystem V-C Summary

Time	Phase	% of Phase						Total
Period		v/c > 1	v/c > .9	v/c > .8	v/c > .7	v/c > .6	v/c < .6	
A M Peak	Left Tur	30	10	5	0	15	40	100
	Through	9	4 5	4 5	6	15	61	100
Noon Peak	Left Tur	35	10	0	30	0	25	100
	Through	4.5	4 5	4 5	8 5	8 5	69 5	100
P M Peak	Left Tur	40	15	0	0	15	30	100
	Through	13	17	4 5	9	4 5	52	100

Data Collection and Preparation

Intersection turning movement counts, intersection sketches, and average running speed data for three peak periods of the average weekday were collected in the field by the consulting team. The collected turning movement volumes were reduced and adjusted to obtain flow rates. The lowest peak hour factors for MF and HOST subsystems were calculated to be 0.25 and 0.3, respectively, indicating that the area experiences very heavy 15-min flows during the peak hour.

In order to minimize the effort required for data collection without sacrificing data accuracy, the a.m. off-peak and p.m. off-peak turning movement volumes were systematically estimated based on peak period flow rates and area adjustment factors for different approaches. Existing timing plan information available in the form of color sequence charts was obtained from the City of Houston (COH) Department of Traffic and Transportation. Both geometric and operational features have been field verified by the engineers because there have been a number of recent updates in the area. Left-turn treatments and phase sequence information were extracted from the COH color sequence charts and prepared for utilization in this study. Most of the saturation flow rates were estimated based on the intersection sketches and field verifications, with a few complex ones being calculated using highway capacity software. Minimum phase time for each phase was calculated based on the curb-to-curb distance and median width information from intersection sketches, field verification and intersection layout for construction. The initial cycle length range for each network was established based on current practices in the Houston area.

Optimization of Signal Timings Using PASSER IV

PASSER IV network node link structures were established based on existing geometric features and some special program requirements. Figure 2 shows a simplified sketch of the MF subsystem. Note, the northern section of Fannin Boulevard is one-way while the southern section is two-way. Since PASSER IV cannot explicitly handle such arterials, Fannin was divided into two separate arterials: Fannin North (one-way) and Fannin South (two-way). Figure 3 shows a sketch of the HOST subsystem. Figures 2 and 3 also show the A-direction that was chosen for entering data into the program. Note that the A-direction is so identified only to differentiate it from the other direction of flow on a two-way arterial. This designation does not imply that this direction has a higher priority; arterial priority is specified separately. For this project, PASSER IV was requested to internally establish the arterial and directional priorities based on total arterial and directional volumes. Speed variation was set to be plus or minus 3 mi/hr.

After the data were coded into PASSER IV, preliminary runs were made to detect and correct coding errors. Optimization runs were then performed for each of the five analysis time periods: a.m. peak, a.m. off-peak, noon peak, p.m. off-peak, and p.m. peak. Optimized timing plans for each analysis period were generated according to the following steps:

Step 1. Select cycle length range.

Step 2. Run PASSER IV to pick the optimal cycle length.

Step 3. Run optimization routine for multiple cycle lengths within the selected range at specified intervals.

Step 4. When two-way progression did not produce a satisfactory solution for an arterial in the network, attempt one-way progression for the heavier direction.

Both artery and subsystem measure of effectiveness summaries, in terms of bandwidth efficiencies and intersection total delays were tabulated and compared. The selection criterion chosen was a combination of maximum bandwidth efficiency and minimal delay. Knowing that a short bandwidth would not be perceived and utilized effectively, a 20 percent bandwidth efficiency was chosen as acceptable threshold value. Further comparisons were made among solutions which yielded bandwidth efficiency values of 20 percent or more. The solutions with highest bandwidth efficiencies were recommended for the five analysis periods. Table 5 gives the recommended cycle lengths for this study, and gives the best cycle lengths and corresponding network efficiency.

From the analysis, it was found that the HOST system would function more efficiently if subdivided into east and west subgroups, with Almeda as the dividing line. Further analysis confirmed that the east subgroup will require lower cycle lengths. The minimum delay cycle lengths for this subgroup are summarized in Table 6.

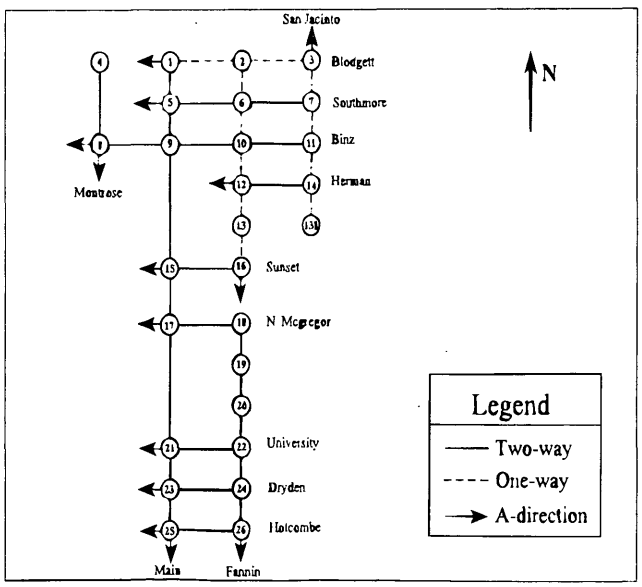


FIGURE 2 MF subsystem.

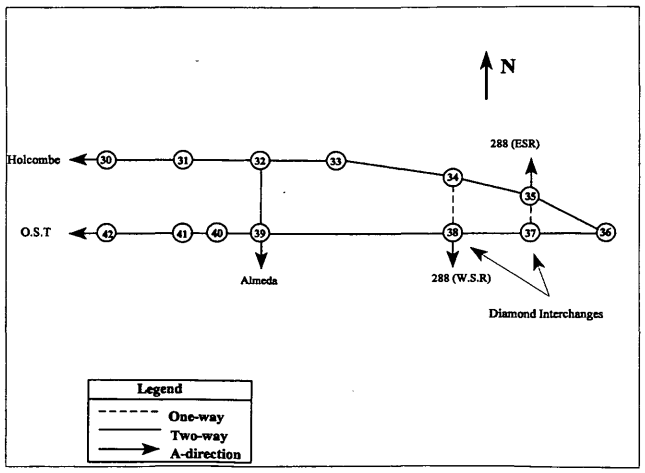


FIGURE 3 HOST subsystem.

TABLE 5 Cycle Length Recommendations

System Name	Analysis Time Period	Cycle Length Range Analyzed	Cycle Length Range Recommended	Best Cycle Length	Average Bandwidth Efficiency
MF	A.M. Peak	65 to 90	82 to 86	82	24.39
MF	A.M. Off	66 to 80	74	74	19.99
MF	Noon	70 to 84	80 to 82	80	27.66
MF	P.M. Off	66 to 80	76 to 80	76	25.19
MF	P.M. Peak	76 to 90	76 to 80	80	27.33
HOST	A.M. Peak	70 to 110	90 to 94	90	27.80
HOST	A.M. Off	70 to 90	76 to 86	78	27.97
HOST	Noon	70 to 100	76 to 80	76	25.06
HOST	P.M. Off	70 to 100	72 to 82	72	26.67
HOST	P.M. Peak	70 to 90	70 to 94	94	26.95

Comparison of Existing and Proposed Signal Timing Plans

In order to quantify the benefits of the signal timing plan improvements, it was necessary to compare the performance of the existing signal timing plans with that of the optimized timing plans generated by PASSER IV-94. TRANSYT-7F, a signal timing simulation and optimization computer package, was used to perform these comparisons. Using a feature provided by PASSER IV, the input files that generated optimal timing plans were converted into TRANSYT-7F format. The existing timing plans were also coded into TRANSYT-7F. Simulation runs were made for six scenarios using existing and PASSER IV optimized timing plans (a.m. peak, noon, and p.m. peak, existing and optimized). For brevity, only the results for the MF subsystem are presented. Table 7 shows the operational performance comparison of the existing timing plans and optimal timing plans generated by PASSER IV for the MF subsystem. Similar results were obtained for the HOST subsystem.

CONCLUSIONS AND RECOMMENDATIONS

The following conclusions and recommendations were drawn, based on all steps of the project and extensive use of the new PASSER IV software on the two signal subsystems.

Conclusions

1. PASSER IV has proved to be a user-friendly and efficient tool for developing signal timing plans for this study. It provides enough flexibility to allow different progression priorities to be set based on the user's professional judgement. Furthermore, it can be used to solve other complex network progression problems.

TABLE 6 HOST East Subgroup Cycle Lengths

Time Period	Minimum Delay Cycle Length
A.M. Peak	70
A.M. Off	70
Noon	72
P.M. Off	72
P.M. Peak	76

TABLE 7 MF Subsystem Operational Performance Comparison

Performance Measures	Units	A.M. System MOEs		Noon System MOEs		P.M. System MOEs	
		Existing	Proposed	Existing	Proposed	Existing	Proposed
Total Travel	veh-km/hr	17910	17910	18085	18085	23692	23692
Total Travel Time	veh-hrs/hr	868	727	937	564	2723	1615
Total Unif. Delay	veh-hrs/hr	309	237	266	205	389	296
Tot. Rand. Delay	veh-hrs/hr	235	166	338	27	1908	892
Total Delay	veh-hrs/hr	544	402	604	232	2296	1188
Average Delay	sec veh	30.9	22.9	34.3	13.1	104.3	54.0
Passenger Delay	pas-hrs/hr	653	483	725	278	2755	1425
Stops: Total	veh/hr	39734	32425	41260	31499	54626	45792
Percentage	%	63	51	65	50	69	58
System Speed	km/h	20.6	24.7	19.4	32.1	8.7	14.7
Fuel Consump.	m^3/hr	4.2	3.5	4.4	3.1	10.0	6.7
Operating Cost	$/hr	6491	5569	6756	5134	12802	9364
Perform. Index	DI	1647.6	1303.1	1750.3	1106.5	3813.5	2459.8

Note: 1 km = 0.62 mile; 1 m^3 = 264.2 gallon

2. The ability of PASSER IV to generate input data files for TRANSYT-7F provides an efficient means of further analysis.
3. The signal timings recommended by this study for the two networks should greatly enhance traffic flow and produce considerable savings in operating costs.

Recommendations

Signal Operations

1. The optimal timing results for HOST and the subsystem adjacent to it on the west need to be compared. It appears that grouping these subsystems would work better at p.m. off-peak periods.
2. The east HOST subgroup should be operated alone at five time periods: a.m. peak, a.m. off-peak, noon peak, p.m. off-peak, and p.m. peak, under the cycle lengths given in Table 2.
3. Signal timings recommended by this study should be implemented to improve the flow of traffic in the two networks.

PASSER IV Future Improvements

1. It would be useful to include total network delay in the network summary section of the output report.
2. The documentation needs to address how to handle triangular networks, and should explain how to deal with conditions in which one intersection is linked to two others in a triangle. In addition, it needs to use examples to address the arterial progression weights and directional weights in more detail, especially when dealing with one-direction progression.
3. The purpose of the objective function value is difficult to understand, and should be explained in more detail. Furthermore, it would be useful to output average network efficiency.
4. Adding the capability of computing saturation flow rates would greatly enhance the program.

5. It would be useful if the program could provide the longest minimum phase requirements for the network.

ACKNOWLEDGMENTS

The authors wish to thank Stephen Ha and Ron Jenson of the City of Houston Traffic and Transportation Department for their assistance in data collection and for their valuable suggestions throughout this study.

REFERENCES

1. Bass, K. G. Another Look at Bandwidth Maximization. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 38–47.
2. Chang, E. C. P. and C. J. Messer. *Arterial Signal Timing Optimization Using PASSER II-90 Program, User's Manual*. Research Report 467-2F. Texas Transportation Institute, Texas A&M University System, College Station, Texas, June 1991.
3. Chaudhary, N. A. and C. J. Messer. *PASSER IV-94 Version 1.0, User/Reference Manual*. Research Report 1271-1F. Texas Transportation Institute, Texas A&M University System, College Station, Texas, August 1993.
4. Chaudhary, N. A. and C. J. Messer. PASSER IV: A Program for Optimizing Signal Timings in Grid Networks. In *Transportation Research Record 1421*, TRB, National Research Council, Washington, D.C., October 1993, pp. 82–91.
5. Wallace, C. E., K. G. Courage, and M. A. Hadi. *TRANSYT-7F User's Guide—Release 7*. Transportation Research Center, University of Florida, Gainesville, Florida, December 1991.
6. Fambro, D. B., N. A. Chaudhary, C. J. Messer, and R. U. Garza. *A Report on the User's Manual for the Microcomputer Version of PASSER III-88*. Research Report 478-1. Texas Transportation Institute, Texas A&M University System, College Station, Texas, September 1988.
7. Wallace, C. E. and K. G. Courage. *Arterial Analysis Package (AAP) User's Guide*. Transportation Research Center, University of Florida, Gainesville, Florida, December 1990.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Uniform and Variable Bandwidth Arterial Progression Schemes

HARI K. SRIPATHI, NATHAN H. GARTNER, AND CHRONIS STAMATIADIS

Compared with conventional uniform bandwidth progressions, variable bandwidth progression schemes offer considerable advantages for arterial traffic signal control. The variable schemes have a traffic-dependent capability that the conventional schemes lack, providing additional design flexibility and superior traffic performance. A simplified and efficient method for calculating variable-bandwidth progressions given optimized uniform bandwidth progressions is presented. Little's half-integer optimization algorithm is used first for uniform bandwidth maximization, coupled with a combinatorial phase-sequence optimization procedure. The method is then extended to calculate variable-bandwidth progressions using the multiband optimization criterion. A principal feature of this method is that it can be applied to any arterial synchronization scheme after the uniform bandwidth has been maximized.

Coordinating traffic signals on arterial streets is vital for transportation systems management. The principal objective of signal coordination is to promote the smooth and efficient flow of traffic throughout the network. Traffic signals tend to group traffic into platoons with more uniform headways than those that would otherwise occur. This platooning effect is more evident on major arterial streets, which have more signalized intersections. Under such circumstances, the uninterrupted movement of vehicle platoons through successive traffic signals can be obtained by synchronizing the signals according to the green bandwidth maximization criterion.

Models that maximize the green bandwidth have been developed by several researchers. The first computer model of the bandwidth maximization problem was developed by Little et al. (*1*). Their model is a search procedure that determines the offsets resulting in the largest two-directional bands at the given green progression speeds and cycle times. Subsequent models developed by Brooks (*2*), Bleyl (*3*), and Leuthardt (*4*) had a similar theoretical basis and similar computational results. All these models maximize the green bandwidth progression on arterial streets with two-phase signal settings. Messer et al. (*5*) developed the PASSER-II model, which is based on Little's and Brooks' algorithms and enhanced by a phase sequence optimization procedure. A mathematical programming formulation of the problem was introduced by Little et al. (*6*) in the MAXBAND model, in which bandwidth, cycle length, phase sequence, and progression speeds are optimized. This approach, later extended to network optimization by Chang et al. (*7*), is based on mixed-integer linear programming and requires a mathematical programming package.

A basic limitation of those early bandwidth maximization models is that the progression schemes that result are based on the total directional arterial traffic volume. Thus, signal settings are not sensitive to the actual traffic flows on the links of the arterial, which can vary significantly due to variations in turn-in and turn-out amounts of traffic at the different intersections of the arterial. Therefore, in a uniform bandwidth progression scheme, the green band may either be wasted at intersections with lower through moving traffic, or deprived from other intersections with higher through moving traffic. An attempt to remedy this problem was made by Tsay and Lin (*8*), who developed an "inverted funnel" progression scheme. However, the band could only grow wider along the arterial, but not be reduced, and hence could not be adequately tailored to variable flows. A more effective variable bandwidth progression scheme was developed by Gartner et al. (*9*) in the MULTIBAND model. MULTIBAND is an extension of the MAXBAND model, which calculates an individual bandwidth for each directional link of the arterial while maintaining main street platoon progression. The individual bandwidth depends on the actual traffic the link carries. By introducing a traffic-dependent capability, which the conventional schemes lack, the model provides additional design flexibility to the traffic engineer as well as improved traffic performance. Because MULTIBAND is based on mixed-integer linear programming, it requires a mathematical programming package similar to the MAXBAND model. Mathematical programming is a formidable optimization tool; however, it is a general purpose tool that can be applied to any mathematical model that has been cast in the required format. Therefore it is not particularly effective for solving the traffic signal synchronization problem per se. The optimization procedure that uses a branch-and-bound algorithm is cumbersome and may take a long time to reach an optimal solution; sometimes the calculation does not converge at all. Recently, attempts have been made to develop heuristic procedures that will speed up the solution process at the expense of achieving suboptimal solutions (*10,11*). As a consequence, one of the principal strengths of the math programming methodology, that of obtaining globally optimal solutions, is being relinquished.

To remedy these limitations, a different approach was used for this study. Instead of using mathematical programming, special-purpose search procedures were developed that are specifically tailored to the arterial synchronization problem and therefore can solve it much more efficiently. Two new simplified models were developed, referred to as U-BAND and V-BAND, to calculate optimal uniform solution and variable bandwidth progression solution, respectively. U-BAND [stands for Uniform Band; i.e., bands of uniform width throughout both directions of the arterial (see Figures 4 and 5)] is based on Little's half-integer optimization algorithm, enhanced by a search procedure for phase-sequence optimization. V-BAND [stands for Variable Band; i.e., continuous bands of variable width along each direction of the arterial (see Figures 6 and 7)] is a further extension of the U-BAND model to calculate variable bandwidth progressions based on the different flow patterns experienced on the individual directional sections of the arterial. The result is a simple and efficient method that is sensitive enough to tai-

H.K. Sripathi, HNTB Corporation, Boston, Mass. 02116. N.H. Gartner and C. Stamatiadis, Department of Civil Engineering, University of Massachusetts, Lowell, Lowell, Mass. 01854.

lor the progression scheme to varying traffic conditions along the arterial street. The development and performance of these models are described in the next section.

UNIFORM BANDWIDTH: THE U-BAND MODEL

The model developed by Little, Martin, and Morgan (*1*) is used as the basic procedure for obtaining a uniform bandwidth. The model finds the optimal offsets that will produce the maximum bandwidth for the simple case of two-phase traffic signals (this restriction is later abandoned). The algorithm uses the half-integer synchronization procedure, in which the middle point of all the intersections' red times are synchronized. Cycle length, signal time splits, traveling speeds on the arterial links, and distances between the intersections are assumed to be known. The middle point of the red time at each intersection is placed in a position to maximize the equal bandwidth in both directions. This position depends mainly on the traveling speed on the links in both directions. If the speeds are different, the position of the middle point of the red time may be placed at any point between zero and the cycle length. For the typical case of equal speeds in both directions, maximum bandwidth will be obtained when the middle point of the red time at each intersection is placed either at the beginning or the middle of the cycle length (hence the term, half-integer synchronization). Thus, the traffic signals of the arterial are synchronized for maximum total bandwidth in both directions by selecting one of the two possibilities. This algorithm will calculate the optimal offsets for equal bandwidths in both directions. If the ratio of the inbound volume to the outbound volume is equal to one, the algorithm will simply give inbound bandwidth equal to outbound bandwidth and the calculated offsets for that bandwidth. If the ratio is different than one, then the total bandwidth is split in proportion to the directional arterial volumes, and the new corresponding offset are calculated.

The preceding algorithm was used as the basis for the U-BAND model, which is further extended to include green split calculations, phase sequence optimization for multiple-phase signalized intersections, arterial progression speed adjustment, and cycle time optimization.

Green Splits

The green splits at each intersection can be calculated as follows. From the input volumes and capacities for the different movements at each intersection, the volume-to-capacity ratios (v/c) for all the movements are calculated. The (v/c) ratio of the main street through inbound movement ($(v/c)_{OML}$) are added. This value is compared with the value obtained by adding the ((v/c) ratio of the main street left outbound movement ($(v/c)_{OMT}$) and left inbound movement ($(v/c)_{IML}$). The maximum of the two values, $(v/c)_M$, is the value that will be used in the green split calculation:

$$(v/c)_M = \max\{[(v/c)_{IMT} + (v/c)_{OML}], [(v/c)_{OMT} + (v/c)_{IML}]\} \quad (1)$$

Similarly, for the cross street $(v/c)_c$ is obtained as the maximum of the (v/c) ratios of the cross street through outbound movement ($(v/c)_{OCT}$) added to the cross street left inbound movement ($(v/c_{ICL}$), and the cross street through inbound movement ($(v/c)_{ICT}$ added to the cross street left outbound movement ($(V/c)_{OCL}$):

$$(v/c)_c = \max\{[(v/c)_{ICT} + (v/c)_{OCL}], [(v/c)_{OCT} + (v/c)_{ICL}]\} \quad (2)$$

The available cycle length is divided between the main street and cross street in proportion to the values of $(v/c)_M$, and $(v/c)_C$. The main street and cross street green times are subsequently divided into through movement and opposing left turning movement proportionally to their (v/c) ratios, and thus the green times for the different movements at each intersection are calculated. The red times for the main street are calculated based on the green times allocated to the cross street and the left turning movement in the opposing direction.

Multi-Phase Sequence Optimization

The four possible phase sequences considered in the model are shown in Figure 1: (*a*) out bound left leads, inbound left lags; (*b*) out bound left lags, inbound left leads; (*c*) out bound left leads, inbound left leads; (*d*) outbound left lags, inbound left lags.

The phase sequence selection is performed after an initial set of offsets and bandwidths has been calculated. This preliminary set is found by Little's half-integer synchronization procedure. The algorithm assumes that inbound red times are equal to outbound red times, which is not applicable in the case of multiple phase signals; therefore, it is used only as the means for establishing the initial settings. Subsequently, the offsets and the bandwidth in one of the two directions (i.e., the outbound direction) are kept constant, whereas offset and bandwidth in the other direction are allowed to vary as the different phase sequences are examined.

In the case of multiple phase sequences, the offsets in the inbound direction vary with respect to the outbound direction offsets, depending only on the green times of the left turning movements. For each phase sequence, the offset in the inbound direction can be calculated based on the outbound offset (Figure 1). Therefore, at each intersection there are four known possible offsets, corresponding to the four different phase sequences:

(*a*) For the first phase sequence, the inbound offset (θ_{IN}) is greater than the outbound offset (θ_{OUT}) by the amount of the left turn green time of the outbound direction (g_{OL}):

$$\theta_{IN} = \theta_{OUT} + R = \theta_{OUT} + g_{OL} \quad (3)$$

(*b*) For the second phase sequence, the inbound offset is less than the outbound offset by an amount equal to the left turning green time of the inbound direction (g_{IL}):

$$\theta_{IN} = \theta_{OUT} + R = \theta_{OUT} - g_{IL} \quad (4)$$

(*c*) For the third phase sequence, the inbound offset is different from the outbound offset by an amount equal to the difference between the green times of the outbound and the inbound directions:

$$\theta_{IN} = \theta_{OUT} + R = \theta_{OUT} + (g_{OL} - g_{IL}) \quad (5)$$

(*d*) For the fourth phase sequence, the inbound offset is equal to the outbound offset:

$$\theta_{IN} = \theta_{OUT} + R = \theta_{OUT} \quad (6)$$

Because there are four possible phase sequences at each intersection, there are 4^n possible combinations of phase sequences on an artery with n intersections. The optimal phase sequence combi-

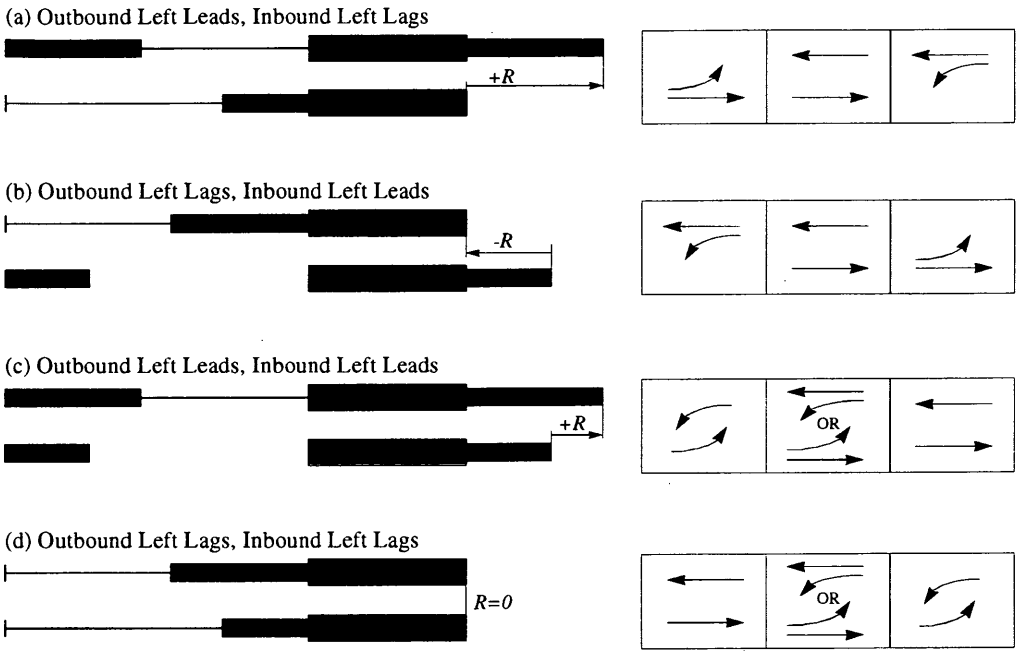


FIGURE 1 Inbound offsets relative to outbound offset for the four different phase sequences.

nation is selected so as to maximize the bandwidth in the inbound direction, through an exhaustive search procedure (Figure 2). Each time a particular phase sequence combination is selected, the inbound bandwidth is recalculated by the NO-OBSTRUCTION technique, described in the next paragraph. This bandwidth is compared with the largest bandwidth found so far from the combinations already examined. If the new bandwidth is larger than the previously largest bandwidth, it replaces the latter; otherwise it is discarded. This procedure is repeated until all phase sequence combinations have been examined.

Inbound-Outbound Bandwidth Calculation

For each phase sequence that is examined, the inbound bandwidth is recalculated with the NO-OBSTRUCTION procedure. Because the offsets and green times are set for each intersection, the procedure involves a simple subtraction of the obstructions to the band at each intersection (Figure 3). The band cannot be greater than the minimum green time, thus the procedure starts from the intersection with the least green time and proceeds in both directions. At this intersection, it is initially assumed that the inbound bandwidth is equal to the green time available for the inbound through movement. The edges of this band are projected to the adjacent intersections with time lags equal to the travel times between the intersections. Travel times between intersections are calculated from the given speeds and distances for each link. If the projected edge does not intersect the red time at the next intersection, the "obstruction" is zero; otherwise, the value of the "obstruction" is calculated as the difference between the point of intersection and the offset at that intersection. The obstruction is then subtracted from the bandwidth and a new bandwidth is calculated, which will be projected to the next intersection. This procedure is repeated until all the intersections are considered, resulting in the adjusted value of the inbound bandwidth.

An adjustment of the previously obtained outbound bandwidth is required to obtain the maximum band for this direction. This is done by adjusting the offsets to the left or right based on the interference values. Since the changes in the offsets should not reduce the inbound bandwidth, the offsets are shifted to the right or left only up to the minimum value of the interferences at that intersection.

Optimization of Cycle Length and Travel Speeds

In the U-BAND model, optimization procedure is repeated for different cycle time values within some specified range. The cycle length is increased by a given increment, the whole procedure is repeated, and the new total bandwidth is recalculated and compared with the best bandwidth from the cycle times examined so far.

Another enhancement to the model is that design speeds on the arterial are allowed to vary slightly. This is achieved by modifying all speeds on the links of the arterial by -1, 0, and $+1$ mph for each cycle time value and selecting the speeds that result in the largest total band.

VARIABLE BANDWIDTHS: THE V-BAND MODEL

The U-BAND model described in the previous section furnishes a uniform bandwidth for the entire arterial for each direction. Therefore, variations in the volumes along the arterial are not taken into

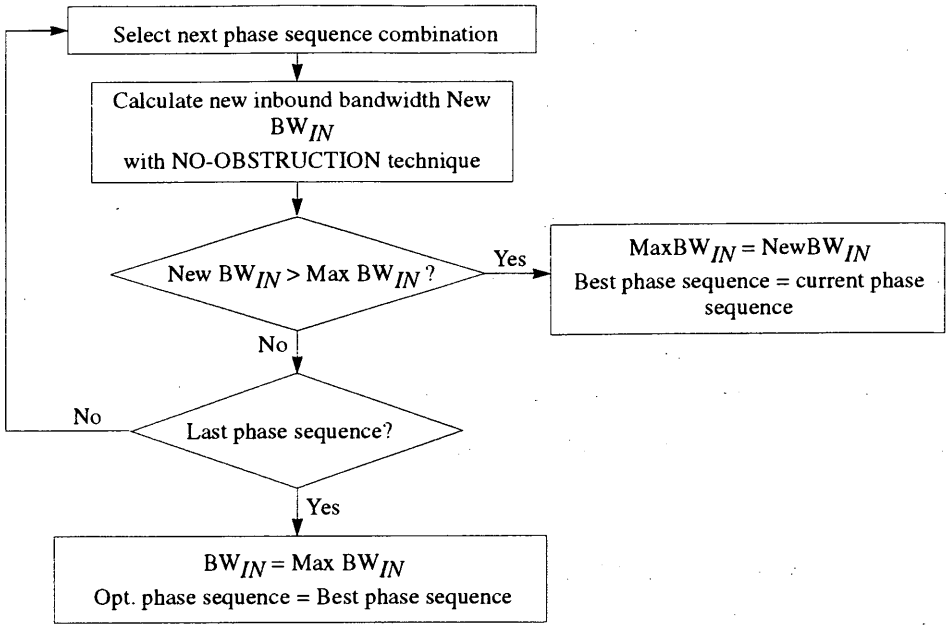


FIGURE 2 Logic for phase sequence optimization procedure.

consideration, which, if significant, may diminish the effectiveness of the bandwidth maximization approach. Because of turn-in and turn-out traffic, such variations in the directional volumes typically exist and must be considered in the model. This is accomplished in the V-BAND model, in which the offset at each intersection is adjusted with the hill-climb search technique to maximize the opportunity for traffic to cross this intersection using the directional green progressions in both directions of the arterial.

In the V-BAND model the total available bandwidth on each link is apportioned to the inbound and outbound directions by giving link-specific weights to the bands that depend on the directional volumes of the link. The link-specific weights that are considered are

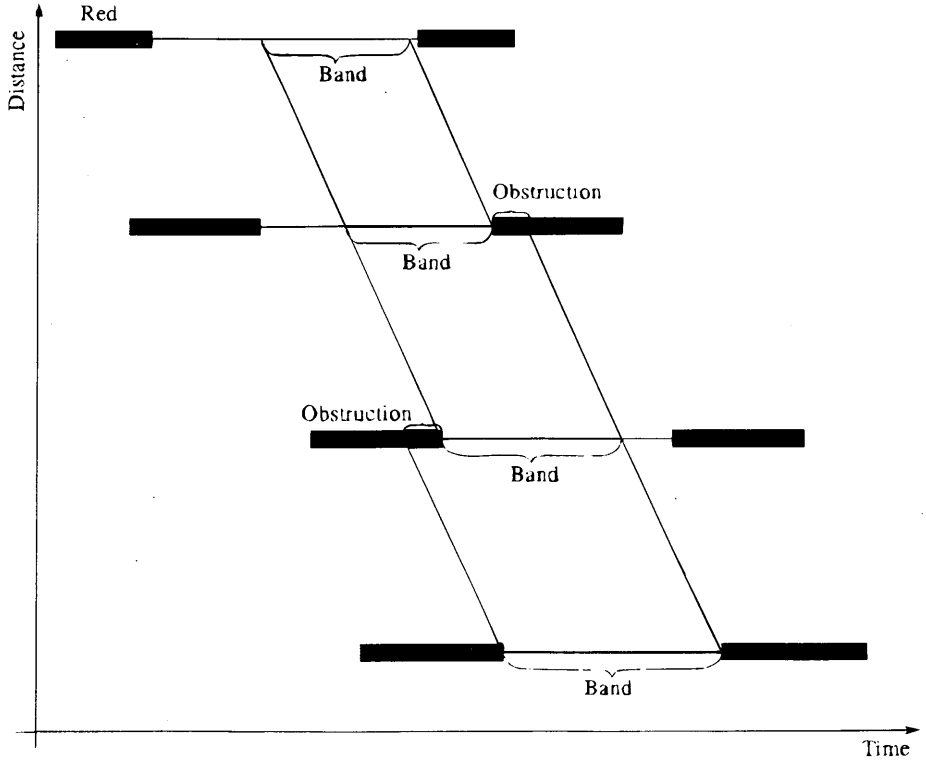


FIGURE 3 NO-OBSTRUCTION procedure for calculating inbound bandwidth.

the volume-to-saturation flow rate ratios of the link (v/s). Therefore, at each intersection the function that must be maximized takes the form:

$$\text{Maximize } Z = b_i(v/s)_i + b_j(v/s)_j + \bar{b}_i\,\overline{(v/s)}_i + \bar{b}_j\,\overline{(v/s)}_j \quad (7)$$

where

b_l = outbound bandwidth of link l,
$\bar{b}_l$ = inbound bandwidth of link l,
i = upstream link,
j = downstream link,
$(v/s)_l$ = (v/s) ratio of the outbound direction of link l, and
$\overline{(v/s)}_l$ = (v/s) ratio of the inbound direction of link l.

Other coefficients also may be used in function Z; that is, on some occasions the (v/s) ratio raised to the fourth power, or exclusion of turning traffic from the (v/s) ratio, has given better results (*9*). The algorithmic approach taken to achieve the objective function described in Equation 7 is to expand or shrink the uniform bands obtained from the U-BAND model on each link symmetrically about the center line of the band. This is called the multiband optimization criterion (*9*).

For this purpose, at each intersection the interference values are calculated. For each link l the interferences of both adjacent intersections are considered, and the minimum value is taken. If this value is equal to 0, the band on link l remains the same; if the value is greater than 0, the band on link l will be increased by up to double the value of the minimum interference to accommodate the increase on both sides of the center line of the band.

At each intersection the offset is then adjusted and the effect of this adjustment on the objective function Z is examined. The initial adjustment depends on the new interference values. If the values of the left interference of the inbound band (w_{il}) or right interference of outbound band (w_{or}) are greater than 0, the offset is shifted to the right by 1 sec. This will increase the outbound bandwidth by 2 sec and reduce the inbound bandwidth by 2 sec. Similarly, if the values of the right interference of the inbound band (w_{ir}) or left interference of outbound band (w_{ol}) are equal to 0, then the offset is shifted to the right by 1 sec, which will increase the inbound bandwidth by 2 sec and reduce the outbound bandwidth by the same amount.

If the new value of Z is smaller than its previous value, the offsets are shifted in the opposite direction; otherwise the offset is shifted in the same direction by one additional second. This process continues until the objective function value decreases, at which point it stops. A constraint that must be satisfied each time an offset adjustment is performed is that the total shift of the offset cannot be greater than the minimum interference in the direction of increase in the bandwidth. This is because the reduction in bandwidth in one direction must be equal to the increase in bandwidth in the other direction. If the offset is moved beyond the minimum interference, the reduction in one direction will not result in increase in the bandwidth in the other direction.

COMPARISON AND EVALUATION

In this section the simplified models previously discussed are compared with their more rigorous brethren. The criterion used in the comparison is the value of the optimization objective. Afterward, simulation is used to evaluate the performance for realistic traffic conditions. To evaluate the effectiveness of the U-BAND model, it is compared with MAXBAND, which can achieve global optimum solutions thus establishing a reliable benchmark. Because the objective of both models is to obtain the maximum bandwidth for a given set of traffic and geometric conditions, the comparison is made in terms of the width of the green band. The test arterials considered in this evaluation include:

1. Canal Street, New Orleans, Louisiana: an arterial street with nine signalized intersections. All intersections have only two phase signal settings;
2. Main Street, Waltham, Massachusetts: an arterial street with nine multiple-phase signalized intersections; and
3. Massachusetts Avenue, Boston, Massachusetts: an arterial street with eight multiple-phase signalized intersections.

MAXBAND can optimize each of the link traveling speeds independently; cycle length is treated as a continuous variable. These features are not currently available in the U-BAND model. Therefore, the U-BAND model was run first, and the optimum traveling speeds and cycle length obtained from these runs were used to set the speeds and the cycle time in MAXBAND. A cycle time of 70 sec and a progression speed of 25 mph were used for all arterials in both models. For all data sets the U-BAND model gave optimal solutions in terms of bandwidth, which are almost identical to the ones obtained from MAXBAND. The phase sequences obtained from the two models for the two arterials with multiple-phase signal settings were not always identical, but it is known that an optimal solution for this type of problem is not unique, and there is a multiplicity of optimal points. The time-space diagrams for Canal and Main streets produced by U-BAND are shown in Figures 4 and 5, respectively.

To evaluate its performance, the V-BAND model was compared with U-BAND, MAXBAND, and MULTIBAND. MULTIBAND (*9*) is an extension of MAXBAND to give link-volume-dependent variable bands, symmetric about their center line. Therefore, like MAXBAND for uniform bandwidth solutions, it can serve as a dependable benchmark for the performance of the U-BAND model. The arterial data sets that were used for this experiment were the ones from Canal and Main streets. For the same reasons as in the previous experiment, the cycle time was set at 70 sec and the progression speeds on all the arterial links were set at 25 mph for all models. The signal settings obtained from the different models were simulated using NETSIM, a microscopic simulation program of traffic in a signalized network. Statistics obtained from NETSIM include average delay per vehicle, average number of stops, average stopped delay per vehicle, and average speed.

The simulation results for these performance measures are shown in Tables 1–4. Table 4 and Table 2 show the performance measures on the arterial streets without taking into consideration the effect of traffic on the side streets, and Table 3 and Table 1 show the same measures with traffic on the side streets. In both cases the advantages of variable bandwidth progression schemes are evident. The V-BAND and MULTIBAND models give better results than the U-BAND and MAXBAND models for both arterials in average delay per vehicle and average stopped delay per vehicle. For example, average delay is reduced by 10 and 11 percent by V-BAND and MULTIBAND, respectively, compared with the MAXBAND results for Canal Street, and average stopped delay is reduced by as much as 13 percent by both models for the same arterial street example.

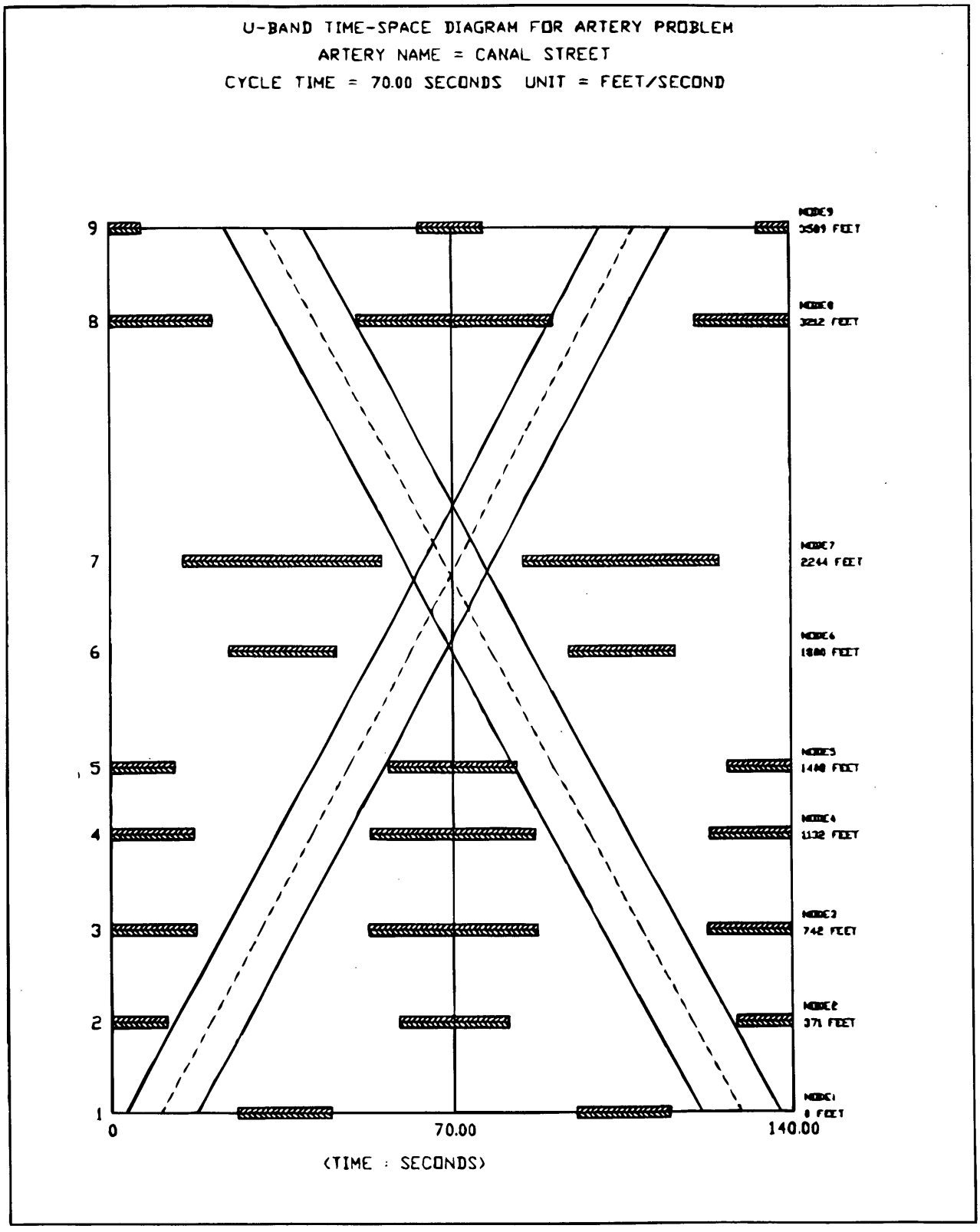


FIGURE 4 Time-space diagram for Canal Street; U-BAND model.

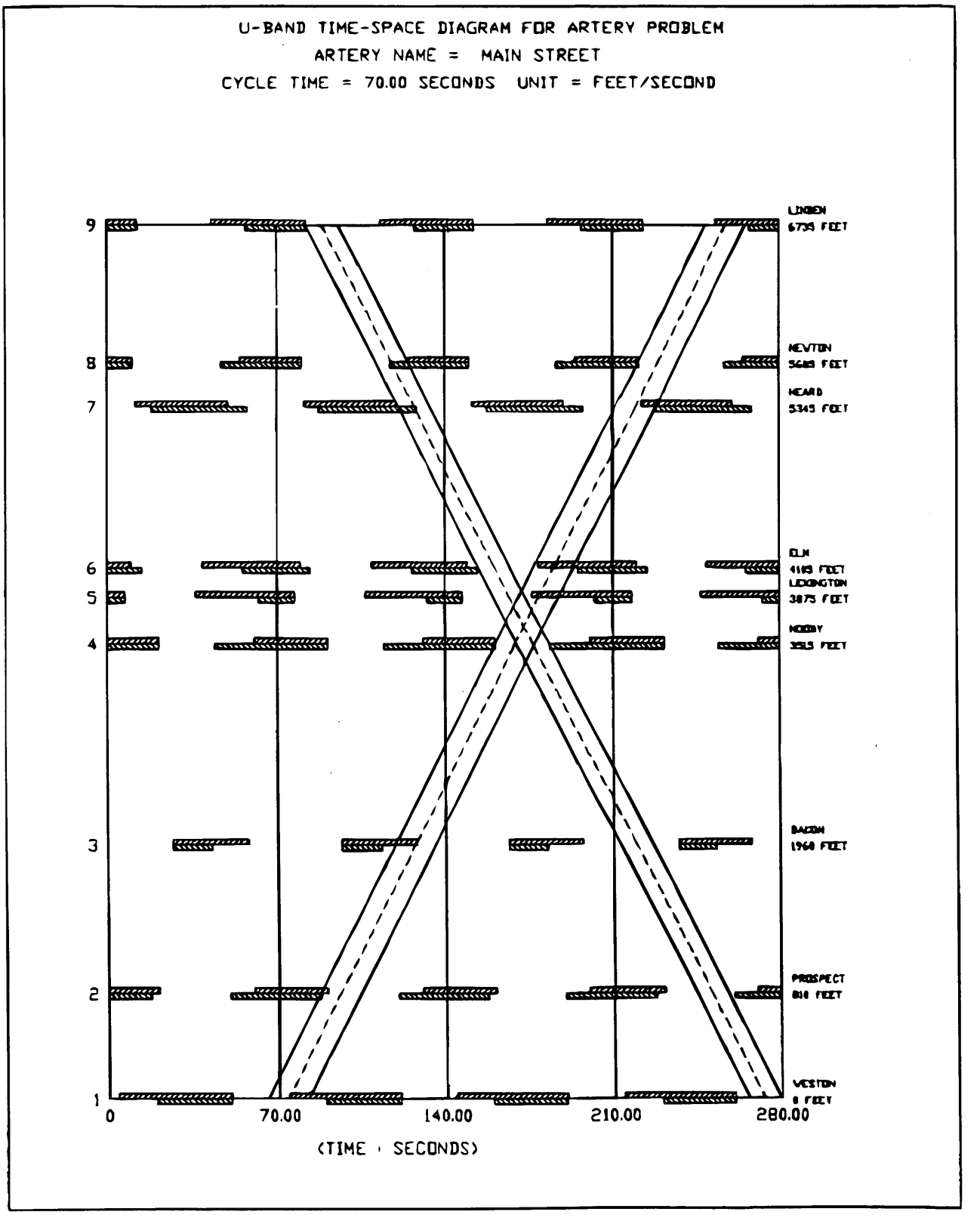


FIGURE 5 Time-space diagram for Main Street; U-BAND model.

TABLE 1 NETSIM Simulation Results: Main Street (With Side Streets)

	Avg. Delay (sec./veh.)	Avg. Stopped Delay (sec./veh.)	Avg. % of Stops	Avg. Speed (mph)
U-BAND	30.46	21.83	69.45	11.08
MAXBAND	29.42	20.85	69.2	11.29
V-BAND	27.75	19.16	68.97	11.22
MULTIBAND	27.3	18.72	67.10	11.38

TABLE 2 NETSIM Simulation Results: Main Street (Without Side Streets)

	Avg. Delay (sec./veh.)	Avg. Stopped Delay (sec./veh.)	Avg.% of Stops	Avg. Speed (mph)
U-BAND	33.09	24.08	69.09	10.58
MAXBAND	31.79	22.84	68.77	10.83
V-BAND	29.57	20.61	68.53	10.77
MULTIBAND	29.51	20.49	66.7	10.84

TABLE 3 NETSIM Simulation Results: Canal Street (With Side Streets)

	Avg. Delay (sec./veh.)	Avg. Stopped Delay (sec./veh.)	Avg.% of Stops	Avg. Speed (mph)
U-BAND	26.07	14.08	55.13	10.64
MAXBAND	23.28	14.53	57.91	10.34
V-BAND	21.23	12.85	55.84	10.51
MULTIBAND	20.89	12.75	55.29	10.63

TABLE 4 NETSIM Simulation Results: Canal Street (Without Side Streets)

	Avg. Delay (sec./veh.)	Avg. Stopped Delay (sec./veh.)	Avg.% of Stops	Avg. Speed (mph)
U-BAND	27.75	13.90	50.48	11.54
MAXBAND	23.82	14.57	54.08	11.11
V-BAND	20.80	12.16	51.66	11.39
MULTIBAND	20.58	12.21	50.51	11.48

The improvements in delay are even more pronounced when only the main street traffic is considered. For example, for Canal Street the average delay is improved by 14.5 and 16 percent by V-BAND and MULTIBAND, respectively, over the MAXBAND results, and average stopped delay is reduced by as much as 19 percent by both models. There are also some improvements in the average number of stops when the average traveling speed is approximately the same for all models.

The results also show that the new simplified V-BAND model performs in a way similar to the more sophisticated MULTIBAND model in terms of delays, number of stops, and average speed. Hence, it may be concluded that V-BAND obtains results that are virtually identical to MULTIBAND. The time-space diagrams for Canal and Main streets produced by the V-BAND model are shown in Figures 6 and 7, respectively.

CONCLUSIONS

A simplified and efficient method to calculate variable-bandwidth progressions given optimized uniform bandwidth progressions is presented. Little's half-integer optimization algorithm was used as the basic tool for the uniform bandwidth maximization, coupled with a combinatorial phase-sequence optimization procedure to develop the U-BAND model. The method was then extended to the V-BAND model to calculate variable-bandwidth progressions

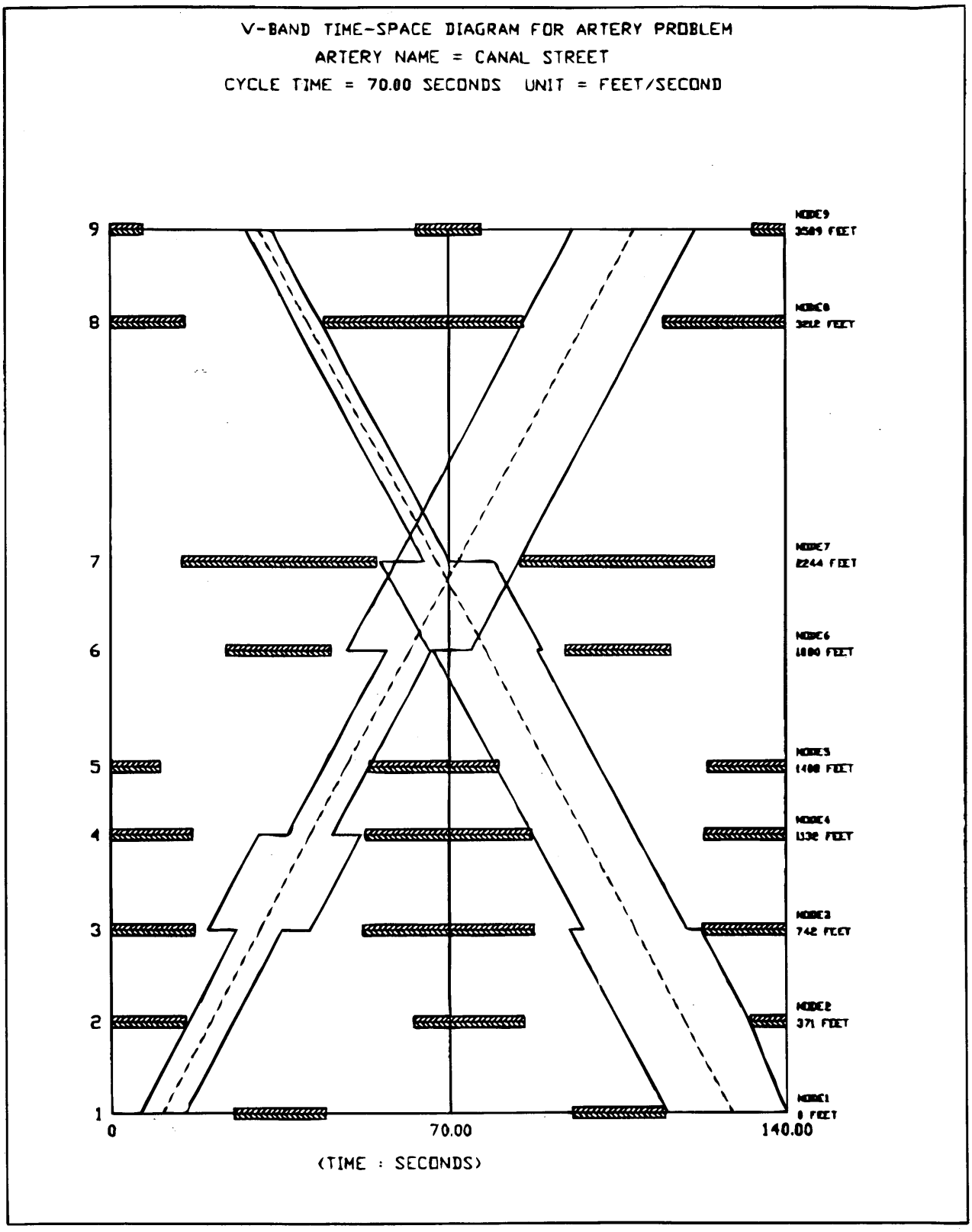


FIGURE 6 Time-space diagram for Canal Street; V-BAND model.

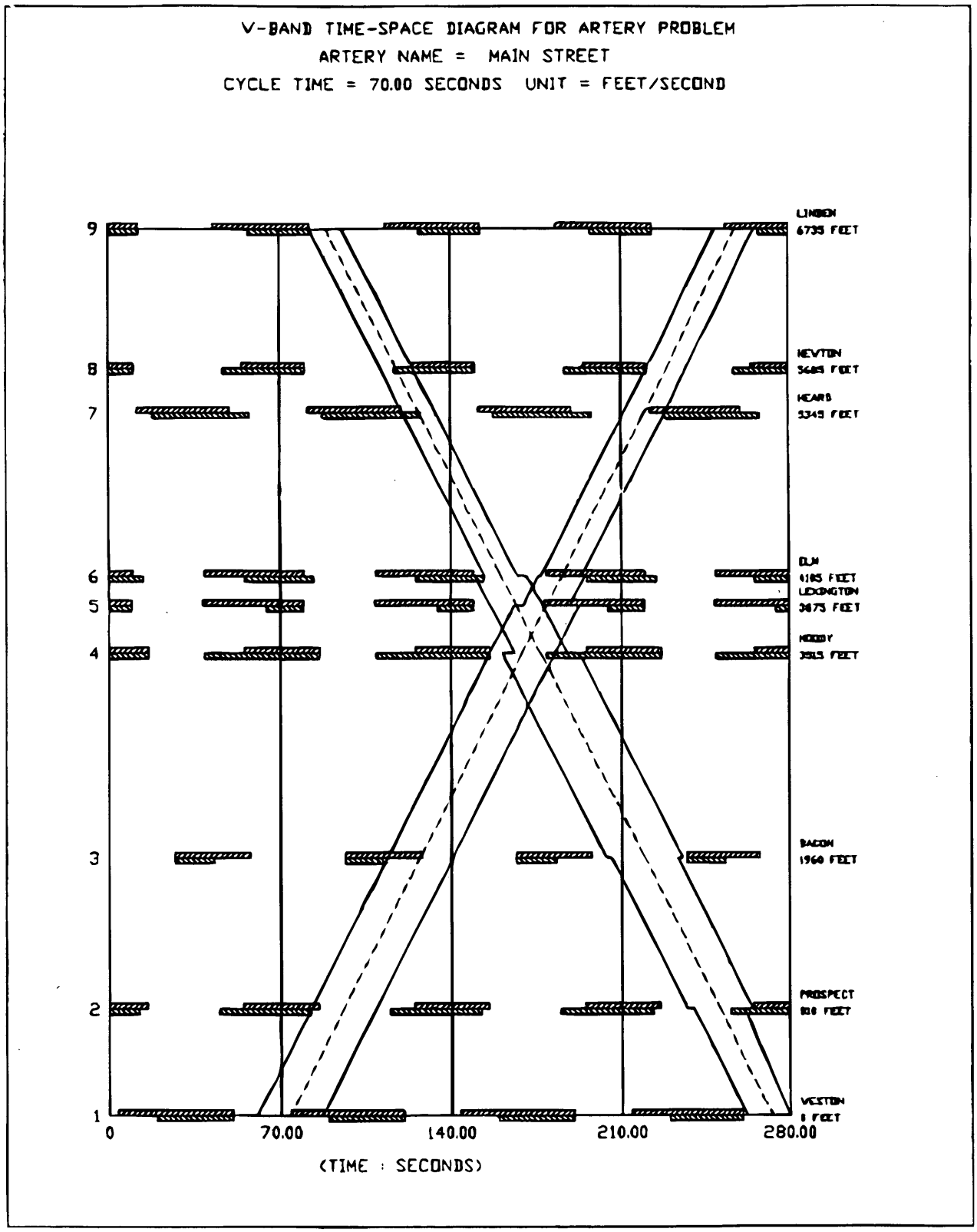


FIGURE 7 Time-space diagram for Main Street; V-BAND model.

using the multiband optimization criterion. An important feature of this method is that it can be applied to any arterial synchronization scheme after the uniform bandwidth progression has been optimized.

Several arterial examples were given to illustrate the effectiveness of the U-BAND and V-BAND models. The results from the V-BAND model were simulated with NETSIM and it was demonstrated that significant benefits can be obtained in traffic performance. An important aspect of this approach is that near-optimal bandwidth progressions can be obtained without sophisticated and cumbersome mathematical programming tools. Further research is under way to extend the approach described in this study to networks of arterials where it is likely to have comparable beneficial effects. Finally, this approach is expected to (*a*) provide advantages compared with established models such as PASSER-II and (*b*) compare favorably with recent versions of TRANSYT.

REFERENCES

1. Little, J. D. C., B. V. Martin, and J. T. Morgan. Synchronizing Traffic Signals For Maximal Bandwidth. *Highway Research Record 118,* 1967, pp. 21–47.
2. Brooks, W. D. Vehicular Traffic Control—Designing Arterial Progressions Using a Digital Computer IBM (undated).
3. Bleyl, R. L. A Practical Computer Program for Designing traffic Signal Timing Plans. *Highway Research Record 211,* 1967, pp. 19–33.
4. Leuthardt, H. R. Design of a Progressively Timed Signal System. *Traffic Engineering,* No. 45, 1974.
5. Messer, C. J., R. H. Whitson, C. L. Dudek, and E. J. Romano. A Variable-Sequence Multi Phase Progression Optimization Program. *Highway Research Record 445,* 1973, pp. 24–33.
6. Little, J. D. C., M. D. Kelson, and N. H. Gartner. MAXBAND: A Program for Setting Signals on Arteries and Triangular Networks. *Transportation Research Record 795,* 1981, pp. 40–46.
7. Chang, E. C-P, S. L. Cohen, C. Liu, N. A. Chaudhary, and C. Messer. MAXBAND-86: A Program for Optimizing Left-Turn Phase Sequence in Multiarterial Closed Networks. *Transportation Research Record 1181,* 1988, pp. 61–67.
8. Tsay, H. S., and L. T. Lin. New Algorithm for Solving the Maximum Progression Bandwidth. *Transportation Research Record 1194,* 1988, pp. 15–30.
9. Gartner, N. H., S. F. Assmann, F. Lasaga, and D. L. Hou. MULTIBAND—A Variable-Bandwidth Arterial Progression Scheme. *Transportation Research Record 1287,* 1990, pp. 212–222.
10. Chaudhary, N. A., A. Pinnoi, and C. J. Messer. Proposed Enhancements to MAXBAND-86 Program. *Transportation Research Record 1324,* 1991, pp. 98–104.
11. Solanki, R. S., and R. S. Pillai, and A. K. Rathi. *A Fast Heuristic for Maximizing Bandwidth in Traffic Network,* Oak Ridge National Lab Report, 1993.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Bus-Preemption Under Adaptive Signal Control Environments

GANG-LEN CHANG, MEENAKSHY VASUDEVAN, AND CHIH-CHIANG SU

To explore the advantages of integrating bus preemption and adaptive signal control, an integrated model for adaptive bus-preemption control in the absence of automated vehicle location systems was developed. In the proposed system, unconditional priority is not given to buses over passenger cars. Instead of using pre-specified strategies such as phase extension, phase early start, or special bus phase, preemption decision is based on a performance index, which includes vehicle delay, bus schedule delay, and passenger delay. An extensive simulation evaluation with respect to the integration of adaptive control with preemption is also presented. The developed model displays promising results.

Finding ways to relieve traffic congestion has long been a priority of transportation and traffic engineers. While advanced traffic management systems (ATMS) and advanced traveler information systems (ATIS) have alleviated some of the problems, these methods alone are not enough. New approaches are vital as the demand on transit systems continues to grow. Hence, to substantially improve urban traffic conditions, effective strategies are needed from both demand and supply sides. Preferential treatment for buses such as signal preemption, devised to encourage the use of public transit systems, is one of the latest demand-side strategies for relieving urban congestion. Since adaptive signal control is one of the latest supply-side methods for relieving urban traffic congestion, integrating the two methods is essential.

Over the past several decades, several studies related to bus-preemption strategies have been conducted, involving experimental testings (*1–11*) and analytical explorations (*12–15*). Some of the transit preemption methods have been implemented in the existing signal systems, such as UTCS/BPS (*16*), UTOPIA (*17*), SCRAM (*18*), and SPPORT (*19,20*). Overall, the potential benefits of properly designed and implemented bus-preemption strategies have been well-justified in these studies.

Because preemption strategies traditionally favor bus users over passenger-car drivers, their implementation is a sensitive issue and has often prompted debate. Therefore, a rigorous evaluation of the trade-offs and complex interactions between transit users and passenger-car users under various traffic conditions is necessary before any strategy can be successfully developed and applied. Although a review of the literature shows that considerable progress has been made, future research should address the following issues:

- Integration of bus preemption with adaptive signal control to ensure that the optimal signal control minimizes not only vehicle delay, but also passenger delay. Most existing studies on bus preemption, with the exception of UTOPIA (*17*), did not operate under acyclic adaptive signal control systems.
- Evaluation of the bus-preemption need from transit system management perspectives. For instance, in comparing the trade-offs between competing signal plans, the status of an approaching bus, either ahead of or behind its schedule, should be considered, along with its loading factors.
- Incorporation of information from automated vehicle location (AVL) systems in the design of bus preemption and adaptive signal control.

G. L. Chang and M. Vasudevan, Department of Civil Engineering, University of Maryland, College Park, Md 20742. C.-C. Su, Central Police University, Taipei, Taiwan.

In this study, the first two issues are discussed with an integrated adaptive system for bus-preemption and signal control. The incorporation of AVL information and its impact on the systems effectiveness will be presented elsewhere (Chang et al., unpublished data). This discussion includes:

1. A description of the proposed adaptive preemption system for intersection control, along with its principal modules and their interrelations.
2. A detailed presentation of the logic and mathematical formulations for each primary system module.
3. An experimental plan for assessing the effectiveness of the proposed system under various traffic conditions, and the results of evaluation.

AN INTEGRATED SYSTEM FOR BUS PREEMPTION AND ADAPTIVE SIGNAL CONTROL

To execute bus preemption effectively in an adaptive signal control environment, the control algorithm should:

- Incorporate bus preemption as one of the adaptive signal control functions;
- Use an adaptive control logic with real-time algorithms instead of using pre-specified strategies, such as phase extension, phase early start, or a special bus phase;
- Impose a minimum green constraint and automatically update it after every switchover decision, based on the existing traffic conditions and driver safety; and
- Have a performance function for system evaluation, based on the current queue length, bus loading factors, and bus schedule delay.

Figure 1 presents the relationship between all principal components of the proposed integrated system, including the bus-preemption module and other local adaptive control components. The integration of these modules enables the system to provide a preventive adaptive control every 3 sec based on the detected real-time

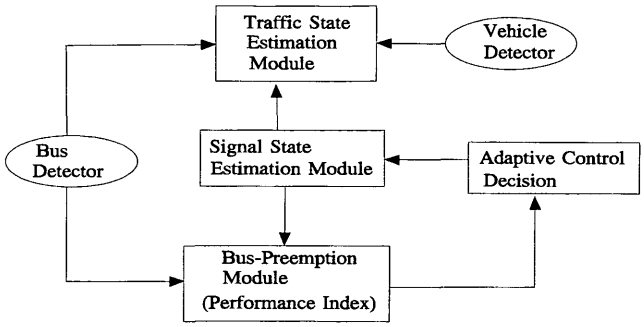


FIGURE 1 The relationship between principal modules of the proposed adaptive control system with bus preemption.

demand. The interaction between the local optimization module and the bus-preemption module will allow the system to operate under both, with and without bus arrival conditions.

In the operating process, shown in Figure 2, the real-time arrival information is provided by detectors for both private vehicles and buses, and is used to estimate the arrival and discharge of flows and the queue lengths at the current time step (in the traffic state estimation module) based on the existing signal state. The estimated traffic state is used to determine the optimal adaptive control strategy with the detected traffic conditions. If any bus has been identified by the surveillance system, then the benefit of offering bus-preemption status is evaluated while making a control decision. The resulting decision is used to calculate the signal state elements in the signal state estimation module for the succeeding time step. The major function of each module is described in the next section.

The proposed adaptive signal control module does not rely on prediction models for arriving traffic over the entire time horizon, and thus is myopic in nature. A modified version of an integrated system that uses AVL information and a neural network model for prediction has been presented elsewhere (Chang et al., unpublished data). The notations used in this discussion are given in Table 1.

Surveillance Systems

The operation of the adaptive control system requires:

- Vehicle detectors placed at the location of 36.6 m (120 ft) per lane from the stop line for estimating queue length and 15.25 m (50 ft) per lane from the upstream intersection for estimating the arrivals when the downstream detectors are occupied; and
- Bus detectors placed at the location of 36.6 m (120 ft) per lane from the stop line for reducing the uncertainty of a bus arrival due to additional delay in loading/unloading, lane-changing behaviors, curb parking or turning movements; and the stop line (per lane) for detecting bus departures.

Traffic State Estimation Module

The estimation of traffic conditions for signal optimization or bus preemption involves determining (*a*) current queue length, (*b*) expected demand, and (*c*) anticipated discharged flow. Computation of queue length is vital for the execution of the bus-preemption function. It is one of the key factors in making a signal control decision, as it is critical in determining the allowable minimum green duration. Hence, in this module, data supplied by the detectors are used to estimate the arrival and discharge of flows, and consequently, the queue lengths for each time step. The estimated queue length is used in the Performance Index (*PI*) module. A simple queue estimation concept, shown in Equation *1*, is used to estimate the short-term queue length at the target intersection.

Queue Length Estimation

$$Q_l^i(k+1) = Max\{Q_l^i(k) + A_l^i(k+1) - d_l^i(k+1), 0\} \quad (1)$$
$$\forall\, l \in P^i;\ \forall\, P^i \in i;\ \forall\, i \in H$$

The queue length at a given time step is computed from (*a*) the queue length of the previous time step, (*b*) the number of new

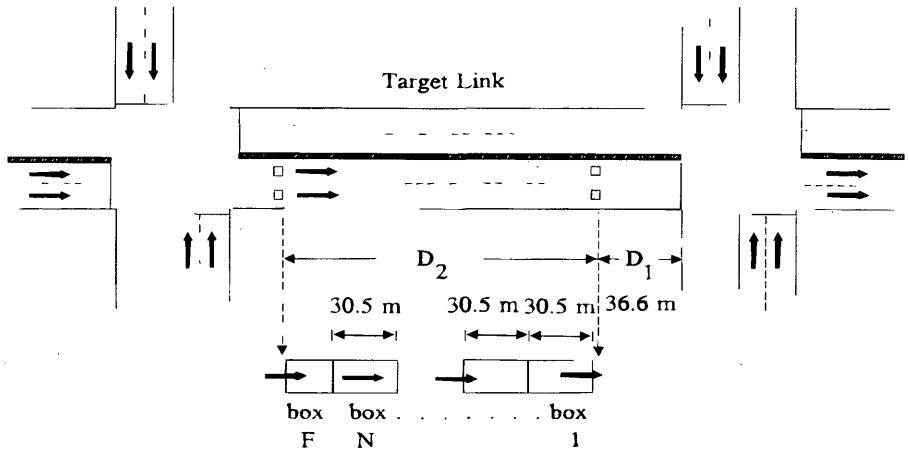


FIGURE 2 The relation between detector placement and arrival estimation.

TABLE 1 Notations Used in the Study

T	:	Duration of a time step (seconds)
H	:	Set of signal phases at the control intersection
P^i	:	Set of lane groups in phase i
$\phi^i(k)$	:	0 if signal state is green for phase i and time step k 1 if signal state is red for phase i and time step k
$\xi^i(k)$	:	1 if existing signal state of phase i is switched at end of time step k 0 otherwise
G^i_{min}	:	Minimum green for phase i (seconds)
G^i_{max}	:	Maximum green for phase i (seconds)
$U^i(k)$	:	Green time used by phase i, at the end of time step k (seconds)
Y^i	:	Yellow time for phase i (seconds)
AR^i	:	All red time for phase i (seconds)
$R^i(k)$	:	Minimum waiting time for a green for phase i (green phase of time step k), if a switchover occurs at the end of time step k
$S^i_{l,g}$	:	Saturation flow rate for green time, for lane l, in phase i
$S^i_{l,g}$	:	Saturation flow rate for yellow time, for lane l, in phase i
q^s_l	:	Given saturation flow rate for lane l
$q_{l,u}(k)$	:	Traffic flow of lane l, detected by upstream detector, u, at time step k, for phase i
$q^i_{l,d}(k)$	:	Traffice flow of lane l, detected by downstream detector, d, at time step k, for phase i
$A^i_l(k)$	:	Number of arrivals in lane l, provided by downstream detectors at time step k, and phase i (vehicles)
$d^i_l(k)$	:	Discharged flow of lane l at time step k, and phase i (vehicles)
$Q^i_l(k)$	:	Estimated queue length of lane l, at time step k, and phase i (vehicles)
D_1	:	Distance between the downstream detector located at 36.6 m (120 ft) and the stop line (m)
D_2	:	Distance between the upstream detector and the downstream detectors (m)
$a^i_{l,1}(k)$	:	Number of arrivals in lane l moving in Box l from the upstream detectors at time step k, for phase i (vehicles) (see Figure 2)
$a^i_{l,N}(k)$	:	Number of arrivals in lane l, moving in box N, fro the upstream detectors at time step k, and phase i (vehicles) (see Figure 2)
N	:	Number of integer boxes from the upstream detectors to the downstream detectors (see Figure 2)
F	:	Fractional part of the box, closest to the upstream detector, which may not be accommodated within D_2 (Figure 2)
L_v	:	Average vehicle length (m)
S_d	:	Distance between the rear of a vehicle and the front of the following stopped vehicle (m)
t^p_{sd}	:	Starting delay for a passenger car (seconds)
t^b_{sd}	:	Starting delay for a bus (seconds)
n_p	:	Average number of passengers in a passenger car
$PQ^i_l(k)$	:	Estimated passenger car queue length of lane l, at time step k, for phase i (vehicles)
$B^i_l(k)$	:	Number of detected buses in lane l, at time step k, and phase i that have not yet cleared the intersection (vehicles)
$P^j_{j,l}(k)$	:	Number of passengers in bus j, in lane l, at time step k, for phase i
$SD^j_{j,l}(k)$	:	Schedule delay of bus j, in lane l, at time step k, and for phase i
$D^j_{j,l}(k)$	:	Total delay of bus j, in lane l, and phase i' (red phase), if green is extended at the end of time step k for the green phase, i
$F^j_{j,l}(k)$	:	Number of vehicles detected ahead of bus j, in lane l, at time step k, and phase i

arrivals, and (c) the discharged flow. However, when the queue length is calculated for a red approach, the discharged flow term in Equation 1 is reduced to zero. The equation is used to determine passenger car and bus queue lengths. $A^i_l(k + 1)$ and $d^i_l(k + 1)$ are estimated from real-time surveillance data and signal control states.

Estimation of Arrivals

Depending on whether the downstream detectors are occupied by the queued vehicles, the system uses either Equation 2 or Equation 3.

$$A^i_l(k) = q^i_{l,d}(k - 1) \text{ if } Q^i_l(k) \leq D_1 \quad (2)$$

$$A^i_l(k) = a^i_{l,1}(k) \text{ if } D_2 \geq Q^i_l(k) \geq D_1 \quad (3)$$

$q^i_{l,d}(k - 1)$ is measured in real time from the downstream detectors, while $a^i_{l,1}(k)$ is estimated from the upstream detector information, based on the following modified PRODYN (*21*) concept:

$$a^i_{l,1}(k) = a^i_{l,2}(k - 1) \quad (4)$$

$$a^i_{l,(N-1)}(k) = a^i_{l,N}(k - 1) + (1 - F)\, q^i_{l,u}(k - 1) \quad (5)$$

$$a^i_{l,N}(k) = F q^i_{l,u}(k - 1) \quad (6)$$

Estimation of Discharged Flows

The discharged flow $d^i_l(k)$ in a control phase i depends on the adaptive control decision and the signal control state (i.e., the green, yellow, and red duration). It can be approximated with the following equation:

$$d^i_l(k) = (1 - \phi^i(k))[S^i_{l,g}(1 - \xi^i(k)) + S^i_{l,y}\xi^i(k)] + S^i_{l,g}\xi^i(k)\phi^i(k) \quad (7)$$

Depending on the signal state (red or green) and the control decision, the discharged flow becomes equal to the saturation flow rate for green or yellow time. For example, when the signal state is green ($\phi^i(k) = 0$) and the control decision is to switch the green ($\xi^i(k) = 1$), then the discharged flow is equal to the saturation flow rate for yellow.

Signal State Estimation Module

This module monitors the signal state, computes the elapsed green time, and estimates the minimum green duration in real time. The logic for all its functions is given in the next section.

Signal State

The signal state of any phase i, $\phi^i(k)$ at time step k is given by (*21*)

$$\phi^i(k) = \xi^i(k - 1) + \phi^i(k - 1) - 2\,\xi^i(k - 1)\,\phi^i(k - 1) \forall i \in H \quad (8)$$

The first term represents the control decision at the end of time step $k-1$. The second term signifies the signal state of phase i at time step $k-1$. $\phi^i(k)$ is a binary variable. If the signal state is red for time step $k-1$, (i.e., $\phi^i(k-1)=1$) and the control decision at the end of the time step is to switchover ($\xi^i(k-1)=1$), then the signal state for time step k from Equation 8 must be 0, which corresponds to a green state.

Elapsed Green

The green time already used up by phase i at time step k is computed with the following equation (*21*):

$$U^i(k) = (U^i(k-1) + T)\,(1 - \xi^i(k-1))\ \forall\, i \in H \tag{9}$$

Based on the control decision, $\xi^i(k-1)$, green time is either increased by a duration of T seconds, or it is reduced to zero.

Minimum Green

Minimum green is recommended to be the shortest green time during which drivers can be expected to react safely to signal changes. It also must be sufficiently long for discharging the average waiting queue during each control phase i. A mathematical representation of such a requirement is given as

$$G^i_{\min} = t^p_{sd} + \left(\text{Max}\left\{\left(\frac{D_l}{L_v + S_d} + 1\right), \underset{l}{\text{Avg}}\, Q^i_l(k)\right\}\right)\left(\frac{3{,}600}{q^l_s}\right) \tag{10}$$
$$\forall\, l \in P^i, \forall\, P^i, \forall\, i \in H$$

Thus, the minimum green ($G^i_{\min}$) for phase i is made up of the following components:

- Starting delay, t^p_{sd}, due to switching of signals, and
- The maximum of the two expressions, for safely discharging the average queue length: first denotes the number of vehicles that will occupy the length D_l, and second indicates the average queue length for all lanes in phase i at time step k.

Maximum Green

A sufficiently long green can be set so the control algorithm can effectively handle oversaturated conditions. It also can be set by the user to respond to demand variations during different periods, such as morning peak, evening peak, day off-peak, night off-peak, and holidays.

Bus-Preemption Module

A review of the literature shows that most adaptive control strategies do not consider the delay in the schedule of a bus while making a signal-state decision for bus preemption. Hence, the decision to switchover to another phase or not may not be an optimal solution. This can be rectified by computing a PI that evaluates the effect of the decision. With this in mind, a *PI* model, allowing for measuring the benefit of the control decision and based on passenger delay ($C_{pd}{}^{i'}$), vehicle delay ($C_{oc}{}^{i'}$) and schedule delay($C_{sd}{}^{i'}$) is formulated in this section.

In a multiphase control intersection, the *PI* value should be computed based on the sum of $PI^{i'}$ for each competing phase i', of phase i, in set H.

$$PI = \sum_{i' \in H} PI^{i'} \tag{11}$$

Each $PI^{i'}$ is the sum of the trade-offs due to the signal control decision in $C_{pd}{}^{i'}$, $C_{oc}{}^{i'}$, and $C_{sd}{}^{i'}$.

$$PI^{i'} = C^{i'}_{pd} + C^{i'}_{oc} + C^{i'}_{sd} \quad \forall\, i' \neq i,\ i' \in H \tag{12}$$

In this module the benefit of giving a green is compare with that of terminating it by computing the trade-offs incurred in passenger, vehicle, and schedule delays. The following equations do not reflect the actual passenger, vehicle, and schedule delays.

Computation of Passenger Delay

$$C^{i'}_{pd} = R^{i'}(k) \sum_{\forall l} \left[n_p\, PQ^i_l(k) + \sum_{j=1}^{B^i_l(k)} P^i_{j,l}(k) \right]$$
$$- T \sum_{\forall l} \left[n_p\, PQ^{i'}_l(k) + \sum_{j=1}^{B^{i'}_l(k)} P^{i'}_{j,l}(k) \right] \tag{13}$$

The minimum waiting time for a green for phase i if a switchover occurs is given by

$$R^{i'}(k) = Y^i + AR^i + G^{i'}_{\min} \tag{14}$$

The computation of the total passenger delay in Equation 13 varies with the following scenarios:

- The first term considers the delay of passengers in the green approach resulting from a switchover. If the current green is terminated, then the passengers in the terminated green phase will have to wait for a duration equal to the minimum green time needed for the previous red phase to compete for a switchover.
- If green is extended for phase i by another time step (i.e., for T seconds), then passengers of vehicles in the waiting queue of the competing phase (red phase, i') will suffer an additional delay of T seconds. This is expressed in the second term.

Computation of Vehicle Delay

$$C^{i'}_{oc} = \left[t^p_{sd} \sum_{\forall l} PQ^i_l(k) + t^b_{sd} \sum_{\forall l} B^i_l(k) \right]$$
$$- \left[t^p_{sd} \sum_{\forall l} PQ^{i'}_l(k) + t^b_{sd} \sum_{\forall l} B^{i'}_l(k) \right] \tag{15}$$

The first term expresses the delay of vehicles in queue in the current green phase, i, if their green is terminated. If green is extended

for the current green, then vehicles in the current red phase, $i,'$ will encounter a delay as given in the second term.

Computation of Schedule Delay

$$C_{sd}^{i'} = \sum_{\forall l} \sum_{\forall j} D_{j,l}^{i}(k) - \sum_{\forall l} \sum_{\forall j} D_{j,l}^{i'}(k) \quad (16)$$

The first term denotes the delay of buses in the green approach if their green is terminated, and the second term gives the delay of buses in the red approach if green is extended. If a bus in the green phase is experiencing a delay in schedule, $SD^{i}_{j,l}(k-1)$, when detected, terminating green will result in a delay of $D^{i}_{j,l}(k)$, which is given by

$$D_{j,l}^{i}(k) = R^{i'}(k) + t_{sd}^{b} + (F_{j,l}^{i}(k) - d_{l}^{i}(k) + 1)\frac{3{,}600}{q_s^l} + SD_{j,l}^{i}(k-1) \quad (17)$$

Terminating green at the end of time step k will result in an additional delay caused by

- Minimum waiting time for a green for phase i if a switchover occurs (first term),
- Starting delay, t^{b}_{sd}, for the bus (second term), and
- Time taken to discharge the number of vehicles ahead of the bus, which did not clear the intersection before the end of green.

However, a bus in the red approach will suffer an additional delay due to the extension of green by T seconds. Thus, the total delay of bus j at current red phase i' can be computed with the following equation:

$$D_{j,l}^{i'}(k) = T + t_{sd}^{b} + (F_{j,l}^{i'}(k) - d_{l}^{i'}(k) + 1)\frac{3{,}600}{q_s^l} + SD_{j,l}^{i'}(k-1) \quad (18)$$

Note that the above $PI^{i'}$ should be computed for every competing phase i', of current green phase i, in H. The net PI is the sum of all $PI^{i'}$ If PI is negative, then the optimal decision, with bus-preemption control, is not favorable to the intersection. Hence, it should be changed. If $PI \geq 0$, then the current green should be extended by T seconds.

SYSTEM CONTROL LOGIC

This section deals with the basic control strategy governing the proposed model for adaptive control with bus preemption. Given the aforementioned system and all the functions of its key elements, the operational procedures may be summarized as

Step 1. At time step k and phase i, the system computes the minimum and maximum green times.

Step 2. Checks the minimum and maximum green constraints:

Condition 1: If green time is less than the minimum green time, then the system extends the green ($\xi^{i}(k) = 0$). $U^{i}(k)$ is updated.

Condition 2: If $U^{i}(k)$, the green time used by phase i at time step k, is greater than $G^{i}_{\max}$, then green is terminated immediately. Both parameters $U^{i}(k)$ and $G^{i}_{\min}$ are updated.

Condition 3: If both conditions are satisfied, then the system proceeds to Step 3.

Step 3. Examines bus presence using the bus detectors. If no bus is present, then the number of passengers, $P^{i}_{lj}(k)$, is reduced to zero. Otherwise, it provides all bus presence information.

Step 4. Computes the net benefit of extending green with the proposed PI function.

Step 5. If PI is negative, then the optimal decision is not favorable to the intersection and a switchover decision is taken. Otherwise, it extends the current green by another T seconds.

In the proposed model, the control decision is made every 3 sec depending on a comparison of the benefits of extending green or terminating it. The control logic uses real-time traffic state conditions instead of pre-stipulated strategies. It is assumed in the logic that no bus stop is located between the 36.6 m (120 ft) detector and the stop line. The adopted control strategy is illustrated with a flow chart in Figure 3.

SAMPLE APPLICATION

This section presents a sample application of the proposed system and evaluates its effectiveness under various traffic conditions. All traffic flow-related data for use in the proposed algorithm were generated with TRAF-NETSIM. The key features of all simulated scenarios and evaluation plans are summarized in the next section.

Simulation Experiment

The network considered had a link length of 305 m (100 ft) with 2 lanes in each direction and a bus stop 183 m (600 ft) from the stop line. There was no bus bay. To facilitate the functioning of the proposed system, the surveillance environment included a stop line detector and detectors at 36.6 m (120 ft) and 289.75 m (950 ft) per lane for each direction. Signal control operations were designed with a two-phase actuated control, permitted left turns, minimum green of duration 15 sec, maximum green of 60 sec, and a yellow of 3 sec.

Two bus route arrivals were simulated for northbound and southbound approaches and one each for east- and westbound approaches. The experimental data were collected for 10 min after the initialization period. The proposed model was tested for 90 time intervals, each of duration 3 sec. The traffic volume varied as 300 vphpl, 500 vphpl, and 1000 vphpl. The mean discharge headways of buses were taken as 180 sec (20 buses/hr) and 120 sec (30 buses/hr).

The layout of the experimental intersection is given in Figure 4. The traffic variables were collected only to provide a meaningful data set for evaluating the performance of the control logic. Because the purpose of the experiment was to test the model, the entering traffic volume was taken as a constant. The algorithm used the following traffic measurements from NETSIM's output:

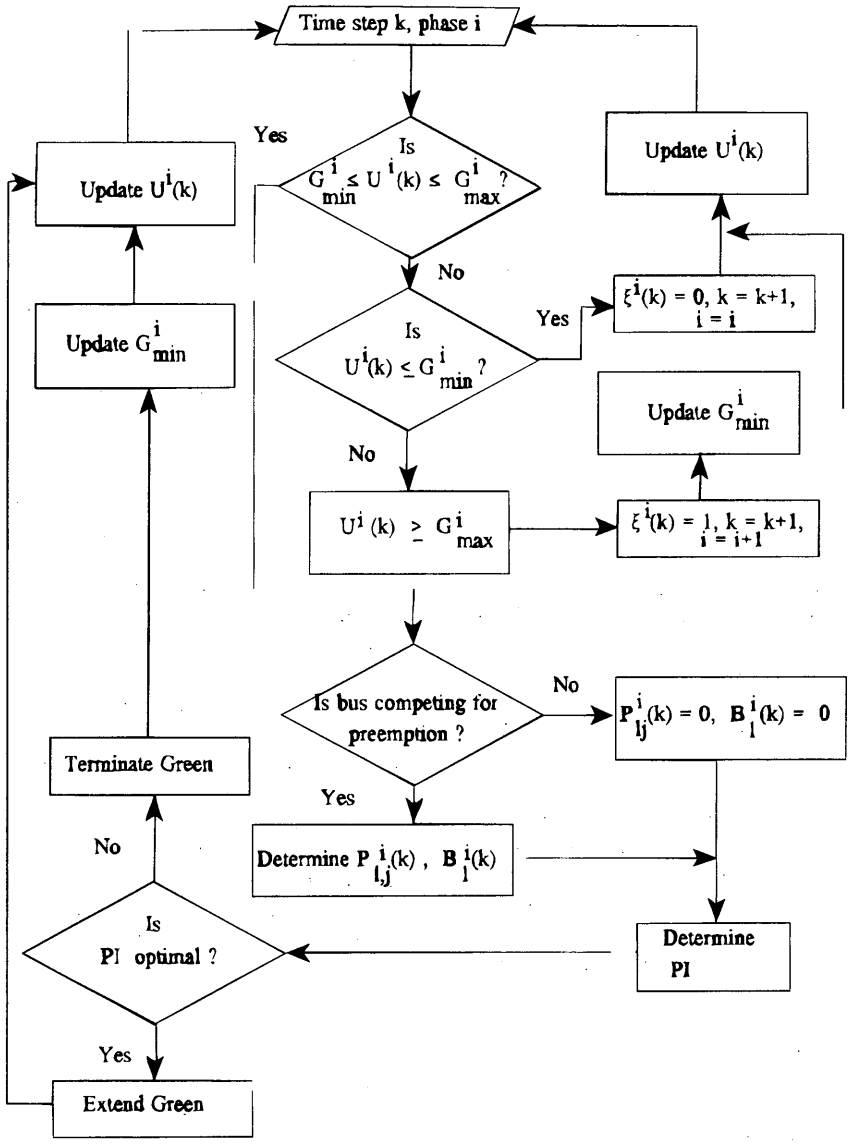


FIGURE 3 The control logic for bus preemption.

- Queue length at the beginning of the first time step in the experiment;
- Number of passenger car arrivals from the information supplied by the 36.6-m (120-ft) and 289.75-m (950-ft) detectors to estimate queue length; and
- Number of bus arrivals from the bus detector at 36.6 m (120 ft) to include in the preemption function and to estimate bus queue length.

To conform with the proposed control logic that a bus shall compete for preemption only when detected by the 36.6 m (120 ft) detector, the number of passengers in a detected bus were assigned according to a normal distribution with mean 15 and standard deviation 2.5. A schedule delay was designated, assumed to be uniformly distributed between 0 and 10 min. If a bus was not detected, the number of passengers and the schedule delay were recorded as zeros in the *PI* function.

Model Performance Evaluation

Based on the simulation output, the following computation procedure was used for testing the algorithm.

Criterion for Testing Performance of Adaptive Control over Actuated Control

The performance was tested based on the total queue length recorded at the end of every 3 sec for the entire intersection. Since

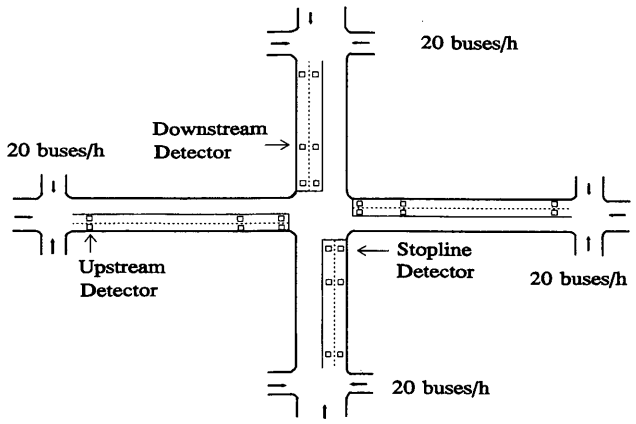


FIGURE 4 Layout of the experimental intersection.

NETSIM does not include a bus-preemption function, the model was first compared, without considering preemption, with the actuated control model of NETSIM for passenger car volumes of 500 vphpl and 1000 vphpl, and mean bus headway of 180 sec.

Criterion for Testing Performance of Adaptive Control With Preemption and Without Preemption

Having analyzed the effectiveness of adaptive control without preemption, the performance of the proposed model was studied. Hence, the total passenger delay at the intersection as a result of the signal control decision, with and without giving bus preemption, was investigated for passenger car volumes of 300 vphpl, 500 vphpl, and 1000 vphpl, and mean bus headways of 120 and 180 sec.

Discussion of Experimental Results

Following the first criterion for evaluating the performance of the proposed model without a preemption function, graphs (Figures 5 and 6) were drawn (*a*) for the total queue length at the intersection, (*b*) for each control time step for 90 time intervals (each of duration 3 sec) for the different listed cases, and (*c*) for both adaptive and actuated control logics.

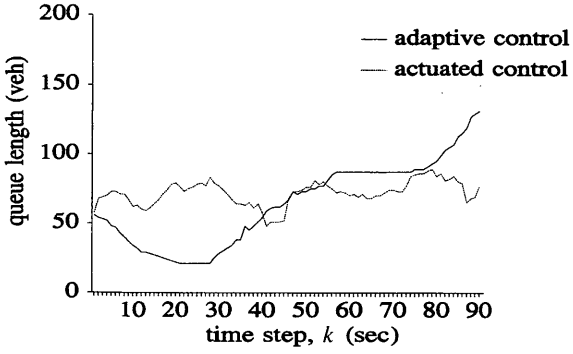


FIGURE 5 Total queue length for the demand level of 500-vphpl and 180-sec bus discharge headway.

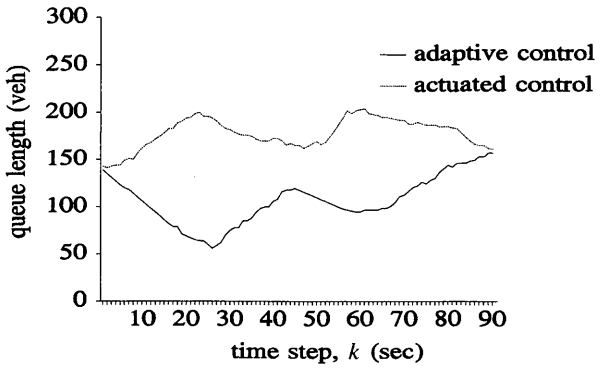


FIGURE 6 Total queue length for the demand level of 1,000-vphpl and 180-sec bus discharge headway.

Figures 5 and 6 show that the adaptive control logic yielded results superior to those of the actuated control simulated by NETSIM. For demand levels of 500 vphpl (Figure 5), and 1000 vphpl (Figure 6), the overall queue length for the actuated control model was more than the adaptive algorithm by 10 to 15 percent and 40 to 45 percent, respectively. For highly congested flow (1,000 vphpl), the adaptive control queue length was found to be less than the actuated control queue length for the entire test period. Thus, it may be concluded that adaptive control even without bus-preemption operation is superior to the actuated control under all traffic conditions.

To investigate the performance of the algorithm with preemption, graphs were drawn for (*a*) the total delay at the intersection for the different traffic scenarios under the second criterion and (*b*) the model with and without preemption. The total delay for the adaptive control logic with and without preemption is listed in Table 2 for all indicated scenarios. As observed in Table 2, the proposed adaptive control model with bus-preemption function was superior to the logic without preemption for all traffic volume conditions.

For Scenarios 1 (300 vphpl and 180 sec) and 2 (300 vphpl and 120 sec) in Figure 7, the algorithm without preemption produced 80 to 90 percent more delay than the one with preemption. For very heavy traffic conditions (1,000 vphpl), with mean bus discharge headways of 180 sec and 120 sec (Figure 8), the control logic with preemption produced adequately better results than the strategy without preemption. There also was an increase in the total delay for the control logic without preemption by 1 to 10 percent for the two discharge headways.

These results show that the proposed model performs well under light-to-moderate traffic volume situations, but exhibits a slight decrease in the benefit as the traffic state becomes highly congested. The reason is that under heavy congestion the total number of bus passengers in the queue have to compete with the long passenger car queue length for priority. Hence, a fair competition for very low bus volume does not exist. Despite the large difference in the two volumes, the proposed system exhibited a better performance than the logic without a preemption function. In the experiment, the random number of passengers assigned to a bus was assumed to follow a normal distribution with mean 15 and standard deviation 2.5. Varying the mean of the distribution from 5 to 30 and the standard deviation from 0.5 to 2.5 did not affect the superior performance of the proposed logic. Thus, the experimental results indicate the superiority of the devised model under all traffic conditions.

TABLE 2 Total Delay for the Adaptive Control Logic With and Without Preemption

Traffic Volume Delay (vphpl)	Mean Bus Discharge Headway (seconds)	Total Delay (seconds)		% Increase in for Model Without Preemption
		Without Preemption	With Preemption	
300	180	24,342	3483	85
300	120	25,470	4833	81.1
500	180	34,044	18,609	45.34
500	120	33,303	22,020	33.88
1000	180	57,990	57,195	1.37
1000	120	60,372	55,869	7.46

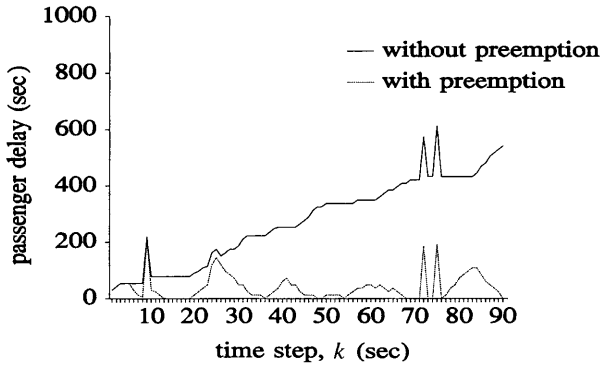


FIGURE 7 Total delay at the intersection for the demand level of 300-vphpl and 180-sec bus discharge headway.

CONCLUSIONS AND FURTHER RESEARCH

A model was formulated for an integrated adaptive control system with bus preemption and signal control functions. In the proposed model, absolute priority was not given to a bus. The model applied real-time algorithms instead of prespecified strategies used by more conventional bus-preemption logic. Driver safety and overall minimization of queue length were the two deciding factors when imposing the minimum green requirement. The control decision for signal setting was based on a performance index, which incorporated bus schedule delay, passenger delay, and vehicle delay.

Real-time traffic variables from the output of TRAF-NETSIM were used to test the performance of the algorithm. The experimental results proved the superiority of the proposed model over the actuated control logic simulated by NETSIM, under all traffic conditions. Hence, it may be concluded that the model performed favorably under all traffic volume states.

It should be noted that the primary focus of this article was to investigate the process of integrating bus-preemption and adaptive signal control. Hence, only a simple myopic adaptive logic was employed in the proposed system. An enhanced version of the proposed system, which uses information from both neural network prediction models and AVL systems for optimizing signal control over a projected time horizon, has also been developed and is available elsewhere (Chang et al., unpublished data).

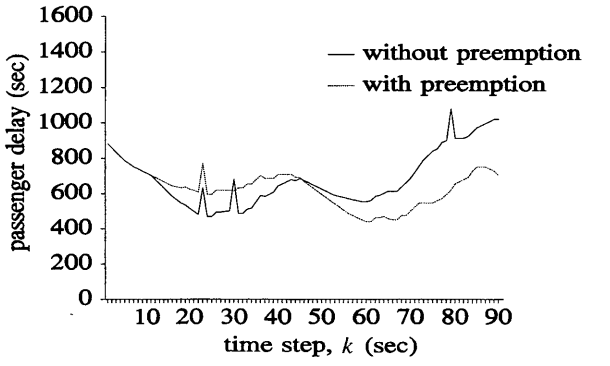


FIGURE 8 Total delay at the intersection for the demand level of 1000-vphpl and 120-sec bus discharge headway.

REFERENCES

1. Wilbur, E. J. *The Greenback Experiment-Signal Preemption for Express Buses: A Demonstration Project.* Report DMT-014. California Department of Transportation, 1976.
2. Ludwick, J. S. *Bus Priority System: Simulation and Analysis.* Report UTMA-VA-06-0026-1. Final Report prepared by the Mitre Corporation for U.S. Department of Transportation, 1976.
3. Courage, K. C., C. E. Wallace, and J. A. Wattleworth. *Effect of Bus Priority System Operation on Performance of Traffic Signal Control Equipment on NW 7th Avenue.* Report UTMA FL-06-0006. U.S. Department of Transportation, 1977.
4. Seward, S. R., and R. N. Taube. Methodology for Evaluating Bus-Actuated, Signal-Preemption Systems. In *Transportation Research Record 630,* TRB, National Research Council, Washington, D.C., 1977, pp. 11–17.
5. Lieberman, E. B., A. Muzyka, and D. Schneider. *Bus Priority Signal Control. Simulation Analysis of Two-Strategies.* Prepared by KLD Associates, Incorporated and Transportation Systems Center for the U.S. Department of Transportation, 1977.
6. Vincent, R. A., B. R. Cooper, and K. Wood. *Bus-Actuated Signal Control at Isolated Intersections-Simulation Studies of Bus Priority.* TRRL Report 814. Crowthorne, England, 1978.
7. EL-Reedy, T. Y., and R. Ashworth. The Effect of Bus Detection on the Performance of a Traffic Signal Controlled Intersection. *Transportation Research,* Vol. 12, No. 5, 1978, pp. 337–342.
8. TJKM. *Evaluation of Bus Priority Signal System.* Prepared for the City of Concord, Calif., 1978.

9. Salter, R. J., and J. Shahi. Prediction of Effects of Bus-Priority Schemes by Using Computer Simulation Techniques. In *Transportation Research Record, 718,* TRB, National Research Council, Washington, D.C., 1979, pp. 1–5.
10. Copper, B. R., R. A. Vincent, and K. Wood. *Bus-Actuated Traffic Signals-Initial Assessment of Part of the Swansea Bus Priority Scheme.* TRRL Report 925, Crowthorne, England, 1980.
11. Khasnabis, S., G. V. Reddy, and B. B. Chaudry. Signal Preemption as a Priority Treatment Tool for Transit Demand Management. *Proc., Vehicle Navigation and Information System Conference,* Paper No. 912865, Dearborn, Mich. 1991.
12. Allsop, R. E. Priority for Buses at Signal-Controlled Junctions: Some Implications for Signal Timings. *Proc., 7th International Symposium on Transportation and Traffic Theory,* Kyoto, Japan, 1977, pp. 247–270.
13. Jacobson, J., and Y. Sheffi. Analytical Model of Traffic Delays under Bus Signal Preemption: Theory and Application. *Transportation Research,* Vol. 15B, No. 2, 1981, pp. 127–138.
14. Heydecker, B. G. Capacity at a Signal-Controlled Junction Where There is Priority for Buses. *Transportation Research,* Vol. 17B, No. 5, 1983, pp. 341–357.
15. Heydecker, B. G. Delay at a Junction Where There Is Priority for Buses. *Proc., 9th International Symposium on Transportation and Traffic Theory,* 1984, pp. 113–132.
16. MacGowan, J., and I. J. Fullerton. Development and Testing of Advanced Control Strategies in the Urban Traffic Control System. *Public Roads,* Vol. 43, No. 3, 1979, pp. 97–105.
17. Mauro, V., and C. Di Taranto. UTOPIA. *IFAC Symposium on Control, Computers, and Communication in Transportation,* Paris, France, 1989, pp. 245–252.
18. Cornwell, P. R. Dynamic Signal Co-Ordination and Public Transport Priority. *IEE, Road Traffic Monitoring and Control,* 1986, pp. 158–161.
19. Han, B., and S. Yagar. Real-Time Control of Traffic with Bus and Streetcar Interactions. *IEE, Road Traffic Monitoring and Control,* 1992, pp. 108–122.
20. Han, B., and S. Yagar. A Procedure for Real-Time Signal Control that Considers Transit Interference and Priority. *Transportation Research-B,* Vol. 28B, No. 4, 1994, pp. 315–331.
21. Henry, J. J., J. L. Farges, and J. Tuffal. The PRODYN Real-Time Traffic Algorithm. *IFAC Symposium on Control in Transportation Systems,* 1983, pp. 305–310.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Testing of Light Rail Signal Control Strategies by Combining Transit and Traffic Simulation Models

THOMAS BAUER, MARK P. MEDEMA, AND SUBBARAO V. JAYANTHI

The Chicago Central Area Circulator (CAC) is a light rail transit (LRT) system scheduled to serve downtown Chicago by the year 2000. It will operate in its own travel lane parallel to automobile traffic; however, it will interfere with other surface transportation modes at intersections. The traffic and train signal system controlling the interface will be crucial for the successful performance of all modes. The signal control strategy must balance the needs of LRT, buses, autos, and pedestrians. For this reason, three LRT priority control strategies were developed. The approach used to analyze train and automobile traffic performance for each of these strategies is described. The CAC design team simulated LRT operation, automobile traffic flow, and intersection control units (ISC) as the interface between the two modes for all three control strategies. Two different microscopic modeling tools performed the simulation. TransSim II™ (registered trademark of James R. Hanks dba JRH Transportation Engineering) was selected for the transit and signal controller simulation because it realistically models LRT operation. TransSim II™ can also simulate priority strategies, which include arrival time estimation capability for trains and two-way communication between trains and ISCs. TRAF-NETSIM was selected for the traffic flow simulation because of its ability to reproduce traffic conditions, such as individual vehicles, queuing impacts, and potential spillbacks across adjacent intersections. The interface between the simulation programs is signal phasing and timing. This information calculated by TransSim II™ was read into TRAF-NETSIM. The two simulation processes yielded LRT performance measures of speed, travel time, and delay statistics, and auto performance measures of delay, queue lengths, and spillbacks. This allowed the design team to choose the most appropriate signal control strategy to provide the best overall system performance.

The public transit industry has experienced a resurgence. After the oil embargo in the 1970s and the recession of the early 1980s, interest in transit had declined. With the emphasis now on the economics of traffic congestion, environmental issues and new commuter travel patterns, more cities are looking to transit as a viable solution. New advances in the industry, such as alternative fuel vehicles, light rail transit, and bus signal preemption, are making transit more attractive.

With these new technologies, transportation engineers are looking for ways to make travel more efficient. A new application of microscopic simulation programs in analyzing transit signal control strategies is presented. The setting for this application is the City of Chicago's proposed Central Area Circulator light rail project. The application of Trans Sim II™ (registered trademark of James R. Hank dba JRH Transportation Engineering) and TRAF-NETSIM to simulate transit and traffic operations in downtown Chicago in a transit signal priority environment is described.

T. Bauer, Access Engineering Inc., 1410 Oak Street, Suite 201, Eugene, Ore. 97401. M.P. Medema and S.V. Jayanthi, Barton-Aschman Associates, Inc., 820 Davis Street, Evanston, Ill. 60201.

TransSim II™ as a microscopic transit simulation model was used in conjunction with TRAF-NETSIM, a microscopic traffic simulation tool, to help the designers measure the effects of several different signal control strategies and provide a recommendation based on quantitative analyses. The combination of these two simulation models allows for a detailed evaluation of transit and traffic impacts subject to the signal control strategy in operation.

SETTING

For the light rail project to be successful, light rail transit (LRT) travel speeds should be higher than conventional bus and auto speeds. The City of Chicago realized the importance of transit to the future of the Central Area and recommended that transit modes be given priority in the street system. Giving transit modes priority enables the street system to move the greatest number of people in the shortest period of time, creating a more efficient transportation system. To accomplish this goal, the LRT was given dedicated travel lanes and a priority signal system. The priority signal system will give priority service to the LRT while maintaining reasonable auto traffic performance and a safe pedestrian environment.

Several different signal control strategies were proposed to meet this requirement. The strategies ranged from a simple fixed time signal controller that would provide progression for LRVs to a preemption-type controller that would immediately respond to an LRV-activated call. Each of these signal control strategies had to be evaluated with respect to LRT performance, auto performance, and pedestrian safety.

It is imperative that pedestrian movements in the Central Area be preserved. Pedestrian traffic, particularly high in downtown Chicago, is the predominant mode of transportation. It is also critical that traffic flow in the city be maintained. Property owners and city officials have stressed the importance of unimpeded traffic flow for employee and customer travel in the marketability of commercial developments, and for the operation of businesses receiving deliveries. For this reason, no streets were closed to automobile traffic. The need to maintain reasonable traffic flow required detailed analysis. Other criteria also played a role in a separate set of analyses. Items such as maintenance of the signal system, cost, risk in development, and vendor acceptability were considered separately.

STUDY AREA

The study area network consists of seven north-south streets from Franklin to Wabash, and five east-west streets from Randolph to

Adams in Chicago's downtown Loop area. The network is basically a grid with an average block spacing of 450 ft. All streets in the study network are one-way streets, with the exception of two two-way streets: LaSalle and State.

The proposed light rail system will operate on Madison and State streets in both directions. The eastbound and westbound LRVs will stop on Madison Street at the station west of LaSalle and the station west of Dearborn. On State Street, southbound LRVs will stop at the station between Washington and Madison, and northbound LRVs will stop at the station between Washington and Randolph. The study area includes seven LRT intersections and one LRT junction. Three different routes will operate in the study area. For the purpose of the simulation, the two routes operating on Madison Street and on State Street north of Madison Street are combined into Route No. 1, while Route No. 2 includes the LRT route operating on State Street. The circled portion in Figure 1 shows the study area.

METHODOLOGY

To perform such a complex analysis, several different approaches were analyzed to determine which applies best to this situation. Four approaches were considered for this project.

The first attempt was a macroscopic look at each of the signal control strategies. The traffic performance was measured using the Highway Capacity Manual (HCM). The HCM was used to conduct intersection capacity analyses to determine auto delay. The amount of lost time to the autos due to the LRT phase was coded as an all-red phase. This procedure provided an estimate of the overall intersection performance but failed to consider (*a*) the effects of upstream and downstream intersections, (*b*) the cumulative effects of queuing on downstream street segments as well as at the intersection; and (*c*) the variable phase lengths that could be generated by an LRV-actuated call. This method also could not predict train performance. Pedestrian safety was considered by providing safe pedestrian clearance times during each cycle.

The second attempt was to create a manual approach to show the network-wide effects of the many intersections by developing time-space (T-S) diagrams. The T-S diagrams were able to show the impacts to autos along street segments by showing the progression along street corridors. This provided an indication of the train performance by showing train progression while including station dwell times at each station stop. Combined with the results of the HCM, this method provided a better understanding of the auto and train performance, but still could not predict the effects of the variations in LRT arrivals and the variations in the signal timings.

The third attempt was the application of a microscopic program to show the effects of the variable signal timings. TRAF-NETSIM, a simulation program developed by the FHWA, was used to determine both LRT and auto performance throughout the network. This program allowed the auto lanes and the LRT lanes to be coded as separate links for most of the intersections. Because of a program limit of five approach links to each intersection, occasionally some of the LRT lanes and auto lanes had to be combined. The intersections were coded as actuated signals with detection loops in the transit lanes. This method provided auto performance statistics and LRT performance statistics that incorporated some of the variability of LRT arrival patterns and ability of the signal system to accommodate the transit calls in the signal cycle. This program also allowed the coding of short-term disruptions to the transit lanes. Examples of disruptions include jaywalkers, vehicles turning into alleys, vehicles turning into parking lots, and pedestrians forming queues extending into the street. However, the real signal control strategies that were being developed for this project had some unique capabilities that TRAF-NETSIM was not able to reproduce. The ability to constantly send information from the LRV to multiple controllers to update the LRV arrival time and cancel calls if a delay was experienced could not be analyzed.

Another alternative had to be developed that could improve TRAF-NETSIM's ability to analyze the different types of advanced signal control strategies but still be able to measure auto performance in the way TRAF-NETSIM could. The fourth attempt, therefore, involved TransSim II™, a microscopic simulation tool for transit operations, that became available for the Central Area Circulator (CAC) project. This method employed a two-step process. The first step was to use TransSim II™ to simulate transit operations and signal controllers, and the second step was to simulate traffic operations with TRAF-NETSIM using signal timing and phasing provided by TransSim II™.

SIMULATION MODELS

TransSim II™

TransSim II™ is a simulation program that models light rail transit or bus transit operations. It is a link-node-based model that treats transit operations on a microscopic level and other traffic on a macroscopic level. The utility of the program lies in the abundance of information that is modeled on transit and traffic signal operation.

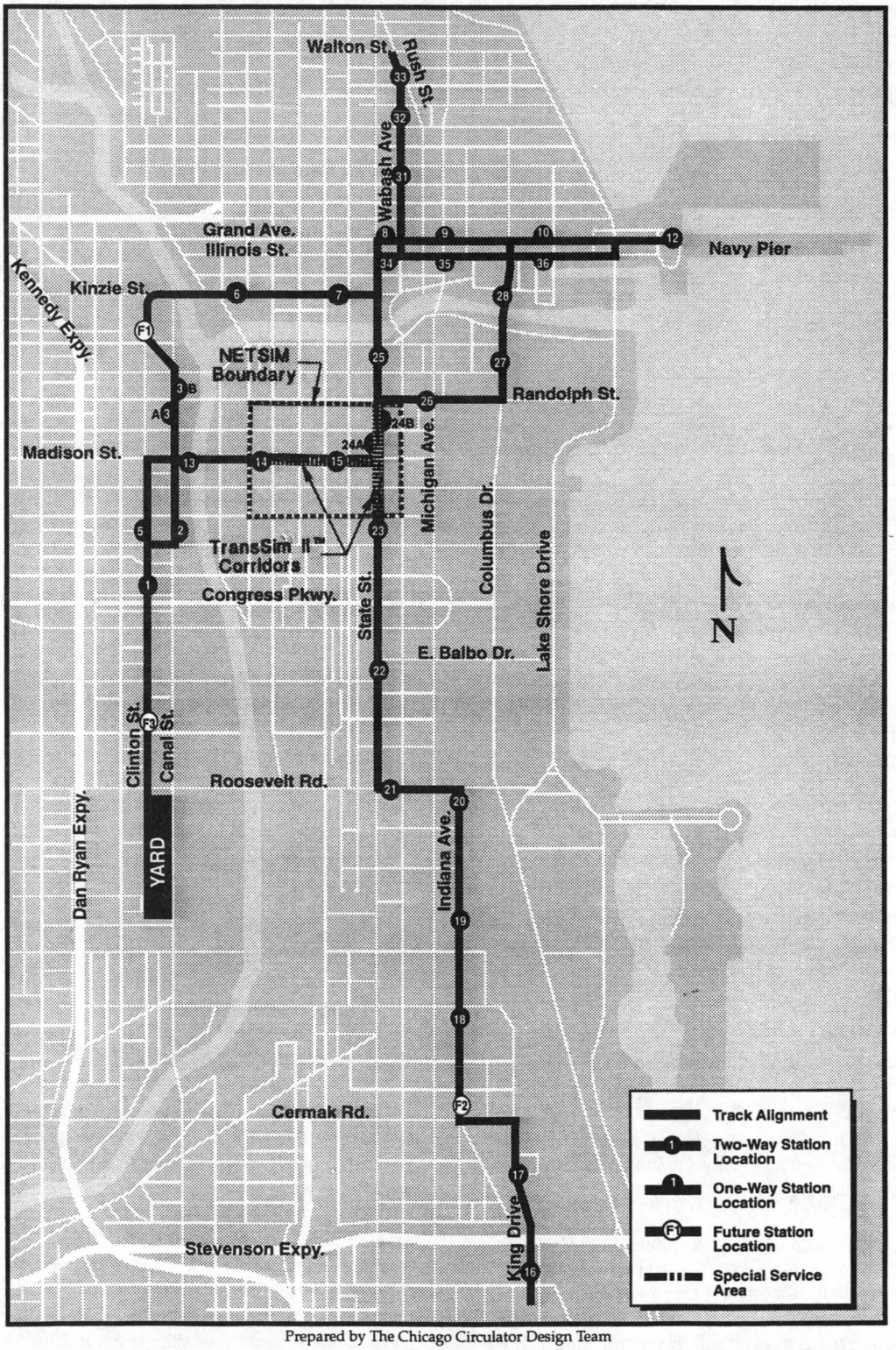
Transit operations are modeled on a real-time basis through a traffic-signal, controlled-street network. The transit operations output shows (*a*) the overall travel time for each transit vehicle and all vehicles, cumulative and averaged; (*b*) detailed point-to-point travel times; (*c*) the time and duration of delays at traffic signals and stations; and (*d*) the time and duration of traffic signal preemption at each intersection. This is accomplished by input data that describe the exact transit route (including the location of stations and intersections) and operating parameters, such as acceleration, deceleration, speed zones, and station dwell times.

The program contains logic that allows the modeling of real-world situations affecting transit operations:

- Maximum operating speeds may vary along the route to account for operating in separate rights-of-way, mixing with automobile traffic, negotiating curves, or other conditions;
- Station dwell times are calculated, taking into account a randomly generated variable, the mean dwell time, and the time gap to the proceeding train; and
- User-defined random delays may be input at any location for any duration.

All common types of controllers can be modeled, from fixed-time to fully actuated control, at isolated intersections or in coordinated systems. Traffic signal controllers are simulated on a second-by-second basis and may be set to provide full preemption and most common types of transit priority treatment.

TransSim II™ can simulate traffic operations on a macroscopic second-by-second basis. However, this traffic model cannot show the effects of queue spillbacks or heavy pedestrian flows. For these



Prepared by The Chicago Circulator Design Team

FIGURE 1 Boundaries of study area (*1*).

reasons, TRAF-NETSIM was used for traffic simulation instead of TransSim II™'s in-built traffic model.

TRAF-NETSIM

TRAF-NETSIM, a microscopic traffic computer simulation model developed by the FHWA, is used to simulate urban surface-street networks. This model tracks individual vehicles as they traverse the study network. After entering the network, vehicles move according to the vehicle-following logic by responding to traffic control devices and other factors, such as pedestrians, buses, etc. Individual driver behavior characteristics also are represented through random sampling from probability distribution to reflect real-world processes.

The physical structure of the roadway is represented as a network consisting of nodes and unidirectional links. The links represent the streets, and the nodes represent the intersections or the points at which a geometric property changes (e.g., a lane drop, a change in grade, or a major mid-block traffic generator). Because of its detailed view of traffic operations, TRAF-NETSIM is a valuable tool for understanding the performance of different transportation system strategies.

The input parameters for the model include: traffic volumes, lane geometrics, lane usage, grades, pedestrian intensity, start-up lost times, mean headways, average free-flow travel speeds, signal phase splits, signal phase movements, bus routes, dwell times, etc. TRAF-NETSIM produces vehicle statistics by individual links. The operational performance of the transportation system can be evaluated by using one or more of the following measures of effectiveness produced by TRAF-NETSIM. They are: average stopped delays, total delays, percentage of stops, average travel speeds, average and maximum length of queues, total travel time, vehicle emissions, and derivatives of these. Individual turn-movement-specific statistics also can be obtained. The output results can be viewed graphically or numerically. The graphical capabilities of TRAF-NETSIM provide easy explanations of traffic performance for laypeople and easy verification of operations.

APPLICATION

Tested Priority Strategies

Three signal control strategies for the LRT operations were proposed and evaluated. The logic behind the three strategies is as follows:

Strategy 1: Operates under fixed-time signal controller logic at intersections and semiactuated at junctions based on a signal timing plan balancing progression for LRT and autos.

Strategy 2: Same as Strategy 1, but LRVs can extend their green window by early termination of the previous phases ("early green") or later termination of their own phase ("green extension").

Strategy 3: LRVs can predict their arrival time at the intersections. Two-way communication between LRVs and signal controllers then allows the signal controllers to optimize signal timing to minimize delay for light rail. Controllers also have the capability to return to coordination in the absence of LRV calls.

TransSim II™

Input Data

In addition to geometric information, several basic assumptions were made for the simulation of the CAC network:

- In the year 2010 all trains will be 55 m (180 ft) long (two-car trains), resulting in LRV clearance times of 9 sec for through movements and 21 sec for turning movements.
- Operating rule requires a minimum spacing of one 128-m (420-ft) long block between two consecutive trains.
- Acceleration and deceleration rates are set to 1.1 m/sec^2 (3.5 ft/sec^2) and 1.4 m/sec^2 (4.5 ft/sec^2), respectively, according to design specifications.
- Entrance times for trains to the simulation network are defined for each train individually based on earlier simulation efforts.
- User-defined random delays are defined by coding delay locations and durations for each train specifically based on the assumption of an exponentially distributed average delay of 7.5 sec per train-kilometer (12 sec/mil) traveled (D. Allen, unpublished data).
- Signal control-related input data includes the base timing plan (phase lengths and offsets). This information was prepared by adjusting an automobile traffic-oriented timing plan to better accommodate LRVs with their exceptional travel characteristics.

Output Data

Output data from TransSim II™ contained a variety of information and included (*a*) measures of effectiveness (MOEs) for the light rail system and (*b*) the lengths of all signal phases for each cycle during the simulated period of time.

The MOEs presented for each light rail vehicle are:

- total travel time in seconds,
- station dwell time in seconds,
- average speed (route length divided by total travel time) in km/h (mph),
- variation from an ideal run (without any delay caused by longer-than-expected dwell times, traffic signals, interference with other LRVs, or user-defined delay) in seconds,
- stop line delay (the accumulated time the LRV was waiting at traffic signal stop lines) in seconds,
- time-to-green delay (the accumulated time from when the LRV passes a decision point, breaking distance to stop line, to the start of LRV GO) in seconds,
- non-station delay (the total delay the LRV receives neglecting any variation of station dwell time) in seconds, and
- user-defined delay (the sum of all random delays defined for the LRV) in seconds.

For all LRVs of each route and direction, TransSim II™ then presents minimum, maximum, mean, and standard deviation for the MOEs. For better interpretation of the results, the ideal travel time and corresponding ideal speed also are presented.

The last part of the comprehensive transit results shows the accumulated time-to-green and stop line delays, and the fraction of vehicles that actually come to a stop for each traffic signal. The mean stop line and time-to-green delay per intersection and train is then displayed as a general MOE for each route and direction.

Signal timing data to be input into TRAF-NETSIM was stored in one data file for each traffic signal of the simulated network. These files included the phase number and its duration of green time and clearance time in the sequence they appeared during the simulation.

TRAF-NETSIM

Input Data

TRAF-NETSIM requires extensive input data, and a description of the important variables is as follows:

- The future lane geometrics, balanced auto volumes, and bus volumes were coded in the network.
- The pedestrian traffic factor that takes a high pedestrian volume of 250 to 500 pedestrians per hour is used.
- The start-up lost time, which is the delay experienced by all lead vehicles in a queue when responding to a phase change from red to green, is set at 2 sec.
- The mean time gap and the free-flow speeds used in the network are 2 sec and 48 km/h (30 mph), respectively.
- Right-turns-on-red are permitted for auto traffic at all intersections.
- Average dwell times for buses are specified as 10 sec.

The signal control strategies were tested with TRAF-NETSIM using the signal timings generated by TransSim II™. Because the format of signal timings generated by TransSim II ™ output is not compatible with that of TRAF-NETSIM, adjustments were made. TransSim II™ generates a series of signal phase sequences and corresponding splits at each intersection for a fixed duration. To use exactly the same signal timings, the time period capability of TRAF-NETSIM had to be used.

Changing conditions with varying signal timings can be simulated using time periods. Each time period should be an integer multiple of a time interval. The time interval is the most commonly used cycle length (in this case 75 sec). The maximum number of phases allowed in each time period is 12. At some intersections, the cycle uses up to 6 phases. These limitations restrict the user to only two time intervals in each time period. This means that it is feasible to input an equivalent of 150 sec of phase splits in each time period under the existing circumstances. TransSim II™'s signal timing output at each intersection is broken down into 150-sec intervals and re-formatted to match TRAF-NETSIM's format. Since TRAF-NETSIM allows the use of a maximum of 19 time periods, it is possible to simulate up to a maximum of 2,850 sec.

The model was thoroughly calibrated to replicate real-world conditions before the testing of the strategies began. This was accomplished by comparing the average stopped delays, queues, and the traffic volumes obtained from the field with TRAF-NETSIM's results.

Output Data

TRAF-NETSIM produces an abundance of MOEs in which the maximum queue lengths and the stopped delays were selected to evaluate the non-LRV traffic operational performance. Stopped delay is the amount of time an average vehicle is forced to stop at the intersection due to traffic conditions. The maximum queue length is the longest queue that has occurred during the simulation. The systemwide MOEs were calculated from the individual link-by-link statistics.

Results

The results of the analyses include average train speeds, delays experienced by auto vehicles, and the maximum queue lengths that develop on each leg of an intersection. Table 1 shows the average train speeds, and Table 2 shows the total systemwide auto delay and the relative differences (in percent) for the three strategies.

Figure 2 shows the relationship between train performance and auto performance for each of the different signal control strategies and their alternatives. A linear relationship is shown illustrating how train performance and auto performance are related. As train performance increases, auto performance decreases. This is intuitive as autos and trains must share a fixed amount of space and time.

Compared with Strategies 1 and 2, the average LRT operating speeds are significantly higher in Strategy 3. The auto performance is better in Strategy 1 and is identical for Strategies 2 and 3.

CONCLUSION

The use of TransSim II™ and TRAF-NETSIM allowed the Chicago Circulator Design Team (CCDT) to identify the most suitable transit priority strategy for the proposed light rail system in downtown Chicago. The detailed analysis that TransSim II™ provided for transit operations and TRAF-NETSIM for automobile traffic enabled the CCDT to predict impacts on light rail and traffic operations from various signal control strategies.

The detailed quantification of the light rail and non-light rail operational performances helped the CCDT select an appropriate

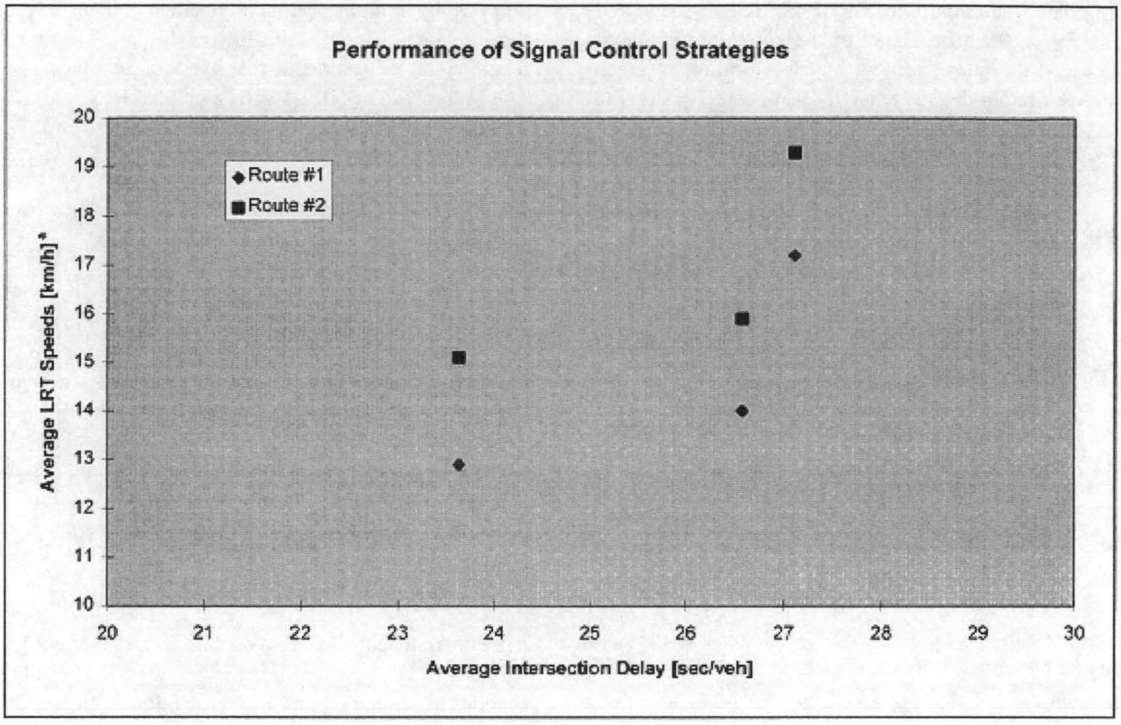
TABLE 1 Average Train Speeds in km/h for pm Peak Period

Route	Strategy 1	Strategy 2	Strategy 3
#1	12.9	14.0	17.2
#2	15.1	15.9	19.3

1 km/h = 0.6 mph

TABLE 2 Summary of Traffic Performance for pm Peak Period

Criteria	Strategy 1	Strategy 2	Strategy
Systemwide Delay [sec]	827	930	950
Change from Alternative 1 [%]	n/a	12.5	14.9



* 1 km/h = 0.6 mph

FIGURE 2 Interdependence between transit and traffic performance.

signal control strategy for the proposed light rail system in downtown Chicago.

The simulation models also helped quickly evaluate several additional variations to the input to understand the effects on train and traffic operational performances given different constraints to the signal control strategies.

REFERENCES

1. Chicago Circulator Design Team. *Train/Traffic Control and Communications System Evaluation Report.* City of Chicago, Central Area Circulator Project, May 1994.

Publication of this paper sponsored by Committee on Light Rail Transit.

Validation of Simulation Software for Modeling Light Rail Transit

STEVEN P. VENGLAR, DANIEL B. FAMBRO, AND THOMAS BAUER

As the engineering and planning communities continue their progress toward managed and integrated transportation systems, transit will play an increasing role. Light rail transit (LRT) has already been selected and implemented by 15 U.S. cities as a rail transit alternative. As new or expanded systems are planned and designed, it is essential that engineers have the means to make the best decisions for LRT placement and operations. The purpose of this research study was to investigate the use of the TRAF-Network Simulator (NETSIM) program and JRH Transportation Engineering's TransSim II™ tools for agencies interested in planning and developing LRT systems. NETSIM is one of the few available traffic analysis programs with the flexibility to model the operations and mobility impacts of transit. Similarly, TransSim II™ can model the impacts of transit and has been developed for this purpose. To evaluate NETSIM and TransSim II™ for simulating traffic in pretimed and actuated arterial networks, outputs from the models were compared with real-world field data from Los Angeles and Long Beach, Calif. and Portland, Oreg. The results indicated that the models could produce moderately accurate estimates of field-stopped delay and percent-stops for individual intersections within studied networks. On a systemwide basis, the models produced reasonably reliable, accurate estimates of network travel times and could reproduce most traffic characteristics observed in the field. The models performed well in simulating the control impacts and behavior of LRT in the modeled systems.

While planning a future light rail transit (LRT) system, or even for examining operational alternatives for an existing LRT system, it is essential that tools are available to assess the impacts of transit on the existing transportation system. Measures of effectiveness (MOEs) describe these effects, which include delay to motorists and transit riders, fuel consumption, emissions, and overall mobility. With such information, selecting the best alternatives for implementing LRT is possible. To produce the necessary data base of MOEs, analysts use models that simulate the LRT system operations. These models can range from mathematical procedures to computer simulation. Computer simulation is often used to process the necessary information and maintain records of the myriad variables describing the interaction between drivers, vehicles, and the roadway.

For traffic engineering applications, the Federal Highway Administration's TRAF-NETSIM (TRAFfic-NETwork SIMulator) is perhaps the most flexible computer simulator. NETSIM can simulate networks under control strategies ranging from sign control to fully actuated signal control. The model can provide MOEs for a variety of traffic scenarios and can simulate LRT in urban environments using a variety of methods. Proprietary software has also been developed to determine the network impacts of LRT. JRH Transportation Engineering's TransSim II™ can simulate LRT using a variety of control and priority schemes for transit and providing MOEs for network traffic.

After the development of a simulation method for computing LRT effects, any shortcomings in the procedure can lead to a failure of the planned system. Therefore, it is essential that the model produce accurate and reliable results. Model calibration and validation ensure that the model outputs accurately represent the effects of the planned LRT system. For this report, calibration consists of adjusting NETSIM and TransSim II™ model inputs and default parameters to model as accurately as possible the true data from field observation. The validation procedure statistically tests and assesses the ability of the model to replicate real-world conditions.

S. P. Venglar and D. B. Fambro, Texas Transportation Institute, Texas A&M University System, CE/TTI Suite 301E/301G, College Station, Tex. 77802, T. Bauer, JRH Transportation Engineering, 1580 Valley River Drive, Suite 160, Eugene, Oreg. 97401.

Considerations for Modeling LRT

The model inputs and embedded parameters for simulation of LRT in an urban street system include the location of the transit line with respect to the roadway, the environment in which LRT will run, general aspects of LRT operations, traffic control devices, and possible priority schemes.

Crossing Configurations

Four major at-grade configurations exist for LRT-roadway intersections: (a) isolated crossings, (b) isolated crossings with a nearby traffic control device, (c) crossings where LRT is adjacent to a parallel street, and (d) crossings for LRT median operation (*1*). For each type of crossing, there are modeling concerns such as the presence and handling of turning vehicles, the need to prevent cross-street vehicles from encroaching on the LRT tracks, the priority provided for light rail vehicles (LRVs), and optimal signal timing. Also important are the effects of altering the signal timing for an LRV when the signal is timed for arterial progression.

The LRT Physical Environment

LRT right-of-way and environment describe the purpose and exclusivity of the corridor in which the LRT line will be located. The land on which the line is or will be constructed may be devoted entirely to the transit facility and its appurtenances, it may be shared with a freight rail line, or it may even be in the right-of-way of a municipal street. Within the corridors, varying at-grade LRT track placements have been used in cities around the country. Despite this diversity, five general classes of track locations define and classify a vast majority of these placements. Ranging from least to greatest interaction with automobile traffic, these locations are: (a) grade

separation, (b) exclusive right-of-way, (c) side of street, (d) median of street, and (e) mixed traffic. Grade separation is included in this discussion as many predominantly at-grade LRT lines are grade-separated at intersections where much automobile congestion exists.

LRT Operations

Providing accurate information about the vehicle's features and operations ensures accurate representation of the LRV within the model. The list here includes vehicle characteristics, headways, dwell time, operating speed, and time factors at roadway crossings (including blockage time, clearance time, and lost time).

Traffic Control Devices

Pursuing the discussion of LRT roadway crossings, another topic is the type of control used at the crossing. The crossing may exhibit crossbucks only, flashing lights with crossbucks, flashing lights with gates and crossbucks, or standard traffic control devices (*1*). Each control option has different blockage, clearance, and lost times, and all differences must be accounted for as accurately as possible within the model.

Control Strategy

In addition to the reproduction of the physical aspects and features of the modeled environment, incorporation of the control strategy found in the network is also necessary. Where LRVs and automobiles are considered equally, no modifications are required; however, where transit is given special treatment, signal priority for the LRV must be considered in the model.

DATA COLLECTION

For each modeled network under investigation, two separate sets of data were collected. Analysts used the first set to calibrate NETSIM and TransSim II™ for use with LRT. They used the second set to validate the model's ability to recreate the modeled environment. Since the data were specifically being collected for input to NETSIM and TransSim II™, the models defined the data collection requirements.

Information gathered at the field data collection sites used in this study consisted of network description data, travel time information collected using a portable computer, and videotapes of at least one major intersection within each of the study networks. The video allowed for later reduction of intersection measures of effectiveness. Study data were organized around the five geographic data collection sites. Networks 1 and 2 were located along Washington Boulevard in Los Angeles, California; Network 3 was located along Pacific Avenue in downtown Long Beach, California; Network 4 was located in Portland, Oregon along Holladay from Martin Luther King to 13th; and Network 5 was located along Burnside in Portland from 102nd to 122nd.

In the Los Angeles and Long Beach networks (Networks 1, 2, and 3), the light rail operated without priority in the median of a pretimed arterial system. In the Portland networks (Networks 4 and 5), light rail operated in the median or on the side of the street with full priority. Light rail approach "calls" were received early enough to ensure that cross-street vehicular and pedestrian minimum times were served. The intersection then dwelled in phases that did not conflict with the LRV until the LRV "checked out" or was "timed out" of the intersection.

NETSIM

The NETSIM model (*2*) performs a microscopic simulation of traffic flow in an urban street network. It is designed for traffic engineers and researchers as an operational tool for evaluating alternative network control and traffic management strategies. NETSIM allows the designer to simulate the performance of traffic under a number of alternative control strategies.

NETSIM application to LRT simulation is not new. Simulation of the Downtown Area Rapid Transit (DART) North Central Light Rail Line was accomplished using a modified version of the software (*3*). The original software did not readily accommodate the complex, frequently changing signal sequences found in the "window"-limited priority scheme proposed for the DART line. Restrictions in NETSIM that limited the signal transition flexibility were identified and their influence on the simulation was mitigated. NETSIM was used, with TRANSYT-7F and the Highway Capacity Software, to identify the delay impacts of LRT and the presence, if any, of residual queues after LRV passage.

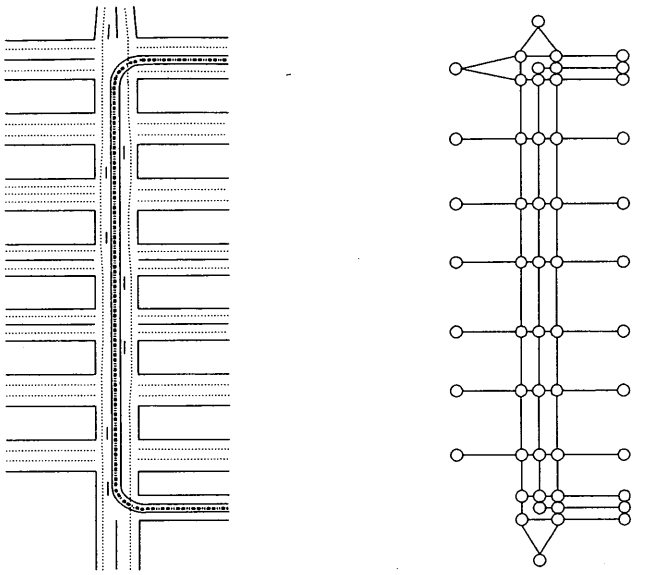
NETSIM was also used (*4*) to evaluate the relationship between an intersection crossing volume and the average automobile delay at an isolated LRT crossing. In NETSIM, the LRT was modeled as a single-lane roadway, and the grade crossing as a two-phase, fully actuated intersection. LRVs' arrivals were modeled as buses using specified headways. The model, however, gave unconditional priority to the LRT vehicles and made no allowances for signals and progression (*4*).

Coding the Modeled Environment in NETSIM

The described geometric, traffic volume, and signal timing information was input into the model using files that contained series of cards. Each card contained information about a particular feature of the modeled environment. Special card types used in the model to simulate bus operations were used to model the LRT in NETSIM.

For each of the pretimed networks (Networks 1, 2, and 3), the required input data was readily processed for entry into the model. Once the necessary information was assembled, the physical features of the roadway environment, the traffic volumes and turning percentages, and the traffic signal data were input via NETSIM's card-type format. The few exceptions to this rule include: (a) any links to the left of left-turn bays cannot be moving links (making it impossible in this scenario to directly model median-running LRT) and (b) links in the model have a minimum length of 15 m (50 ft). Modeling the median-running (or side of street running) LRT given the constraint of the minimum link length requirement produced a network that not only was more complex to model, but also one that required cross-street vehicles and arterial street left-turning vehicles to travel distances that were not present in the modeled environment (see Figure 1).

The coordinated actuated (Network 4) and fully actuated (Network 5) networks used the same LRT node format as the pretimed



Field Environment NETSIM Representation

FIGURE 1 NETSIM Representation of an existing median-running LRT network.

networks. Since the LRT and traffic nodes were separated, the approach of LRVs did not directly influence signal control at the traffic nodes. While vehicles conflicting with the LRT still received green time at modeled vehicular nodes in the presence of an LRV, the vehicles were not able to advance across the median "tracks" at the LRT node. This coding allowed reasonably accurate modeling of field traffic, LRV, and controller behavior except the dwell time in coordinated phases found in the field in Network 4. Coordinated phase dwell time was used in the field environment to "resync" controllers that were unsynchronized by the priority of the approaching LRV, giving extra green to the coordinated cross-street phases. Since dwell time could not be replicated in the model, some green time for the cross streets was not reproduced in the model.

Calibration

Initial calibration of the model consisted of using field-observed means and distributions of start-up lost time and queue discharge headway rather than NETSIM default values for these parameters. Also, repeated link "free flow" speed adjustments were made to the model to coordinate downstream arrivals in the model with patterns observed in the field. Improvement caused by changes to the model was monitored by comparing the modeled output with a calibration field data set. Changes were easily noted since components of the summary output provided by NETSIM were directly comparable to observed calibration field data MOEs. The primary cause of discrepancies between the model and the calibration field data appeared to involve the queue discharge and platoon dispersion behavior in the model. NETSIM tended to "spread out" the platoon earlier and to a greater extent than observed behavior in the field.

Validation

Following calibration, the model was run to produce a simulation data set for comparison to the validation field data. Three categories of comparisons were made for traffic: (a) individual link travel times, (b) network directional travel times, and (c) individual intersection MOEs. Individual link and directional travel time analyses were also performed for LRT.

Analysis showed that 40 percent of modeled individual links displayed travel times within ±20 percent (judged an acceptable range of accuracy) of the validation field data. Link travel times from the field and NETSIM are presented in Figure 2. Network directional

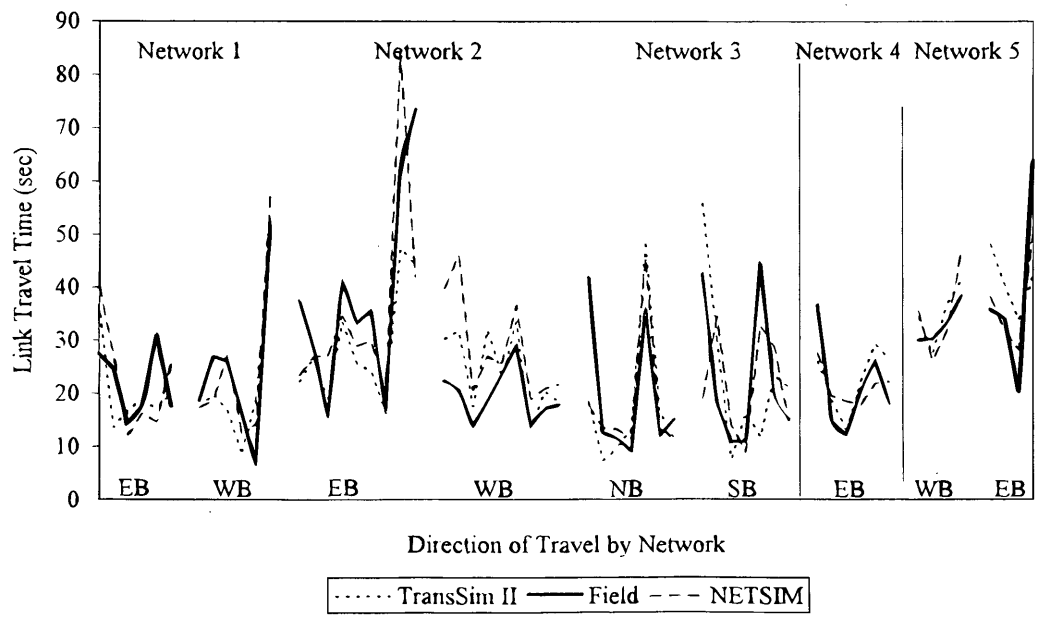


FIGURE 2 NETSIM and TransSim II™ traffic travel time comparison to validation data.

travel time analysis showed that eight out of nine systemwide travel times were accepted at the 95-percent confidence level. Platoon effects present in the field environment that could not be wholly accounted for in the calibration procedure were identified as the major cause of the discrepancy between the model and field link travel times. As the model tended sometimes to predict arrivals earlier and sometimes later than the field, the directional travel times "averaged" out these effects and the model estimates of directional travel time were more accurate than the link measures.

Individual intersection MOE analysis was performed on the stopped delay and percent stops output from NETSIM. Correlation analysis indicated a moderately strong correlation between field- and model-stopped delay and a moderate correlation between field and model percent stops. The stopped delay estimates from NETSIM and the comparison field data are shown in Table 1.

Priority for the LRT in Networks 4 and 5 made the travel times for LRVs vary from the travel times for traffic. The calibration for transit was similar to the calibration for traffic. To validate field travel times for transit, LRT travel time through the actuated networks was compared to the same values from the model (Figure 3). As with the analysis for traffic, the individual link travel time estimates from the model were not as accurate as the directional travel times. Thirteen of the 20 individual LRT link travel times, or 65 percent, were within the ± percent criteria, while three of the four directional LRT travel times were accepted at the 95-percent confidence level.

The graphics component, GTRAF, included in the TRAF software was an invaluable asset throughout the research investigation. Both the static and animated graphics supplied by the model assisted in describing how the input data were accepted by the model, in finding coding errors in the input data sets, and in clarifying the queue discharge behavior of the model.

TRANSSIM II™

TransSim II™ is a program developed by JRH Transportation Engineering of Eugene, Oreg. After identifying the shortcomings mentioned in current software for modeling LRT, JRH developed a program specifically designed for modeling LRT or bus transit in urban networks. The program is microscopic with respect to LRT (or bus) behavior and movement within the modeled system and macroscopic with respect to traffic performance. The computation of MOEs for traffic is accomplished within TransSim II™ using a methodology similar to the TRANSYT program.

Inputs to the program include features of the roadway environment, including geometrics, traffic volumes, and signal phasing, and information about the transit route, including stations and intersections. Operating speeds and station dwell times can vary to simulate realistic transit operations. The analyst enters data in a pull-down menu format under the entries of system data, route data, link data, and signal data. A variety of types and degrees of priority are available and each can be easily selected by the user, facilitating the evaluation of alternative control strategies for the networks.

No unusual configuration was necessary for the five modeled networks and the signal control type was specified by selecting the priority level (a defined code with a variety of control types possible for selection) for the intersection. The selection of a priority level for transit and the entry of subsequent control and phasing information for this priority level were the main differences in coding between the pretimed, nonpriority networks (1, 2, and 3) and the semi-actuated and fully actuated priority networks (4 and 5, respectively).

TABLE 1 NETSIM and Field Intersection MOEs

	Mean Stopped Delay		Mean Percent Stops	
Intersection	Validation Data	Model w/ LRT	Validation Data	Model w/ LRT
Flower & Washington				
EB Approach	2.27	4.47	13	12
NB Approach	29.99	16.74	83	77
Central & Washington				
EB Approach	6.13	5.22	24	38
NB Approach	21.28	32.62	80	68
First & Pacific				
NB Approach	10	6.37	56	78
SB Approach	7.36	5.32	45	29
Broadway & Pacific				
NB Approach	16.19	5.03	71	20
SB Approach	20.86	18.28	68	51
MLK & Holladay				
SB Approach	6.01	5.61	28	33
122nd & Burnside				
NB Approach	31.41	25.53	73	74

Coding the Modeled Environment in TransSim II™

Geometric, traffic volume, signal timing, and LRT information necessary for input into TransSim II™ was entered using the pull-down menu driven data entry format of the program. The program main screen displays five menu options; (a) File, (b) Edit, (c) Schedule, (d) Run, and (e) Result and Graphics *(5)*. Data were entered using the Edit and Schedule menus.

Calibration

Following the entry of the geometric, traffic volume, and signal timing data, few adjustments were required to run the model. Several inputs, including entries for LRV acceleration and deceleration, start-up lost time, and average speeds for LRVs and automobiles, enabled adjustment of the model's environment parameters to field conditions. The one model parameter that did require adjustment through iterative runs of the program was the location of the detector that notified the downstream intersection of an approaching LRV in the priority networks (Networks 4 and 5). This distance was nominally the braking distance of the LRV plus any remaining distance required to produce the time equivalent of the minimum phase duration on the cross street.

A number of information elements were required to model LRT in TransSim II™ accurately. Because the program treats LRV behavior microscopically (i.e., LRVs are tracked through the system and detected to receive priority calls), any physical or control elements impacting the LRV had to be identified and entered.

Validation

Following data entry and detector calibration for the priority networks, the final TransSim II™ runs were made. The output data set was then statistically compared with the validation data.

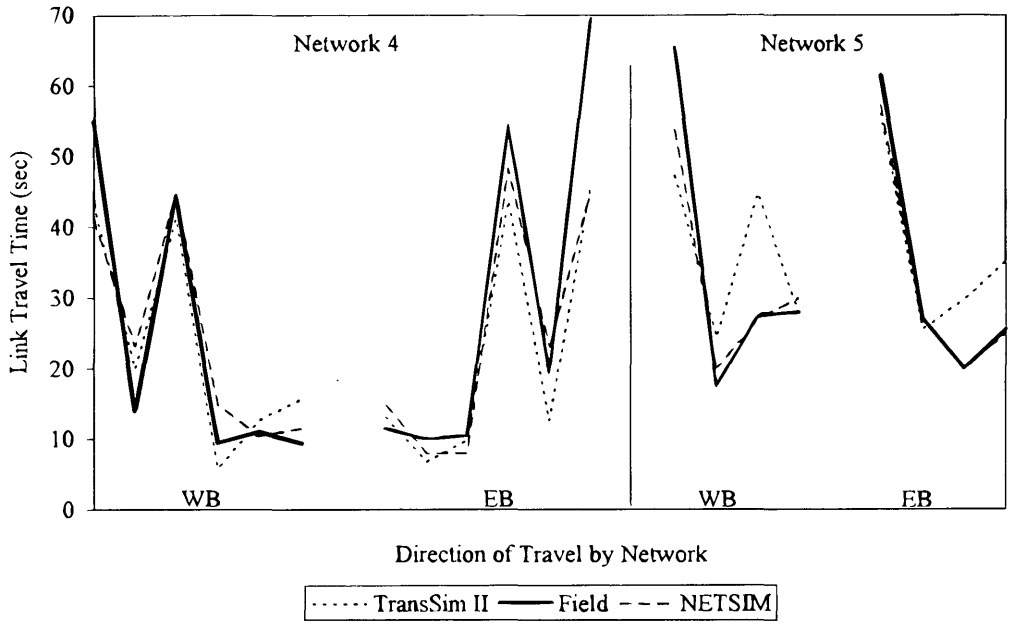


FIGURE 3 NETSIM and TransSim II™ LRT travel time comparison to validation data.

Traffic travel time comparisons showed that the model moderately replicated individual link travel times and accurately represented the system directional travel times (Figure 2). Thirty-eight percent of the individual link travel times were accepted at the ±20-percent criteria established for comparison with the field data. In the directional travel time comparison, however, eight of the nine modeled directional traffic travel times were accepted at the 95-percent confidence level. As a measure of the individual intersection modeling performance of TransSim II™, stopped delay from the model was compared with the field data (Table 2). Correlation analysis indicated a moderate relationship between model and field stopped delay.

TABLE 2 TransSim II™ and Field Intersection MOEs

	Mean Stopped Delay	
Intersection	Validation Data	Model w/ LRT
Flower & Washington		
EB Approach	2.27	9.74
NB Approach	29.99	9.91
Central & Washington		
EB Approach	6.13	9.95
NB Approach	21.28	29.01
First & Pacific		
NB Approach	10	12.38
SB Approach	7.36	12.04
Broadway & Pacific		
NB Approach	16.19	10.07
SB Approach	20.86	19.12
MLK & Holladay		
SB Approach	6.01	1.78
122nd & Burnside		
NB Approach	31.41	17.29

LRT link travel time in Networks 4 and 5, presented in Figure 3, consisted of LRV travel time at ideal speed plus time delayed at signals in the network, LRV acceleration and deceleration at signals and stations, and the dwell times at stations to service passengers. This information was taken from the TransSim II™ output by adding the LRT delay at each intersection to the ideal travel time along transit links and, for links with stations, also adding the time for passenger service and time lost during deceleration and acceleration. Similar to the results for the traffic analysis, the directional travel times for LRT were more accurate than the individual link travel times. Eight of the 20 individual link travel times were within ±20 percent of the field data, and three of the four directional travel times were accepted at the 95-percent confidence level.

The graphics could be viewed for an individual intersection or for the entire transit corridor being modeled. Inspection of the graphics for each intersection showed the simulation time, signal status for each approach, queue buildup during red indications, presence of LRVs, and priority calls and recovery periods attributable to transit. The systemwide view afforded by the graphics helped identify coding errors and contributed to an understanding of LRT treatment in the model.

CONCLUSIONS

The performance of the calibrated models investigated in this study was assessed by comparisons of link travel times, network directional travel times, and individual intersection MOEs to field-observed values. NETSIM and TransSim II™ could replicate the general trends of link travel times, but were only able to reproduce roughly 50 percent of link travel times within ±20 percent. Both models performed well for directional travel time comparison to the validation data. Eight out of nine directional travel times were accepted at the 95-percent confidence level for each model.

Individual intersection model-stopped delay and percent-stops output from NETSIM correlated with their field counterparts in a moderate and strong relationship, respectively. TransSim II™ was a moderate predictor of individual intersection-stopped delay. For the LRT modeling investigation, both models were accurate in replicating systemwide travel time and moderately accurate in estimating link travel times.

Based on the results of this research, analysts concluded that both models could simulate the systems and control behavior of the LRT networks simulated. Model outputs were more representative of field data for systemwide travel times than for MOEs at individual intersections. Strengths of NETSIM include the ability to monitor queue spillback conditions and provide realistic modeling of oversaturated traffic conditions. Advantages of using TransSim II™ are realized in the ease of modeling the LRT environment and the explicit modeling of controller behavior and LRT priority algorithms.

This research has simulated LRT in nonpriority pretimed networks and full priority semiactuated and fully actuated networks. Other types of priority exist between these extremes and there are a variety of means to recover green on cross streets given up during priority calls. These additional priority types should be investigated and simulated using NETSIM and TransSim II™ to determine the best simulation configuration and format for each model.

ACKNOWLEDGMENTS

This report was derived from research conducted at the Texas Transportation Institute sponsored by the Texas Department of Transportation and the FHWA. Contributions to the research were made by Carol Walters, Ed Collins, Jim Cotton, and Greg Krueger. The data collection for this report involved the collaboration of many individuals. In Los Angeles, assistance was provided by Linda Meadow, James Curry, and Brian Gallagher. In Long Beach, assistance was provided by James P. B. Chen and Larry Bass. In Portland, data were collected with the help of William Kloos, Kent Lall, and Bruce Robinson.

REFERENCES

1. Berry, R. A. Estimating Level of Service of Streets with At-Grade Light Rail Crossings. In *ITE 1987 Compendium of Technical Papers,* Institute of Transportation Engineers, Washington, D.C., Aug. 1987, pp. 167–172.
2. *TRAF User Reference Guide, Publication No. FHWA-RO-92-060.* Federal Highway Administration, Office of Safety and Traffic Operations Research and Development, U.S. Department of Transportation, McLean, Va., May 1992.
3. Luedtke, P., S. Smith, H. Lieu, and A. Kanaan. Simulating DART's North Central Light Rail Line Using TRAF-NETSIM. In *63rd Annual Meeting Compendium of Technical Papers,* Institute of Transportation Engineers, Washington, D.C., Sept. 1993, pp. 60–64.
4. Rymer, B., J. C. Cline, and T. Urbanik. *Delay at Isolated Light Rail Transit Grade Crossings.* Texas Transportation Institute Report 339–10, Texas State Department of Highways and Public Transportation, Austin, Tex., 1987.
5. *TransSim II™ Data Input Instructions.* JRH Transportation Engineering, Eugene, Oreg., 1993.

Publication of this paper sponsored by Committee on Light Rail Transit.

Techniques To Assess Delay and Queue Length Consequences of Bus Preemption

BILL ALAN CISCO AND SNEHAMAY KHASNABIS

Two deterministic methods for assessing delay and queue length consequences of bus preemption at signalized intersections are presented. The procedures are adapted from queueing theory and address three types of preemption strategies: green extension, red truncation, and red interruption. Method 1 macroscopically simulates groups of vehicles at the intersection using regular signal timing and timing under preempted conditions. Method 2 uses microscopic simulation in which each vehicle is treated individually and traffic flow patterns are evaluated for the regular signal timing and timing under preemption condition. Both methods are applied to three intersections in Ann Arbor, Michigan, representing different volume levels at the cross street. Data on vehicle arrival, service, queue lengths, and delays were compiled from videotapes made at the intersections during the spring of 1994. The algorithms developed were used to assess changes in queue lengths and delays resulting from the revised signal timing. The two methods appear to be viable tools for evaluating traffic flow consequences of preemption. The case studies indicate some variations in the results between the three strategies tested, between the two methods used, and between the intersections representing different volume levels. Method 2 (microscopic) is preferred for lighter volume levels, and Method 1 (macroscopic) should be used for higher volume levels. Further research is recommended to validate the proposed methods.

Preemption is a method of providing preferential treatment to buses at signalized intersections. Because the number of passengers boarding and unboarding at bus stops varies, predicting the exact arrival time of buses at intersections is difficult. A preemption strategy, if properly designed, can ensure continuous green phases to buses at successive intersections.

The technology uses instrumented buses, detectors, sensors, and a real-time traffic control system that can detect an approaching bus, predict its exact arrival time at the intersection, and communicate the information to the signal control system. The advent of intelligent transportation systems (ITSs) has made preemption a more viable tool for providing priority to buses than any time in the past. A description of available technologies for signal preemption and system logic is available in the literature (*1–3*).

Three categories of preemption strategies include green extension, red truncation, and red interruption. During green extension the green phase on the bus street is prolonged by a fixed amount of time. Red truncation allows premature termination of the red phase on the bus street. In red interruption, a short green phase, not contiguous with the adjacent green, is injected within the red phase along the bus street. In all the cases, the result is an increase in green time along the bus street allowing the bus to cross the intersection (*4*).

Experience with signal preemption in the United States and in Europe, although limited, suggests that signal preemption is a viable tool and, if properly implemented, may result in significant operational improvements along bus routes, including reduced delays and queue lengths, and increased throughput. However it also may adversely affect the traffic operation along the cross street by increasing delays and queue lengths and reducing throughput. Unfortunately no technique is available that can be used to assess the possible consequences of preemption. Without such an assessment tool, the only way to evaluate a preemption strategy is to actually implement the program and conduct a before-and-after study. Such an approach is not considered viable because of difficulties associated with conducting such field experiments under controlled conditions.

Wayne State University, Detroit, Mich. 48202.

Two deterministic methods are presented to assess some of the operational consequences of bus preemption at isolated signalized intersections. Both methods are adapted from queueing theory and are designed to assess the three types of preemption strategies mentioned earlier (green extension, red truncation, and red interruption). Initial results of the application of the two methods on one intersection were reported at the March 1995 ITS National Conference in Washington, D.C. (*5*). A more complete description of the application of the two methods on three intersections is presented in the following sections.

The intersections selected are on Washtenaw Avenue (Route 4 on the Ann Arbor Transportation Authority System) in Ann Arbor, located in southeast Michigan approximately 50 mi west of Detroit. This street is a major transit corridor connecting the central business district (CBD) of a small town (Ypsilanti) with the western end of the city of Ann Arbor using a transfer point at the Ann Arbor CBD. The case study applications were based on actual vehicular arrival and service patterns at the intersection. The three intersections represent light, medium, and heavy traffic volumes on the cross street.

METHODOLOGY

The approach used to assess the operational consequences of preemption is adapted from queueing principles used in undersaturated situations (*6–8*). It was assumed that at each cycle the number of arrivals is less than the capacity of the approach, resulting in no overflow of vehicles from one cycle to the next. Thus, all vehicles queued during a given red phase cleared the intersection before the end of the green phase. Actual vehicular arrival, service data, and queue lengths were recorded on videotape for all approaches during the peak period of 5 p.m. to 6 p.m. These records were analyzed to determine the following:

- Vehicle arrival patterns at the intersection during the red and green phases;
- Vehicle service or processing patterns through the intersection during the first part of the green phase, until the queue was totally discharged; and

• Simultaneous vehicle arrival and service patterns after the queue was completely discharged during the later part of the green phase.

The information obtained was then used to simulate the possible traffic flow (in both bus and cross street directions) if signal preemption of a specified amount (10 sec) was granted to the bus street, following each of the three strategies separately.

Within the analytic framework of queueing discipline, two methodologies were developed. In Method 1, macroscopic simulation was used to represent the flow of a group of vehicles by fixed arrival rates, service rates, and simultaneous arrival and service rates. The rates were determined from repeated observations of the traffic flow and from queueing patterns recorded at the site.

Method 2 is based on microscopic simulation, in which actual traffic flow patterns (arrival, service, and simultaneous arrival and service) over a three consecutive-cycle period were examined for each vehicle individually. Traffic flow consequences on all approaches were assessed based on the "superimposed" conditions of green extension, red truncation, and red interruption.

Method 1

Method 1 uses average rates of arrivals, services, and concurrent arrival-services, with the revised signal timings (resulting from preemption) "superimposed" on a time-rate diagram to assess traffic flow consequences. The amber phase was assumed to be essentially a part of the green phase. The approach used for green extension is explained in the next section. The following notations were used:

λ = vehicle arrival rate (vehicles/second) starting at t_1 and ending at t_2;
μ = vehicle service rate (vehicles/second) starting at t_3 (beginning of the green phase) and ending at t_4, when the queue is totally discharged;
k = simultaneous arrival-service rate (vehicles/second) beginning at t_4 (after the queue is discharged) and ending at t_5;
c = cycle length (seconds);
r = red phase (seconds);
g = green phase (seconds), so that
$c = r + g$ (ignoring amber phase);
t_1 = time of arrival of the first vehicle during the red phase;
t_2 = time of arrival of the last vehicle during the red phase;
t_3 = end of the red phase when the first vehicle in the queue starts moving;
t_4 = time when last vehicle in the queue clears the intersection, denoting the beginning of the simultaneous arrival-service process; and
t_5 = time when the last vehicle in the simultaneous arrival-service mode clears the intersection.

(Note: all t_i values are measured in seconds from the start of the red phase along the bus street.)

For Bus Street (Figure 1)

The number of vehicles arriving for service during time period $(t_2 - t_1)$ is $\lambda_b(t_2 - t_1)$ and the number of vehicles serviced during

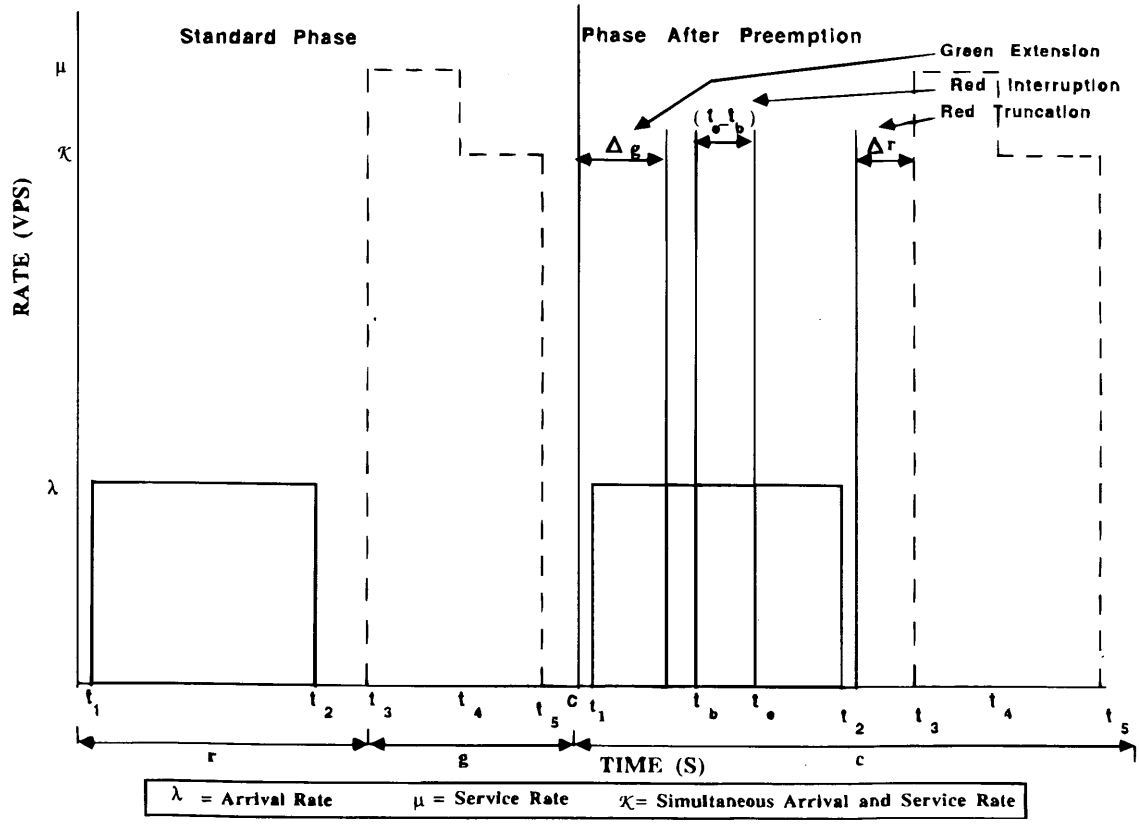


FIGURE 1 Arrival-service rate diagram (bus street).

time period $(t_4 - t_3)$ is $\mu_b(t_4 - t_3)$, until the queue is completely discharged.

Further, since the queue is totally discharged,

$\lambda_b(t_2 - t_1) = \mu_b(t_4 - t_3)$ (Note: the subscript "b" represents the bus street) and $Q_m = \lambda_b(t_2 - t_1)$ where Q_m is maximum queue length (in number of vehicles) and $t_4 - t_1$ is time duration of the queue.

The number of vehicles processed during simultaneous arrival and service is $k_b(t_5 - t_4)$. If Δg is the amount of green extension (seconds), then $\Delta g - t_1$ is the amount of green time effectively utilized by arriving vehicles so that the traffic consequences of Δg seconds of green extension per cycle can be derived as

$\lambda_b(\Delta g - t_1)$ = savings in the maximum queue length in number of vehicles (1)

$\lambda_b(\Delta g - t_1) \times (t_3 - \Delta g)$ = savings in delay in vehicle seconds (2)

For Cross Street (Figure 2)

The traffic consequences of Δg seconds of green extension for the bus street on the cross street are as follows:

$$\text{Increased delay to cross street} = \mu_c [(t_4 - t_3)\,\Delta r + (t_5 - t_4)\Delta r] = \mu_c \Delta r (t_5 - t_3) \quad (3)$$

(Note: the subscript "c" represents the cross street)

where Δr is additional red time along the cross street due to preemption (for all practical purposes, $\Delta r = \Delta g$, identified with the bus street).

Further, if the time needed for all the vehicles to clear the intersection (both those queued as well as those arriving during the green phase) exceeds the net green time (i.e., the original green time minus the lost time due to preemption) by an amount Δt, then additional delay and increase in queue length can be computed with Equations 4 and 5. In these equations, Δt_m is that portion of Δt comprising service events, and Δt_k is that portion of Δt comprising simultaneous arrival and service events. Additional delay and queue length can be computed as follows:

$$\text{Additional delay} = (\mu_c \Delta t_\mu + k_c \Delta t_k)\,(c\text{-}g) \quad (4)$$

$$\text{Increase in queue length} = \mu_c \Delta t_\mu + k_c \Delta t_k = \text{loss in throughput} \quad (5)$$

Similar algorithms for estimating delay and queue length consequences for red truncation and red interruption were developed separately for the bus street and cross street; these are not included in the text. However, results for all the three strategies are presented.

Method 2

Method 2 uses microscopic simulation, in which arrival and service rates are found by regression from the individual vehicle data points for three consecutive cycles, with the revised signal timings (resulting from preemption) altering the original signal timing for the first of the three consecutive cycles to assess traffic flow consequences. The amber phase was assumed to be part of the green phase.

The base condition data containing the arrival and service times of individual vehicles are used to find average arrival and service

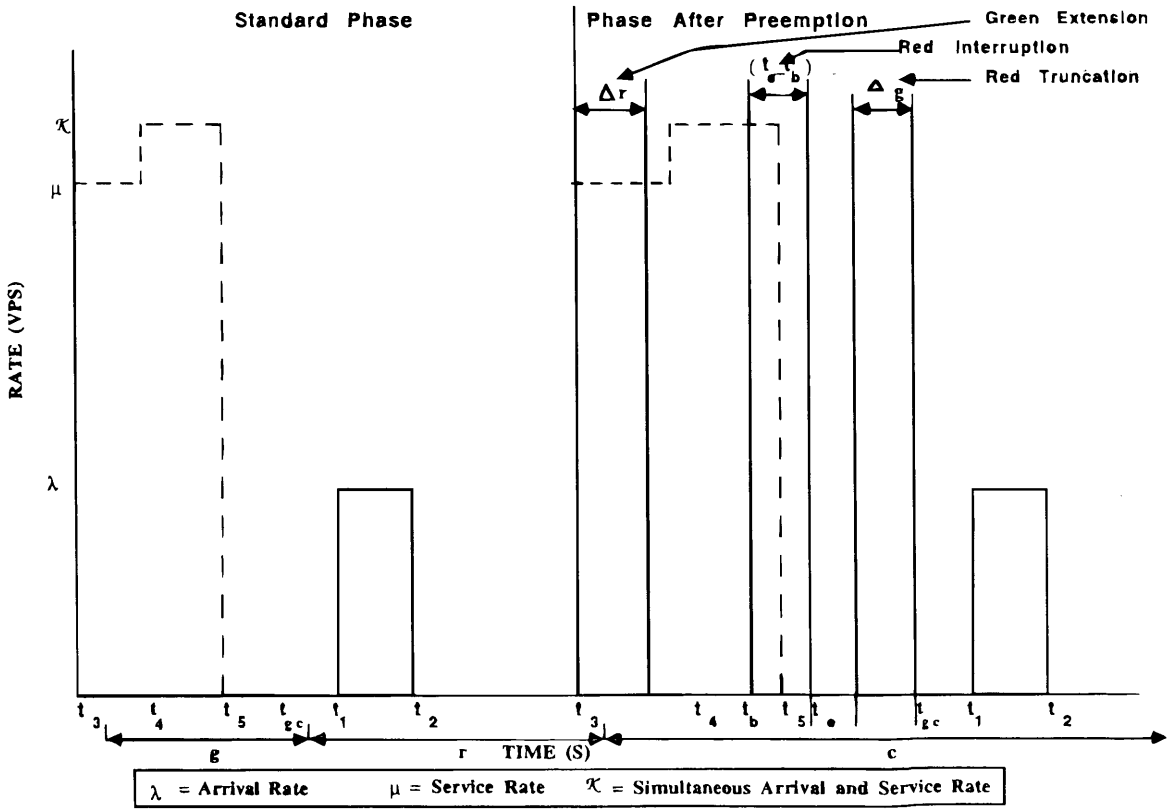


FIGURE 2 Arrival-service rate diagram (cross street).

rates by regression for all three cumulative cycles. These three values for arrival and service rates, as well as the number of vehicles that arrived during each cycle and the effective red times of each cycle, are then averaged. These values are then used to find the maximum queue, Q_m, the time duration of the queue, t_q, the average individual delay, and the total delay, *TD,* from the following equations:

$$\text{Maximum queue, } Q_m = \lambda r \quad (6)$$

where λ is the average of the arrival rates found by regression for the three cycles.

$$\text{Time duration of the queue, } t_q = \mu r / (\mu - \lambda) \quad (7)$$

where μ is the average of the service rates found by regression for the three cycles and r is the average of the effective red times for the three cycles.

$$\text{Average individual delay, } d_a = (r\, t_q) / 2c \quad (8)$$

where c is the average cycle length of the three cycles.

$$\text{Total delay, } TD = d_a N$$

where N is the average number of vehicles arriving in the three consecutive cycles.

The effect of a 10-sec green extension is found by analyzing the individual vehicular data for the second of the three cycles to determine the number of vehicles that arrive within the first 10 sec of this cycle, as these vehicles would now be processed during the first cycle due to the green extension. New values of λ and r are found, and the total delay is calculated as in the base condition using these new values. The change in the total delay is then the effect of the green extension. Essentially the same technique is used for the bus street and for the cross street, with the basic provision that a gain in Δg seconds of green time along the bus street would imply a loss of Δg seconds along the cross street.

RESULTS

Results of the application of the two methods on the three candidate intersections are presented in this section. The three intersections are designated as follows:

Volume Level on Cross St.	*Intersection*	*Description*	*Cycle Length (sec)*
Low	1	Washtenaw with Manchester/Sheridan	70
Medium	2	Washtenaw with Forest/Observatory	70
High	3	Washtenaw with Golfside	70

Compilation of Arrival and Service Rates

In Table 1, the t_i values for Intersection 1 are presented based on 10 cycles of observations. Table 1 shows that for the bus street through lane, the first vehicle arrived 10.3 sec after the start of the red phase. The last vehicle in the queue arrived at 23.4 sec. The first vehicle in the queue started moving at the beginning of the green phase at 26 sec and the last queued vehicle cleared the intersection at 42.1 sec. Between 42.1 and 64.9 sec, simultaneous arrivals and services occurred during the green phase. No arrivals were recorded between 64.1 sec and the end of the cycle at 70 sec.

TABLE 1 Values of t_i (seconds) for the Intersection of Washtenaw-Manchester/Sheridan

Approach (1)	Lane (2)	t_i - values (3) t_1	t_2	t_3	t_4	t_5
Washtenaw EB (Bus Street)	Thru 1	10.3	23.4	26.0	42.1	64.9
	Thru 2	10.8	23.9	26.0	43.0	64.4
	Right	17.9	20.4	26.0	43.6	52.5
	Left	x	x	26.0	61.8	63.5
Washtenaw WB (Bus Street)	Thru 1	9.7	25.2	26.0	41.7	57.3
	Thru 2	9.2	15.9	26.0	38.6	58.5
	Thru 3/Right	5.0	22.2	26.0	41.4	57.4
	Left	10.8	11.7	26.0	52.9	x
Manchester (Cross Street)	Thru/Right	46.3	54.3	0.0	10.5	10.8
	Left	41.3	41.9	0.0	2.6	x
Sheridan (Cross Street)	Thru/Right	39.9	54.3	0.0	5.5	x
	Left	35.3	49.6	0.0	5.8	x

Table 2 gives the average rates of arrival (λ), service (μ), and simultaneous arrival-service (k) compiled from the observation of 10 consecutive cycles. Expressed in vehicles per second, these rates are computed for each lane from average time intervals between successive arrivals and service. Similar information for the other two intersections was also derived from the data collected.

Traffic Operational Consequences

Intersection 1 A low volume

Tables 3, 4, and 5 show the operational consequences of the three preemption strategies for a 10-sec interval for the low-volume intersection using the models presented earlier. The negative signs represent reductions and the positive signs represent increases. Table 3 shows that the use of Method 1 will result in an increase of 31 vehi-

TABLE 2 Average Arrival and Service Rates (vehicles/second) for All Lanes at the Washtenaw-Manchester/Sheridan Intersection

Approach (1)	Lane (2)	Arrival Rate (λ) (3)	Service Rate (μ) (4)	Simultaneous Arrival/Service Rate (k) (5)
Washtenaw EB	Thru 1	0.53	0.51	0.46
	Thru 2	0.56	0.48	0.49
	Right	0.64	0.37	0.37
	Left	x	0.03	0.14
Washtenaw WB	Thru 1	0.22	0.38	0.24
	Thru 2	0.28	0.24	0.18
	Thru 3 & Right	0.30	0.41	0.19
	Left	0.74	1.0	x
Manchester	Thru & Right	0.11	0.32	0.67
	Left	0.50	0.29	x
Sheridan	Thru & Right	0.17	0.31	x
	Left	0.16	0.13	x

TABLE 3 Traffic Operational Consequences of Green Extension(10 sec) on the Intersection 1(Low-volume) for Every Cycle Preempted

Approach	Lane	ΔDelay (veh-sec)		ΔQueue (veh)	
		Method 1	Method 2	Method 1	Method 2
Washtenaw EB (Bus Street)	Thru 1	0	-35	0	-0.9
	Thru 2	0	-43	0	-1.1
	Right	0	0	0	0
	Left	0	0	0	0
Washtenaw WB (Bus Street)	Thru 1	-1	-13	-0.1	-0.5
	Thru 2	-4	0	-0.2	0
	Thru 3/Right	-24	-5	-1.5	-0.3
	Left	0	0	0	0
Total - Washtenaw		-29	-96	-1.8	-2.8
Manchester (Cross Street)	Thru/Right	35	14	0	0.7
	Left	8	0.6	0	0.6
Sheridan (Cross Street)	Thru/Right	9	2.7	0	0.4
	Left	8	0	0	0
Total - Manchester/Sheridan		60	17.3	0	1.7
Total - Intersection		31	-78.7	-1.8	-1.1

cle-sec of delay and a reduction of 1.8 vehicles of queue length for the intersection as a whole for every cycle preempted by 10 sec of green extension. Corresponding figures for each approach are also presented in Table 3. Method 2 predicts reductions in delay of 78.7 vehicle-sec and in queue length by 1.1 sec. The following observations may be made:

- Table 4 shows that with Method 1, for every cycle preempted by a 10-sec red truncation, 662 vehicle-sec of savings in delay and 13.2 vehicles of queue length savings will result. All of these savings will result from the bus street with no adverse effect on the cross street. Predictions by Method 2 are considerably smaller, with reductions in delay and in queue length of 100.5 vehicle-sec and 6.1 vehicles, respectively.
- Table 5 shows that as a result of 10 sec of red interruption, 925 vehicle-sec of delay and 15.6 vehicles of queue length will be saved per cycle preempted, as predicted by Method 1. A minimal adverse effect will be observed on the cross street (10 vehicle-sec). With Method 2, savings in delay are considerably lower.

Intersection 2: Medium Volume

Tables 6, 7, and 8 show the operational consequences of the three preemption strategies on Intersection 2 (medium volume). These three tables may be interpreted in the same manner as Tables 3, 4, and 5. The differences in the model output predicted by Method 1 versus Method 2 appear to have decreased some-

TABLE 4 Traffic Operational Consequences of Red Truncation(10 sec) on the Intersection 1(Low-volume) for Every Cycle Preempted

Approach	Lane	ΔDelay (veh-sec)		ΔQueue (veh)	
		Method 1	Method 2	Method 1	Method 2
Washtenaw EB (Bus Street)	Thru 1	-82	-37	-3.9	-1.0
	Thru 2	-82	-43	-3.8	-1.1
	Right	-65	0	-1.6	0
	Left	-11	0	0	0
Washtenaw WB (Bus Street)	Thru 1	-60	-14	-2.0	-0.4
	Thru 2	-30	-1.5	0	-3.4
	Thru 3/Right	-63	-5.0	-1.9	-0.2
	Left	-269	0	0	0
Total - Washtenaw		-662	-100.5	-13.2	-6.1
Manchester (Cross Street)	Thru/Right	0	0	0	0
	Left	0	0	0	0
Sheridan (Cross Street)	Thru/Right	0	0	0	0
	Left	0	0	0	0
Total - Manchester/Sheridan		0	0	0	0
Total - Intersection		-662	-100.5	-13.2	-6.1

TABLE 5 Traffic Operational Consequences of Red Interruption(10 sec) on the Intersection 1(Low-volume) for Every Cycle Preempted

Approach	Lane	ΔDelay (veh-sec)		ΔQueue (veh)	
		Method 1	Method 2	Method 1	Method 2
Washtenaw EB (Bus Street)	Thru 1	-123	-67	-4.1	-1.3
	Thru 2	-120	-46	-4.0	-1.2
	Right	-95	0	-0.1	0
	Left	-13	0	0	0
Washtenaw WB (Bus Street)	Thru 1	-90	-33.0	-1.8	-1.1
	Thru 2	-49	-3.1	-1.9	-9.2
	Thru 3/Right	-96	0	-3.0	0
	Left	-349	-5.8	-0.7	-0.1
Total - Washtenaw		-935	-155	-15.6	-12.9
Manchester (Cross Street)	Thru/Right	10	0	0	0
	Left	0	0	0	0
Sheridan (Cross Street)	Thru/Right	0	0	0	0
	Left	0	0	0	0
Total - Manchester/Sheridan		10	0	0	0
Total - Intersection		-925	-155	-15.6	-12.9

what for the red interruption and red truncation strategies, compared with the low-volume intersection presented in the previous section.

Intersection 3: High Volume

Tables 9, 10, and 11 give the traffic operational consequences predicted by the two methods on the high-volume intersection. The differences in the operational consequences predicted by the two methods appear to have decreased somewhat compared with the medium volume level intersection, particularly for the green extension and red truncation strategies than for the red interruption strategy.

Delay In Person-Seconds

The delay results presented in this study are expressed in vehicle-seconds, where each vehicle is the basic unit of measurement. However, the overriding impetus for preemption is the fact that a bus generally carries many more passengers than a car. Preemption is viewed as a means to increase the throughput of persons rather than vehicles. To account for this, the delay data were converted from vehicle-seconds to person-seconds using the following assumptions:

- Preemption is triggered only if at least one of the approaching vehicles is a bus,
- No more than one bus benefits from the preemption,

TABLE 6 Traffic Operational Consequences of Green Extension(10 sec) on the Intersection 2(Medium-volume) for Every Cycle Preempted

Approach	Lane	ΔDelay(veh-sec)		ΔQueue (veh)	
		Method 1	Method 2	Method 1	Method 2
Washtenaw EB (Bus Street)	Thru	0	-14	0	0
	Thru/Right	-9.3	-53	-0.4	-0.7
	Left	0	0	0	0
Washtenaw WB (Bus Street)	Thru	0	-31	0	-0.8
	Thru/Right	0	0	0	0
	Left	0	0	0	0
Total - Washtenaw		-9.3	-98	-0.4	-1.5
Observatory (Cross Street)	Thru	0	0	0	0
	Thru/Right	0	0	0	0
	Left	183	6	2.9	0.6
Forest (Cross Street)	Thru	0	0	0	0
	Thru/Right	0	0	0	0
	Left	103	0	1.6	0
Total - Forest/Obsevatory		286	6	4.5	0.6
Total - Intersection		277	-92	4.1	-0.9

TABLE 7 Traffic Operational Consequences of Red Truncation(10 sec) on the Intersection 2(Medium-volume) for Every Cycle Preempted

Approach	Lane	ΔDelay (veh-sec)		ΔQueue (veh)	
		Method 1	Method 2	Method 1	Method 2
Washtenaw EB (Bus Street)	Thru	-47	-38	0	-0.6
	Thru/Right	-56	-38	0	-0.5
	Left	-8	0	0	0
Washtenaw WB (Bus Street)	Thru	-43	0	0	0
	Thru/Right	-33	-15	0	-0.4
	Left	0	0	0	0
Total - Washtenaw		-187	-91	0	-1.5
Observatory (Cross Street)	Thru	0	6	0	0.2
	Thru/Right	0	4	0	0.6
	Left	0	0	0	0
Forest (Cross Street)	Thru	0	0	0	0
	Thru/Right	0	0	0	0
	Left	14	0	0.3	0
Total - Forest/Observatory		14	10	0.3	0.8
Total - Intersection		-173	-81	0.3	-0.7

- The candidate bus is traveling on the rightmost through lane on Washtenaw eastbound,
- Because of the unidirectional peak flow (p.m.) in the easterly direction, no bus on Washtenaw westbound benefits from preemption,
- Average automobile occupancy is 1.3 passengers/vehicle,
- Average bus occupancy is 20 persons/bus, and
- Bus stop location is at far side.

Revised Output for Red Interruption

Because the system is burdened with an additional amber phase for every red interruption granted, the authors decided that a correction should be made to the results of red interruption. Although the amber phase is considered part of the green phase, the correction is recommended to make up for a basic inconsistency in the assumption. In the cases of green extension and red truncation, the system is not subjected to the burden of an additional amber phase because the extension or truncation periods are contiguous to the regular green or red phase. This is not true for the red interruption strategy. For every interruption granted, an additional amber phase is required. Therefore the results were corrected by a factor of 0.65 (6.5 net green-sec out of a total of 10 sec, to provide for a 3.5-sec amber phase).

The delay and queue length data based on person-seconds and persons are shown in Table 12. It should be noted that for the green extension strategy, no bus qualified for preemption for the low-volume intersection. However, for the purpose of computing delay in person-seconds, an assumption was made that one bus (carrying 20 passengers) benefited from preemption. This assumption is justified because unless a bus preempts the signal, the increase in delay for the cross street would not materialize either.

TABLE 8 Traffic Operational Consequences of Red Interruption(10 sec) on the Intersection 2(Medium-volume) for Every Cycle Preempted

Approach	Lane	ΔDelay (veh-sec)		ΔQueue (veh)	
		Method 1	Method 2	Method 1	Method 2
Washtenaw EB (Bus Street)	Thru	-56	45	-2.0	-1.6
	Thru/Right	-64	-117	-2.3	-0.8
	Left	-10	0	-0.3	0
Washtenaw WB (Bus Street)	Thru	-49	0	-2.7	0
	Thru/Right	-37	-33	-1.6	-1.5
	Left	0	0	0	0
Total - Washtenaw		-216	-105	-8.9	-3.9
Observatory (Cross Street)	Thru	16	7	0	0.2
	Thru/Right	25	4	0	0.7
	Left	8	0	0	0
Forest (Cross Street)	Thru	18	0	0	0
	Thru/Right	0	0	0	0
	Left	0	0	0	0
Total - Forest/Observatory		67	11	0	0.9
Total - Intersection		-149	-94	-8.9	-3.0

TABLE 9 Traffic Operational Consequences of Green Extension(10 sec) on the Intersection 3(High-volume) for Every Cycle Preempted

Approach	Lane	ΔDelay (veh-sec)		ΔQueue (veh)	
		Method 1	Method 2	Method 1	Method 2
Washtenaw EB (Bus Street)	Thru	-87	-82	-1.6	-0.9
	Thru/Right	-48	-43	-0.9	-0.6
	Left	0	0	0	0
Washtenaw WB (Bus Street)	Thru	0	-34	0	-0.3
	Thru/Right	-46	-56	-0.8	-0.7
	Left	0	0	0	0
Total - Washtenaw		-181	-215	-3.3	-2.5
Golfside SB (Cross Street)	Thru	0	0	0	0
	Thru/Right	0	0	0	0
	Left	129	28	0.8	0.3
Golfside NB (Cross Street)	Thru	0	0	0	0
	Thru/Right	0	0	0	0
	Left	103	12	0.3	0.2
Total - Golfside		232	40	1.1	0.5
Total - Intersection		51	-175	-2.2	-2.0

TABLE 11 Traffic Operational Consequences of Red Interruption (10 sec) on the Intersection 3(High-volume) for Every Cycle Preempted

Approach	Lane	ΔDelay (veh-sec)		ΔQueue (veh)	
		Method 1	Method 2	Method 1	Method 2
Washtenaw EB (Bus Street)	Thru	-237	-267	-2.2	-2.6
	Thru/Right	-186	-183	-1.8	-2.4
	Left	0	0	0	0
Washtenaw WB (Bus Street)	Thru	-207	-100	-1.5	0.4
	Thru/Right	-223	-207	-1.9	-2.4
	Left	0	0	0	0
Total - Washtenaw		-843	-757	-7.4	-7.0
Golfside SB (Cross Street)	Thru	249	0	2.4	0
	Thru/Right	418	468	4.4	0.8
	Left	0	0	0	0
Golfside NB (Cross Street)	Thru	56	0	0.5	0
	Thru/Right	277	0	2.9	0
	Left	0	0	0	0
Total - Golfside		1000	468	10.2	0.8
Total - Intersection		157	-289	2.8	-6.2

Results of Method 1 versus Method 2

A review of the results presented in Table 12 indicates significant differences in the delay output derived by the two methods. However the differences are less significant when comparing the queue data. Also, differences in delay appear to be more significant in the case of the low-volume, undersaturated intersection. Method 2 generally appears to under-predict savings in delay compared with Method 1. Further research is necessary before these differences can be fully explained. Current literature suggests that macroscopic simulation (Method 1) is more effective under steady-state conditions than under random flow conditions. When vehicular arrival rates are high, any randomness in the individual arrivals around the mean becomes insignificant. In the present study (for the low-volume intersection in particular), arrival rates cannot be considered high. Thus, the assumption of uniform arrival rate by Method 1 (based on data collected for 10 cycles) can be questioned.

By contrast, Method 2 uses individual vehicular arrival data over a three-cycle period. Because of the randomness in arrivals at low volumes, microscopic simulation may be considered more effective for Intersection 1. A major disadvantage of the results obtained by Method 2 is that simulation was conducted over a three-cycle period, and data collected over such a limited period may not be representative of the majority of arrivals during peak hours of operation. Both methods have advantages and disadvantages, and a decision to use Method 1 or Method 2 should be made based on traffic conditions. Intuitively, microscopic simulation (Method 2) should work better under low-volume conditions because of the implicit assumption of constant arrival patterns under macroscopic simulation (Method 1). From the point of view of statistical reliability, for high-volume conditions (when arrival patterns are likely to be more homogeneous) Method 1 appears to be a better approach. However, all of these observations require validation through further research.

TABLE 10 Traffic Operational Consequences of Red Truncation (10 sec) on the Intersection 3(High-volume) for Every Cycle Preempted

Approach	Lane	ΔDelay(veh-sec)		ΔQueue (veh)	
		Method 1	Method 2	Method 1	Method 2
Washtenaw EB (Bus Street)	Thru	-114	-55	0	-0.5
	Thru/Right	-87	-40	0	-0.5
	Left	146	29	1.5	0.3
Washtenaw WB (Bus Street)	Thru	-99	-31	-1.1	-0.4
	Thru/Right	-118	-42	-0.2	-0.5
	Left	0	6	0	-0.5
Total - Washtenaw		-272	-133	0.2	-2.1
Golfside SB (Cross Street)	Thru	0	0	0	0
	Thru/Right	0	0	0	0
	Left	0	0	0	0
Golfside NB (Cross Street)	Thru	0	0	0	0
	Thru/Right	0	0	0	0
	Left	0	0	0	0
Total - Golfside		0	0	0	0
Total - Intersection		-272	-133	0.2	-2.1

TABLE 12 Traffic Operational Consequences Expressed at the Personal Level for Three Preemption Strategies

Strategy	Intersection	ΔDelay		ΔQueue Length	
		Method 1	Method 2	Method 1	Method 2
Green Extension	1	-280	-389	-22.3	-20.1
	2	-991	-1292	-22.9	-21.8
	3	-168	-628	-14.7	-21.0
Red Truncation	1	-1048	-318	-35.9	-26.6
	2	-541	-360	-19.7	-22.1
	3	-412	-292	-19.6	-20.3
Red Interruption Original	1	-1394	-356	-39.0	-35.5
	2	17	-563	-15.1	-26.8
	3	-381	-309	-30.3	-22.8
Red Interruption Corrected	1	-906	-234	-25.4	-23.1
	2	13	-366	-12.0	-17.4
	3	-248	-201	-19.7	-15.0

Comparative Results of the Three Strategies

It is difficult, if not impossible, to make any comparative evaluation of the three preemption strategies from the limited data developed in this study. The strategies' actual effectiveness depends on how many vehicles clear the intersection along the bus street behind the bus that triggers the preemption device, and how many vehicles are made to stop along the cross street. If vehicular arrivals are random (for low-volume conditions), a large sample size would be required to discern any trends. For uniform arrivals (because of homogeneity in arrival patterns), general conclusions may be derived with a smaller sample size.

CONCLUSIONS

The purpose of this study is to present a procedure for assessing delay and queue length consequences of bus preemption at signalized intersections. Two methods are presented that are adapted from queueing theory and that use a deterministic approach to simulate traffic flow at the intersection by superimposing a revised signal phasing on the regular signal phasing. Constant rates of arrivals, services, and simultaneous arrivals and services were used in the manual simulation of Method 1, and a regression analysis was performed on the data points to obtain arrival and service rates for Method 2. Separate procedures for green extension, red truncation, and red interruption were developed. The procedures developed were applied to three signalized intersections on a major bus corridor in Ann Arbor, Michigan, representing various volume levels. The conclusions of the study are:

1. The procedures developed appear viable, and the case studies presented indicate some variations in the results from the three strategies, as well as from the three intersections. Such variations are expected due to the inherent differences in the nature of the strategies as well as in the vehicular arrival patterns at the intersection.
2. The case studies appear to indicate more significant differences in the delay results for Method 1 and Method 2. However, differences in queue lengths as derived by the two methods for compatible strategies are less significant.
3. The validity of the assumption of constant arrival, service, and simultaneous arrival and service rates made by Method 1 can be questioned for the low-volume intersection. Under such circumstances, Method 2 appears to be more effective.
4. For higher volume levels, Method 1 may be more appropriate.
5. Further research is needed to validate the proposed methods.

ACKNOWLEDGMENT

Support for this study was provided by the U.S. Department of Transportation through the University of Michigan, Ann Arbor, under the Great Lakes Center for Truck and Transit Research Scholars Program. Matching support was provided by the College of Engineering and the College of Urban Labor and Metropolitan Affairs, Wayne State University. Much of the data used was obtained through the support of the Michigan Department of Transportation and the Ann Arbor Transportation Authority. The authors express their appreciation to these agencies for their support and cooperation. The authors also express their appreciation to Rajasekhar Reddy Karnati, graduate research assistant for a companion project, for his assistance in the collection and compilation of the data.

REFERENCES

1. Casey, R. F. et al. *Advanced Public Transportation Systems: The State of the Art.* Report DOT-VNTSC-UMTA-91–2. U.S. Department of Transportation, 1991.
2. *Assessment of Advanced Technologies for Transit and Rideshare Applications.* Final Report. NCTRP Project 60-1A. Castle Rock Consultants, July 1991.
3. Vuchic, V. R. *Urban Public Transportation Systems and Technology.* Prentice-Hall, 1981.
4. Khasnabis, S., et al. Evaluation of Operating Cost Consequences of Signal Preemption as an IVHS Strategy. *TRB Record 1390,* Transportation Research Board, National Research Council, 1993 pp. 3–9.
5. Cisco, B. A., and S. Khasnabis. *A Comparative Analysis of Two Methods to Assess Operational Consequences of Bus Preemption.* Presented at the 5th Annual Meeting of ITS America, March 1995.
6. May, A. D. *Traffic Flow Fundamentals.* Prentice-Hall, Inc., New Jersey, 1990.
7. Newell, G. F. Approximation Methods for Queue's with Application to the Fixed Cycle Traffic Light. *SIAM Review,* Vol. 7, No. 2, April 1965 pp. 223–240.
8. Lee, A. M. *Applied Queuing Theory.* McMillan, New York, 1966.

Publication of this paper sponsored by Committee on Traffic Signal Systems.